CHAPTER

11

PROCESS IMPROVEMENT AND PATIENT FLOW

Operations Management in Action

Cambridge Health Alliance Whidden Hospital in Everett, Massachusetts, is a safety net hospital whose emergency department (ED) was experiencing long waits, inefficient processes, and poor patient satisfaction. Its leaders undertook two projects to improve patient flow: an ED facility expansion, and, two years later, a reorganization of patient flow and the establishment of a rapid assessment unit (RAU).

In the period following the ED expansion, significant negative trends were observed: decreasing Press Ganey patient satisfaction percentiles (–4.1 percentile per quarter), increasing door-to-provider time (+4.9 minutes per quarter), increasing duration of stay (+13.2 minutes per quarter), and increasing percentage of patients leaving without being seen (+0.11 per quarter).

After the RAU was established, significant immediate impacts were observed for door-to-provider time (–25.8 minutes) and total duration of stay (–66.8 minutes). The trends for these indicators further suggested the improvements continued to be significant over time. Furthermore, the negative trends for the Press Ganey outcomes observed after ED expansion were significantly reversed and continued to move in the positive direction after the RAU. The major conclusion from the project team was that the impact of process improvement and RAU implementation is far greater than the impact of renovation and facility expansion.

Source: Sayah et al. (2016).

OVERVIEW

At the core of all organizations are their operating systems. Excellent organizations continuously measure, study, and make improvements to these systems. This chapter provides a methodology for measuring and improving systems using a select set of the tools presented in the preceding chapters.

The terminology associated with process improvement can be confusing. Typically, tasks combine to form subprocesses, subprocesses combine to form processes, and processes combine to form a system. The boundaries of a particular system are defined by the activity of interest. For example, the boundaries of a supply chain system are more encompassing than those of a hospital system that is part of that supply chain.

The term *process improvement* refers to improvement at any of these levels, from the task level to the systems level. This chapter focuses on process and systems improvement.

Process improvement follows the classic plan-do-check-act (PDCA) cycle (chapter 9), with the following, more specific, key steps:

- *Plan:* Define the entire process to be improved using process mapping. Collect and analyze appropriate data for each element of the process.
- *Do:* Use a process improvement tool(s) to improve the process.
- *Check:* Measure the results of the process improvement.

(continued)

Problem Types

Continuous process improvement is essential for organizations to meet the challenges of today's healthcare environment. The theory of swift and even flow (TSEF) (Schmenner 2001, 2004; Schmenner and Swink 1998) asserts that a process is more productive as the stream of materials (customers or information) flows more swiftly and evenly. Productivity rises as the speed of flow through the process increases and the variability associated with that process decreases.

OVERVIEW *(Continued)*

- *Act to hold the gains:* If the process improvement results are satisfactory, hold the gains (chapter 15).

If the results are not satisfactory, repeat the PDCA cycle.

This chapter discusses the types of problems or issues faced by healthcare organizations, reviews many of the operations tools discussed in earlier chapters, and illustrates how these tools can be applied to process improvement. The relevant tools include the following:

- Basic process improvement tools
- Six Sigma and Lean tools
- Simulation software

Note that these phenomena are not independent. Often, decreasing system variability increases flow, and increasing flow decreases variability. For example, advanced-access (same day) scheduling increases flow by decreasing the elapsed time between when a patient schedules an appointment and when she has completed her visit with the provider. Applying this concept of interdependence to patient no-shows, advanced-access scheduling can decrease variability by decreasing the number of no-shows.

Solutions to many of the problems facing healthcare organizations can be found in increasing flow or decreasing variability. For example, a key operating challenge in most healthcare environments is the efficient movement of patients in a hospital or clinic, commonly called *patient flow*. Various approaches to process improvement can be illustrated using the patient flow problem. Optimizing patient flow through EDs has become a top priority of many hospitals; therefore, the Vincent Valley Hospital and Health System (VVH) example at the end of this chapter focuses on improving patient flow through that organization's ED.

Another key issue facing healthcare organizations is the need to increase the level of quality and eliminate errors in systems and processes. In other words, variation must be decreased. Finally, increasing cost pressures result in the need for healthcare organizations to improve processes and do so while reducing costs.

The tools and techniques presented in this book are aimed at enabling cost-effective process improvement. Although this chapter focuses on patient flow and elimination of errors related to patient outcomes, the discussion is equally applicable to other types of flow problems (e.g., information, paperwork)

and other types of errors (e.g., billing). Some tools are more applicable to increasing flow and others to decreasing variation, eliminating errors, or improving quality, but all of the tools can be used for process improvement.

Patient Flow

Efficient patient movement in healthcare facilities can significantly improve the quality of care patients receive and substantially improve financial performance. A patient receiving timely diagnosis and treatment has a higher likelihood of obtaining a desired clinical outcome than a patient whose diagnosis and treatment are delayed. Because most current payment systems are based on fixed payments per episode of treatment, a patient moving more quickly through a system tends to generate lower costs and, therefore, higher margins.

Patient flow optimization opportunities occur in many healthcare settings. Examples include operating suites, imaging departments, urgent care centers, and immunization clinics. Advanced-access scheduling is a special case of patient flow and is examined in depth in chapter 12.

Poor patient flow has several causes; one culprit discovered by many investigators is variability of scheduled demand. For example, if an operating room is scheduled for a surgery but the procedure does not take place at the scheduled time, or it takes longer than scheduled to complete, the rest of the surgery schedule becomes delayed. These delays ripple through the entire hospital, including the ED.

As explained by Eugene Litvak, PhD (2003):

> You have two patient flows competing for hospital beds—ICU or patient floor beds. The first flow is scheduled admissions. Most of them are surgical. The second flow is medical, usually patients through the emergency department. So when you have a peak in elective surgical demand, all of a sudden your resources are being consumed by those patients. You don't have enough beds to accommodate medical demand.

If scheduled surgical demand varies unpredictably, the likelihood of inpatient overcrowding, ED backlogs, and ambulance diversions increases dramatically.

A number of management solutions have been introduced to improve patient flow. Separating low-acuity patients into a unique treatment stream can reduce the time these patients spend in the ED and improve overall patient satisfaction (Rodi, Grau, and Orsini 2006). Other tools and methods that have been employed to improve flow once a patient is admitted to the hospital relate to the discharge process. These approaches include creating a uniform discharge time (e.g., 11:00 a.m.), writing discharge orders the night before

release, communicating discharge plans early in the patient's care, centralizing oversight of census and patient movement, changing physician rounding times, alerting ancillary departments when their testing procedures are critical to a patient's discharge, and improving discharge coordination with social services (Clark 2005).

Investments in health information technology (IT) can improve patient flow as well. Devaraj, Ow, and Kohli (2013) studied 576 US hospitals to investigate the relationship between IT and investments in smooth and even flow. Using risk-adjusted length of stay (LOS) as their measure of smooth and even flow, they found that IT investments were positively related to smooth and even flow (shorter LOS) at the .05 level of significance.

They provide an example of how this result occurs (Devaraj, Ow, and Kohli 2013, page 190):

> When the patient record is complete, the discharge IT system prompts the attending physician to access the patient record from the cloud. After reviewing the record, the attending physician can digitally sign the record and issue orders to discharge the patient. Because the entire patient record resides in the cloud, the attending physician can complete the entire process through a mobile device and discharge the patient from anywhere. If a hospital automated the current process that requires attending physicians to physically come to the hospital, often the next day, in order to review and sign discharge orders, the LOS may not be significantly reduced. Therefore, it is important for hospital managers to understand such complementarities (e.g., TSEF) to ensure that IT is appropriately placed in the patient care "system."

For patient flow to be carefully managed and improved, the formal methods of process improvement outlined in the next section need to be widely employed.

Process Improvement Approaches

Process improvement projects can use a variety of approaches and tools. Typically, they begin with process mapping and measurement. Some simple tools can be initially applied to identify opportunities for improvements. Identifying and eliminating or alleviating bottlenecks in a system (theory of constraints) can quickly improve overall system performance. In addition, the Six Sigma tools described in chapter 9 can be used to reduce variability in process output, and the Lean tools discussed in chapter 10 can identify and eliminate waste. Finally, simulation (discussed later in this chapter) is a powerful tool that enables understanding and optimization of flow in a system.

All major process improvement projects should use the formal project management methodology outlined in chapter 5. An important first step is to identify a system's owner: For a system to be managed effectively over time, it must have a designated individual who monitors the system as it operates, collects performance data, and leads teams to improve the system.

Many systems in healthcare do not have an owner and, therefore, operate inefficiently. For example, a patient may enter an ED, be assessed by the triage nurse, move to the admitting department, take a chair in the waiting area, be moved to an exam room, be seen by a floor nurse, have his blood drawn, and finally be examined by a physician. From the patient's point of view, this is one system, but these various hospital departments may be operating autonomously. System ownership problems can be remedied by multidepartment teams with one individual designated as the overall system or process owner.

Problem Definition and Process Mapping

Once the process owner is identified, the first step in improving a system is generally considered to be problem description and mapping of that process. However, the team should first ensure that the correct problem is being addressed. Mind mapping or root-cause analysis should be employed to ensure that the problem is identified and framed correctly; much time and money can be wasted finding an optimal solution to a process that is not problematic.

For example, suppose a project team is given the task of improving customer satisfaction with the ED. The team assumes that customer satisfaction is low because of high throughput time. It proceeds to optimize patient flow in the ED. Patient satisfaction does not improve.

Now, imagine that a second project team is assigned to improve customer satisfaction. It conducts an analysis of customer satisfaction, which reveals that customers are dissatisfied because of a lack of parking. The team solves the problem by following a different path than the first team because it has clearly understood and defined the issue, allowing team members to determine what process to map.

Processes can be described in a number of ways. The most common is the written procedure or protocol, typically constructed in the "directions" style. This type of process is sufficient for simple procedures—for example, "Turn right at Elm Street, go two blocks, and turn left at Vine Avenue." Clearly written procedures are an important part of defining standardized work, as described in chapter 10.

However, when processes are linked to form systems, they become complex. These linked processes benefit from process mapping because process maps

- provide a visual representation that allows process improvement through inspection,

- enable branching in a process,
- provide the ability to assign and measure the resources in each task in a process, and
- are the basis for modeling the process via computer simulation software.

Chapter 6 provides an introduction to process mapping. To review, the steps in process mapping are as follows:

1. Assemble and train the team.
2. Determine the boundaries of the process (where it starts and ends) and the level of detail desired.
3. Brainstorm the major process tasks, and list them in order. (Sticky notes are often helpful here.)
4. Generate an initial process map (also called a flowchart).
5. Draw the formal flowchart using standard symbols for process mapping.
6. Check the formal flowchart for accuracy by all relevant personnel.
7. Depending on the purpose of the flowchart, collect data needed or include additional information.

Process Mapping Example

A basic process map illustrating patient flow in VVH's emergency department is displayed in exhibit 11.1.

Here, the patient arrives at the ED and is examined by the triage nurse. If the patient is very ill (high complexity level), she is immediately sent to the intensive care section of the ED. If not, she is sent to admitting and then to the routine care section of the ED.

The simple process map shown in exhibit 11.1 ends with the routine care step. In actuality, other processes now begin, such as admission into an inpatient bed or discharge from the ED to home with a scheduled clinical follow-up. The VVH emergency department process improvement project is detailed at the end of this chapter.

Process Measurements

Once a process map is developed, relevant data are collected and analyzed. The situation at hand dictates which specific data and measures should be employed. Important measures and data for possible collection and analysis include the following:

- *Capacity of a process* is the maximum possible amount of output (goods or services) that a process or resource can produce or transform.

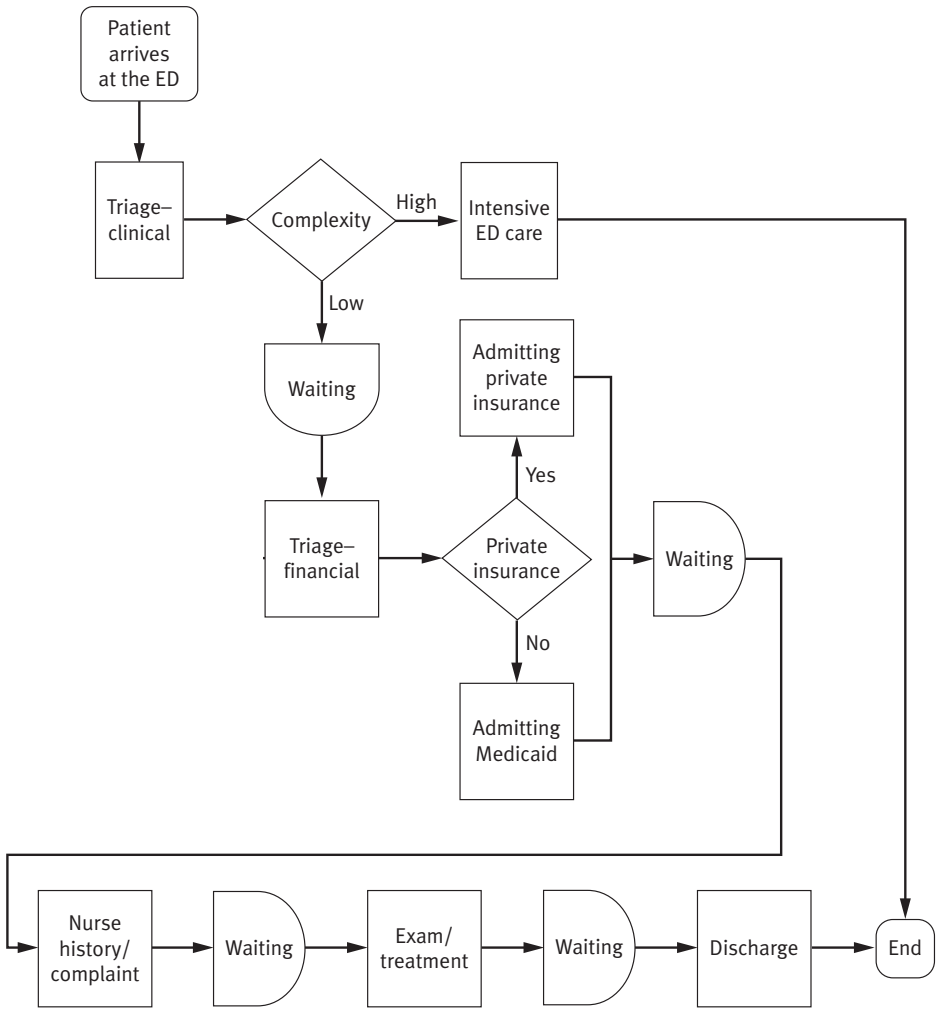


EXHIBIT 11.1 VVH Emergency Department (ED) Patient Flow Process Map

Note: Created with Microsoft Visio.

Capacity measures can be based on outputs or on the availability of inputs. The capacity of a series of tasks is determined by the lowest-capacity task in the series.

- *Capacity utilization* is the proportion of capacity actually being used. It is measured as actual output divided by maximum possible output.
- *Throughput time* is the average time a unit spends in the process. Throughput time includes both processing time and waiting time and is determined by the critical (longest) path through the process.
- *Throughput rate,* sometimes referred to as drip rate, is the average number of units that can be processed per unit of time.

- *Service time* or *cycle time* is the time to process one unit. The cycle time of a process is equal to the longest task cycle time in that process. The probability distribution of service times may also be of interest.
- *Idle time* or *wait time* is the time a unit spends waiting to be processed.
- *Arrival rate* is the rate at which units arrive to the process. The probability distribution of arrival rates may also be of interest.
- *Work-in-process, things-in-process, patients-in-process,* or *inventory* describes the total number of units in the process.
- *Setup time* is the amount of time spent getting ready to process the next unit.
- *Value-added time* is the time a unit spends in the process where value is being added to the unit.
- *Non-value-added time* is the time a unit spends in the process where no value is being added. Wait time is non-value-added time.
- *Number of defects or errors.*

The art in process mapping is to provide enough detail to be able to measure overall system performance, determine areas for improvement, and measure the impact of these changes.

Tools for Process Improvement

Once a system has been mapped, several techniques can be considered for improving the process. These improvements should result in a reduction in the duration, cost, or waste in a system.

Eliminate Non-Value-Added Activities

The first step after a system has been mapped is to evaluate every element to ascertain whether each is necessary and provides value (to the customer or patient). If a system has been in place for a long period and has not been evaluated through a formal process improvement project, elements of the system can likely be easily eliminated. This step is sometimes referred to as "harvesting the low-hanging fruit."

Eliminate Duplicate Activities

Many processes in systems have been added on top of existing systems without formally evaluating the total system, frequently resulting in duplicate activities. The most infamous redundant process step in healthcare is asking patients repeatedly for their contact information. Duplicate activities increase both time and cost in a system and should be eliminated whenever possible.

Combine Related Activities

Process improvement teams should examine both the process map and the activity and swim lane map. If a patient moves back and forth between departments, the movement should be reduced by combining these activities so he only needs to be in each department once.

Process in Parallel

Although a patient can only be in one place at one time, other aspects of her care can be completed simultaneously. For example, medication preparation, physician review of tests, and chart documentation can all be performed at the same time. As more tasks are executed simultaneously, the total time a patient spends in the process is reduced. Similar to a chef who has a number of dishes on the stove synchronized to be completed at the same time, much of the patient care process can be completed simultaneously.

Another element of parallel processing is the relationship of subprocesses to the main flow. For example, a lab result may need to be obtained before a patient enters the operating suite. Many of these subprocesses can be synchronized through the analysis and use of takt time (chapter 10). This synchronization enables efficient process flow, thereby optimizing the process.

Balance Workloads

If similar workers perform the same task, a well-tuned system can be designed to balance the work among them. For example, a mass-immunization clinic should develop its system so that all immunization stations are active at all times. This aim can be accomplished by using a single queue that feeds into multiple immunization stations.

Load balancing (or load leveling, heijunka) is difficult when employees can only perform a limited set of specific tasks (a consequence of the superspecialization of the healthcare professions). Load balancing is easier in environments that feature cross-training of employees than in those that limit employee tasks to singular functions.

Develop Alternative Process Flow Paths and Contingency Plans

The number and placement of decision points in the process should be evaluated and optimized. A system with few decision points has few alternative paths and, therefore, does not respond well to unexpected events. Alternative paths or contingency plans should be developed for these types of events. For example, a standard clinic patient rooming system should designate alternative paths for when an emergency occurs, a patient is late, a provider is delayed, or medical records are absent.

Establish the Critical Path

For complex pathways in a system, identifying the critical pathway with tools described in chapter 5 can be helpful. If a critical path can be identified, execution of processes on the pathway can be improved (e.g., reduce average service time). In some cases, the process can be moved off the critical path and be performed in parallel to it. Either technique decreases the total time on the critical pathway. In the case of patient flow, moving this process off the critical pathway decreases the patient's total time spent in the system.

Embed Information Feedback and Real-Time Control

Some systems have a high level of variability in their operations because they experience variability in the arrival of jobs or customers (patients) into the process and variability of the cycle time of each process in the system. High variability in the system can lead to poor performance. One tool to reduce variability is the control loop. Information can be obtained from one process and used to drive change in another. For example, the number of patients in the ED waiting area can be continuously monitored, and if it reaches a certain level, contingency plans—such as floating in additional staff from other portions of the hospital—can be initiated.

Ensure Quality at the Source

Many systems contain multiple reviews, approvals, and inspections. A system in which the task is performed correctly the first time should not require these redundancies. Deming (1998) first identified this problem in the process design of manufacturing lines that had inspectors throughout the assembly process. This expensive and ineffective system was one of the factors that gave rise to the quality movement in Japan and, later, the United States.

Systems should be designed to embed quality at their source or beginning to eliminate inspections. For example, a billing system that requires a clerk to inspect a bill before it is released does not have quality built into the process.

Match Capacity to Demand

A common problem in 24-hour healthcare operations is having too few or too many staff for patient care demand. This problem is exacerbated if an organization only allows set shifts (e.g., eight hours).

To solve this problem, first graph and analyze demand on an hourly and daily basis. Then develop staffing patterns that match this demand. For example, a five-hour or seven-hour shift might be needed to correctly meet the demand.

Using the tools in chapter 7, you should be able to identify patterns of demand (e.g., high ED demand on Friday and Saturday evenings). Chapter 12 also provides details on capacity planning.

Let the Patient Do the Work

The Internet and other advanced information technologies have allowed for increased self-service in service industries. Individuals are now comfortable booking their own airline reservations, buying goods online, and checking themselves out at retailers. This trend can be exploited in healthcare with tools that enable patients to be part of the process. For example, online tools are now available that allow patients to make their own clinic appointments. Letting the patient do the work reduces the work of staff and provides an opportunity for quality at the source—the data are more likely to be correct if the patients input them than if a staff member does so.

Use Technology

The electronic health record and other IT tools provide a platform to automate many tasks that were once performed manually. A good rubric through which to identify these tasks is to examine every daily task and ask where it ranks in complexity on the basis of your professional training. For those tasks that are low on this list, consider ways to automate them.

Today, work is an activity—not a place. The widespread use of smartphones and tablets enables work to be performed outside the traditional workplace. Consider moving some tasks to these devices to improve your personal productivity.

Apply the Theory of Constraints

Chapter 6 discusses the underlying principles and applications of the theory of constraints, which can be used as a powerful process improvement tool. First, the bottleneck in a system is identified, often through the observation of queues forming in front of it. Once a bottleneck is identified, it should be exploited and everything else in the system subordinated to it. Specifically, other nonbottleneck resources (or steps in the process) should be synchronized to match the output of the constraint. Idleness at a nonbottleneck resource costs nothing, and nonbottlenecks should never produce more than can be consumed by the bottleneck resource. Often, this synchronization causes the bottleneck to shift and a new bottleneck is identified. However, if the original bottleneck remains, the possibility of elevating the bottleneck needs to be considered. Elevating bottlenecks requires additional resources (e.g., staff, equipment), so a comprehensive financial and outcomes analysis needs to be undertaken to determine the trade-offs among process improvement, quality, and costs.

Identify Best Practices and Replicate

Although this tip does not describe a formal operations management tool, it must be mentioned as a highly recommended management approach. As

health systems expand, they are likely to have many similar activities replicated in separate geographic sites. Good management practice is to identify high-performing sites (e.g., the best primary care clinic in a system) and replicate their core processes throughout the organization.

A similar approach can be taken with individual employees. For example, study the best billing clerk in a hospital to understand her processes and then replicate them with all the billers in a department.

The Science of Lines: Queuing Theory

Queuing theory
The mathematical study of wait lines.

Although most people are familiar with waiting in line, few are familiar with, or even aware of, **queuing theory**, or the theory of waiting lines. Most people's experience with waiting lines is when they are actually part of those lines, for example, when waiting to check out in a retail environment. In a manufacturing environment, items wait in line to be worked on. In a service environment, customers wait for a service to be performed.

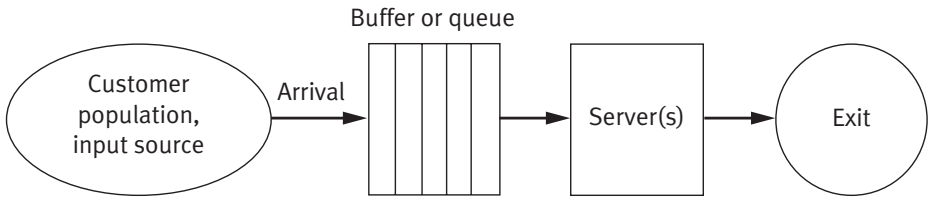
Queues, or lines, form because the resources needed to serve them (servers) are limited—deploying unlimited resources is economically unfeasible. Queuing theory is used to study systems to determine the best balance between service to customers (short or no waiting lines, implying many resources or servers) and economic considerations (few servers, implying long lines). A simple queuing system is illustrated in exhibit 11.2.

Customers (often referred to as entities) arrive and either are served (if there is no line) or enter the queue (if others are waiting to be served). Once they are served, customers exit the system.

The customer population, or input source, can be either finite or infinite. If the source is effectively infinite, the analysis of the system is easier than if it is finite because simplifying assumptions can be made.

The arrival process is characterized by the arrival pattern—the rate at which customers arrive (number of customers divided by unit of time)—or by the interarrival time (time between arrivals) and the distribution in time of those arrivals. The distribution of arrivals can be constant or variable. A

EXHIBIT 11.2
Simple Queuing System



constant arrival distribution has a fixed interarrival time. A variable, or random, arrival pattern is described by a probability distribution. The **queue discipline** is the method by which customers are selected from the queue to be served. Often, customers are served in the order in which they arrived—first come, first served. However, many other queue disciplines are possible, and choice of a particular discipline can greatly affect system performance. For example, choosing the customer whose service can be completed most quickly (shortest processing time) usually minimizes the average time customers spend waiting in line. This result is one reason urgent care centers are often located near an ED—urgent issues can usually be handled more quickly than true emergencies can.

Queue discipline In queuing theory, the method by which customers are selected from the queue to be served.

The service process is characterized by the number of servers and service time. Like arrivals, the distribution of service times can be constant or variable. Often, the exponential distribution (M) is used to model variable service times, μ is the mean service rate, λ is the mean arrival rate, and ρ is capacity utilization. (An exponential distribution creates data points that simulate a purely random process.)

Queuing Notation

The type of queuing system is identified with a specific notation in the form of A/B/c/D/E. The A represents the interarrival time distribution, and B represents the service time distribution. A and B together are represented as either a deterministic or a constant rate. The c represents the number of servers, D is the maximum queue size, and E is the size of the input population. When both queue and input population are assumed to be infinite, D and E are typically omitted. An M/M/1 queuing system, therefore, has an exponential service time distribution, a single server, an infinite possible queue length, and an infinite input population; it assumes only one queue. An M/M/1 queue for VVH is used as an example throughout the remainder of the chapter.

Queuing Solutions

Analytic solutions for some simple queuing systems at equilibrium or steady state (after the system has been running for some time and is unchanging, often referred to as a stable system) have been determined; however, the derivation of these results is outside the scope of this text. Refer to Cooper (1981) for a complete derivation and results for many other types of queuing systems.

Here, we focus primarily on the M/M/1 queuing system by presenting the results for an M/M/1 queue where $\lambda < \mu$—the arrival rate is less than the service rate. Note that if $\lambda \geq \mu$ (customers arrive faster than they are served), the queue becomes infinitely long, the number of customers in the system becomes infinite, waiting time becomes infinite, and the server experiences 100 percent capacity utilization (percentage of time the server is busy). The

following formulas can be used to determine some characteristics of the queuing system at steady state.

Capacity utilization:

$$\rho = \frac{\lambda}{\mu} = \frac{\text{Mean arrival rate}}{\text{Mean service rate}} = \frac{1/\text{Mean time between arrivals}}{1/\text{Mean service time}} = \frac{\text{Mean service time}}{\text{Mean time between arrivals}}$$

Average waiting time in queue:

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

Average time in the system (average waiting time in queue plus average service time):

$$W_s = W_q + \frac{1}{\mu} = \frac{1}{\mu - \lambda}$$

Average length of queue (average number in queue):

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \left(\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu - \lambda}\right)$$

Average total number of customers in the system:

$$L_s = \frac{\lambda}{\mu - \lambda} = \lambda W_s = \text{Arrival rate} \times \text{Time in the system}$$

Little's law
The relationship between the arrival rate to a system, the time an item (e.g., a patient) spends in the system, and the number of items in a system.

This last result is called **Little's law** and applies to all types of queuing systems and subsystems. To summarize this result in plain language, in a stable system or process, the number of things in the system is equal to the rate at which things arrive to the system multiplied by the time they spend in the system. In a stable system, the average rate at which things arrive to the system is equal to the average rate at which things leave the system. If this were not true, the system would not be stable.

Little's law can also be restated using other terminology:

Inventory (things in the system) = Arrival rate (or departure rate) × Throughput time (flow time)

or

$$\text{Throughput time} = \text{Inventory} \div \text{Arrival rate}$$

Knowledge of two of the variables in Little's law allows calculation of the third variable. Consider a clinic that serves 200 patients in an eight-hour day, or an average of 25 patients an hour. The average number of patients in the clinic (waiting room, exams rooms, etc.) is 15. Therefore, the average throughput time is

$$\begin{aligned} T &= I/\lambda \\ &= \frac{15 \text{ patients}}{25 \text{ patients/hour}} \\ &= 0.6 \text{ hour,} \end{aligned}$$

where T is throughput time, λ is patients per hour, and I is number of patients. Hence, each patient spends an average of 36 minutes in the clinic.

Little's law has important implications for process improvement and can be seen as the basis of many improvement techniques. Throughput time can be decreased by decreasing inventory or increasing departure rate. Lean initiatives often focus on decreasing throughput time (or increasing throughput rate) by decreasing inventory. The theory of constraints (chapter 6) focuses on identifying and eliminating system bottlenecks. The departure rate in any system is equal to 1 ÷ task cycle time of the slowest task in the system or process (the bottleneck). Decreasing the amount of time an object spends at the bottleneck task therefore increases the departure rate of the system and decreases throughput time.

Vincent Valley Hospital and Health System M/M/1 Queue

VVH began receiving complaints from patients related to crowded conditions in the waiting area for magnetic resonance imaging (MRI) procedures. The organization has determined a goal to average just one patient waiting in line for the MRI. It has collected data on arrival and service rates and sees that, for MRIs, the mean service rate (μ) is four patients per hour, exponentially distributed. VVH also finds that the mean arrival rate (λ) is three patients per hour. To find the capacity utilization of MRI (percentage of time the MRI is busy), VVH uses the following formula:

$$\rho = \frac{\lambda}{\mu} = \frac{3}{4} = 75\% \quad \text{or} \quad \rho = \frac{1/\mu}{1/\lambda} = \frac{15 \text{ minutes}}{20 \text{ minutes}} = 75\%.$$

If one customer arrives every 20 minutes and assuming each MRI takes 15 minutes to complete, the MRI is busy 75 percent of the time.

Next, VVH calculates patients' average time waiting in line,

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{3}{4(4-3)} = \frac{3}{4} = 0.75 \text{ hour},$$

and average time spent in the system,

$$W_s = \frac{1}{\mu - \lambda} = \frac{1}{4-3} = 1 \text{ hour}.$$

Finally, it determines average total number of patients in the system,

$$L_s = \frac{\lambda}{\mu - \lambda} = \frac{3}{4-3} = 3 \text{ patients}$$

or

$L_s = \lambda W_s$ = Arrival rate × Time in the system = 3 patients/hour × 1 hour = 3 patients,

and average number of patients in the waiting line,

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \left(\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu - \lambda}\right) = \left(\frac{3}{4}\right)\left(\frac{3}{4-3}\right)$$
$$= \frac{3^2}{4(4-3)} = \frac{9}{4} = 2.25 \text{ patients}.$$

To decrease the average number of patients waiting, VVH needs to decrease the utilization, $\rho = \lambda \div \mu$, of the MRI process. In other words, the service rate must be increased or the arrival rate decreased. VVH may increase the service rate by making the MRI process more efficient so that the average time to perform the procedure is decreased and MRIs can be performed on a greater number of patients in an hour. Alternatively, the organization may decrease the arrival rate by scheduling fewer patients per hour.

To achieve its goal (assuming that the service rate is not increased), VVH needs to decrease the arrival rate to

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\lambda^2}{4(4-\lambda)} = 1$$
$$\lambda^2 = 4 \times (4 - \lambda) = 16 - 4\lambda$$
$$\lambda^2 + 4\lambda - 16 = 0$$
$$\lambda = 2.47.$$

Alternatively (assuming that the arrival rate is not decreased), VVH may increase the service rate to

$$L_q = \frac{3^2}{\mu(\mu - 3)} = \frac{3^2}{\mu(\mu - 3)} = 1$$
$$\mu(\mu - 3) = \mu^2 - 3\mu = 3^2 = 9$$
$$\mu^2 - 3\mu - 9 = 0$$
$$\mu = 4.85.$$

VVH may also implement some combination of decreasing arrival rate and increasing service rate. In all cases, utilization of the MRI will be reduced to $\rho = \lambda \div \mu = 2.47 \div 4.00$, or $3.00 \div 4.85 = 0.62$.

Real systems are seldom as simple as an M/M/1 queuing system and rarely reach equilibrium. Often, simulation is needed to study these more complicated systems.

Discrete Event Simulation

Discrete event simulation (DES) is typically performed using commercially available software packages. As with Monte Carlo simulation, performing DES by hand is an option, albeit a tedious one. Two popular simulation software packages are Arena (Rockwell Automation 2016) and Simul8 (Simul8 Corporation 2016).

The terminology and general logic of DES are built on queuing theory. A basic simulation model consists of entities, queues, and resources, all of which can have various attributes. Entities are the objects that flow through the system; in healthcare, entities typically are patients, but they can be any object on which some service or task will be performed. For example, blood samples in the hematology lab are entities. Queues are the waiting lines that hold the entities while they await service. Resources (previously referred to as servers) can be people, equipment, or space for which entities compete.

The specific operation of a simulation model is based on states (variables that describe the system at a point in time) and events (variables that change the state of the system). Events are controlled by the simulation executive, and data are collected on the state of the system as events occur. The simulation jumps through time from event to event.

A simple example from the Vincent Valley Hospital and Health System M/M/1 MRI queuing discussion helps show the logic behind DES software. Exhibit 11.3 contains a list of the events as they happen in the simulation. The arrival rate is three patients per hour, and the service rate is four patients per hour. Random interarrival times are generated using an exponential distribution with a mean of 0.33 hours. Random service times are generated using an exponential distribution with a mean of 0.25 hours (shown at the bottom of exhibit 11.10 later in this chapter).

EXHIBIT 11.3
Simulation Event List

Just Finished			Variables		Attributes		Statistics				Upcoming Events		
Entity No.	Time	Event Type	Length of Queue	Server Busy	Arrival Time in Queue	Arrival Time in Service	Number of Complete Waits in Queue	Total Wait Time in Queue	Average Queue Length	Utilization	Entity No.	Time	Event
1	0.00	Arr	0	1	0.00	0.00	0	0	0	0	2	0.17	Arr
											1	0.21	Dep
2	0.17	Arr	1	1	0.17		0	0	0	1.00	1	0.21	Dep
											3	0.54	Arr
1	0.21	Dep	0	1	0.00	0.00	1	0.04	0.19	1.00	3	0.54	Arr
											2	0.77	Dep
3	0.54	Arr	1	1	0.54		1	0.04	0.07	1.00	2	0.77	Dep
											4	0.90	Arr
2	0.77	Dep	0	1	0.17	0.21	2	0.27	0.35	1.00	3	0.79	Dep
											4	0.90	Arr
3	0.79	Dep	0	0		0.77	3	0.27	0.34	1.00	4	0.90	Arr
											4	1.27	Dep
4	0.90	Arr	0	1		0.90	3	0.27	0.30	0.88	4	1.27	Dep
											5	1.49	Arr
						1	2	3	4	5	6	7	8
Interarrival time			Expon (0.33)			0.17	0.37	0.36	0.59	0.14	0.17	0.24	0.06
Time of arrival					0.00	0.17	0.54	0.90	1.49	1.63	1.80	2.04	2.10
Service time			Expon (0.25)		0.21	0.56	0.02	0.37	0.34	0.11	1.02	0.01	0.20

Note: Arr = arrival; Dep = departure; Expon = exponent.

The simulation starts at time 0.00. The first event is the arrival of the first patient (entity); there is no line (queue), so this patient enters service. Upcoming events are the arrival of the next patient at 0.17 hours (the interarrival between patients 1 and 2 is 0.17 hours) and the completion of the first patient's service at 0.21 hours.

The next event is the arrival of patient 2 at 0.17 hours. Because the MRI on patient 1 is not complete, patient 2 enters the queue. The MRI has been busy since the start of the simulation, so the utilization of the MRI is 100 percent. Upcoming events are the completion of the first patient's service at 0.21 hours and the arrival of patient 3 at 0.54 hours (the interarrival between patients 2 and 3 is 0.37 hours).

When the first patient's MRI is completed at 0.21 hours, no one is waiting in the queue because once patient 1 has completed service, patient 2 can enter service. The total waiting time in the queue for all patients is 0.04 hours (the difference between when patient 2 entered the queue and entered service). The average queue length is 0.19 patients. No people were in line for 0.17 hours, and one person was in line for 0.04 hours:

$$\frac{0 \text{ people} \times 0.17 \text{ hours} + 1 \text{ person} \times 0.04 \text{ hours}}{0.21 \text{ hours}} = 0.19 \text{ people.}$$

Upcoming events are the arrival of patient 3 at 0.54 hours and the departure of patient 2 at 0.77 hours (patient 2 entered service at 0.21 hours, and service takes 0.56 hours).

Patient 3 arrives at 0.54 hours and joins the queue because the MRI is still busy with patient 2. The average queue length has decreased from the previous event because more time has passed with no one in the queue—only one person has been in the queue for 0.04 hours, but total time in the simulation is 0.54 hours. Upcoming events are the departure of patient 2 at 0.77 hours and the arrival of patient 4 at 0.90 hours.

Patient 2 departs at 0.77 hours. No one is waiting in the queue at this point because patient 3 has entered service. Two people have departed the system. The total wait time in the queue for all patients is 0.04 hours for patient 2 plus 0.17 hours for patient 3 (0.77 hours – 0.54 hours) for a total of 0.21 hours. The average queue length is

$$\frac{0 \text{ people} \times 0.50 \text{ hours} + 1 \text{ person} \times 0.21 \text{ hours}}{0.77 \text{ hours}} = 0.35 \text{ people.}$$

The MRI utilization is still at 100 percent because the MRI has been busy constantly since the start of the simulation. Upcoming events are the departure

of patient 3 at 0.79 hours (patient 3 arrived at 0.54 hours, and service takes 0.25 hours) and the arrival of patient 4 at 0.90 hours.

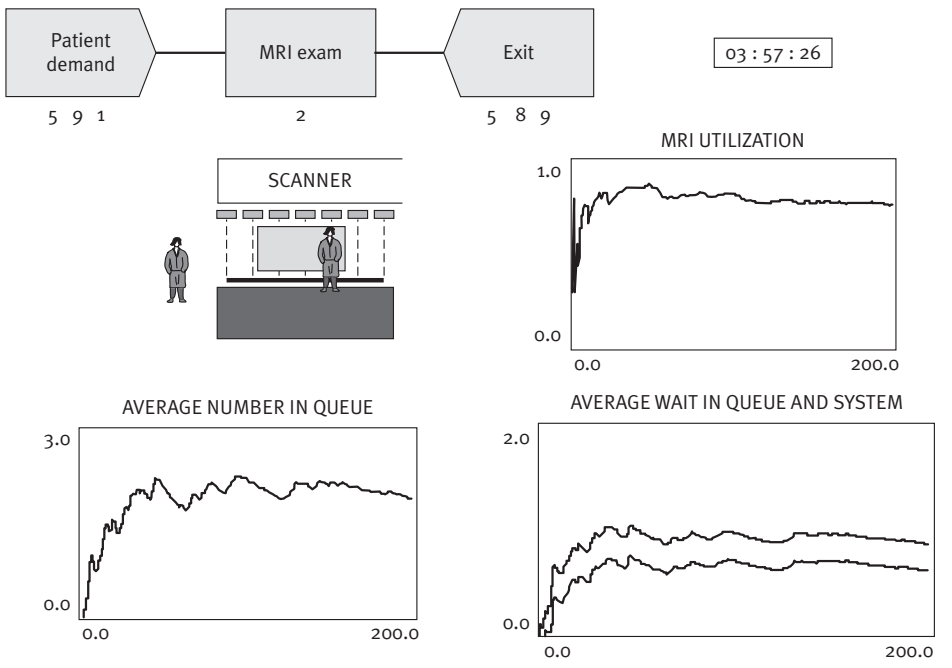
Patient 3 departs at 0.79 hours. Because no patients are waiting for the MRI, it becomes idle. Upcoming events are the arrival of patient 4 at 0.90 hours and the departure of patient 4 at 1.27 hours.

With patient 4 arriving at 0.90 hours and entering service, the utilization of the MRI has decreased to 88 percent because it was idle for 0.11 hours of the 0.90 hours the simulation has run. Upcoming events are the departure of patient 4 at 1.27 hours and the arrival of patient 5 at 1.49 hours. The simulation continues in this manner until the desired stop time is reached.

Even for this simple model, performing these calculations by hand takes a long time. Additionally, an advantage of simulation is that it uses process mapping; many simulation software packages are able to import and use Microsoft Visio process and value stream maps. DES software allows process improvement teams to build, run, and analyze simple models in limited time; Arena software was used to build and simulate the present model (exhibit 11.4).

As before, the arrival rate is three patients per hour, the service rate is four patients per hour, and both rates are exponentially distributed. Averages over time for queue length, wait time, and utilization for a single replication are

EXHIBIT 11.4
Arena Simulation of VVH MRI M/M/1 Queuing Example



Note: Created with Arena simulation software. M = exponential distribution; MRI = magnetic resonance imaging.

shown in the plots in exhibit 11.12 later in the chapter. Each of 30 replications of the simulation is run for 200 hours. Replications are needed to determine confidence intervals for the reported values. Some of the output from this simulation is shown in exhibit 11.5. The sample mean plus or minus the half-width gives the 95 percent confidence interval for the mean. Increasing the number of replications reduces the half-width. The results of this simulation agree fairly closely with the calculated steady-state results because the process was assumed to run continuously for a significant period, 200 hours. A more realistic assumption might be that MRI procedures are only performed ten hours every day. The Arena simulation was rerun with this assumption, and the results are shown in exhibit 11.6. The average wait times, queue length, and utilization are lower than the steady-state values.

EXHIBIT 11.5
Arena Output for VVH MRI M/M/1 Queuing Example: 200 Hours

Category Overview

8:22:36 AM July 26, 2011

Values across all replications

MRI Example

Replications: 30 Time unit: Hours

Key Performance Indicators

System

	Average
Number out	601

Entity

Time

Wait Time	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Patient	0.7241	0.08	0.5009	1.3496	0.00	7.3900

Total Time	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Patient	0.9734	0.08	0.7427	1.6174	0.00001961	7.4140

Queue

Other

Number Waiting	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
MRI exam queue	2.1944	0.25	1.4326	4.2851	0.00	29.0000

Resource

Usage

Instantaneous Utilization	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
MRI	0.7488	0.01	0.6767	0.8513	0.00	1.0000

Arrival rate = 3 patients/hour; service rate = 4 patients/hour.

Note: Created with Arena simulation software. M = exponential distribution; MRI = magnetic resonance imaging.

EXHIBIT 11.6
Arena Output for VVH MRI M/M/1 Queuing Example: 10 Hours

Category Overview

12:19:03 PM July 26, 2011

Values across all replications

MRI Example

Replications: 30 Time unit: Hours

Key Performance Indicators

System

	Average
Number out	28

Entity

Time

Wait Time	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Patient	0.4778	0.15	0.02803444	1.4312	0.00	2.9818

Total Time	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Patient	0.7304	0.16	0.2407	1.7611	0.00082680	3.3129

Queue

Other

Number Waiting	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
MRI exam queue	1.5265	0.46	0.2219	4.5799	0.00	10.0000

Resource

Usage

Instantaneous Utilization	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
MRI	0.7167	0.05	0.4088	0.9780	0.00	1.0000

Arrival rate = 3 patients/hour; service rate = 4 patients/hour.

Note: Created with Arena simulation software. M = exponential distribution; MRI = magnetic resonance imaging.

Vincent Valley Hospital and Health System M/M/1 Queue

VVH has determined that a steady-state analysis is not appropriate for its situation because MRIs are only offered ten hours a day. The process improvement team assigned to this system decides to analyze the situation using simulation. Once the model is built and run, the model and simulation results are compared with actual data and evaluated by relevant staff to ensure that the model accurately reflects reality. All staff agree that the model is valid and can be used to determine how to achieve the stated goal. If the model had not been considered valid, the team would have needed to build and validate a new model.

The results of the simulation (refer to exhibit 11.14 later in the chapter) indicate that VVH has an average of 1.5 patients in the queue. To reach the desired goal of only one patient waiting on average, VVH needs to decrease the arrival rate or increase the service rate. Using trial and error in the simulation,

the organization finds that decreasing the arrival rate to 2.7 or increasing the service rate to 4.4 will allow the goal to be achieved.

However, even using the improvement tools in this text, the team believes that the organization will only be able to increase the service rate of the MRI to 4.2 patients per hour. Therefore, to reach the goal, the arrival rate must also be decreased. Again using the simulation, VVH finds that it needs to decrease the arrival rate to 2.8 patients per hour. Exhibit 11.7 shows the results of this simulation.

The team recommends that (1) a kaizen event be held for the MRI process to increase service rate and (2) appointments for the MRI be reduced to decrease the arrival rate. However, the team also notes that implementing these changes will reduce the average number of patients served from 28 to 26 and reduce the utilization of the MRI from 0.72 to 0.69. More positively, average patient wait time will be reduced from 0.48 hours to 0.35 hours.

EXHIBIT 11.7 Arena Output for VVH MRI M/M/1 Queuing Example: Decreased Arrival Rate, Increased Service Rate

Category Overview

8:24:44 AM July 26, 2011

Values across all replications

MRI Example

Replications: 30 Time unit: Hours

Key Performance Indicators

System

	Average
Number out	26

Entity

Wait Time	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Patient	0.3507	0.12	0.02449931	1.4202	0.00	3.4973

Total Time	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Patient	0.6008	0.14	0.1899	1.7825	0.00097591	4.2210

Queue

Other

Number Waiting	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
MRI exam queue	1.0342	0.36	0.0928	4.2272	0.00	9.0000

Resource

Usage

Instantaneous Utilization	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
MRI	0.6682	0.06	0.3314	0.9456	0.00	1.0000

Arrival rate = 2.8 patients/hour; service rate = 4.2 patients/hour; 10 hours simulated.

Note: Created with Arena simulation software. M = exponential distribution; MRI = magnetic resonance imaging.

VVH is able to increase the service rate to 4.2 patients per hour and decrease the arrival rate to 2.8 patients per hour, and the results are as predicted by the simulation. The team now begins to investigate other solutions enabling VVH to increase MRI utilization while maintaining wait times and queue length.

Simulation and Queuing Theory Findings

Simulation is a powerful tool for modeling processes and systems to evaluate choices and opportunities. As is true of all of the tools and techniques presented in this text, simulation can be used in conjunction with other initiatives, such as Lean or Six Sigma, to enable continuous improvement of systems and processes.

In a series of studies, queuing theory has been used to analyze flow of EDs and operating rooms (Butterfield 2007; McManus et al. 2004). In many instances, surgical suites more than doubled the number of surgeries they are able to complete in a short time. Because surgeries are a prime source of revenue and margin for most hospitals, this improvement makes the hospital more profitable.

Process Improvement in Practice

In this section, we review methods and tools that, in addition to simulation, are key approaches to process improvement, and we apply them to an emergency department scenario at VVH.

Review of Methodologies

Six Sigma

If the primary goal of a process improvement project is to improve quality (reduce the variability in outcomes), the Six Sigma approach and tools described in chapter 9 yield the best results. As discussed previously, Six Sigma uses seven basic tools: fishbone diagrams, check sheets, histograms, Pareto charts, flowcharts, scatter plots, and run charts. It also includes statistical process control to provide an ongoing measurement of process output characteristics to ensure quality and enable the identification of a problem situation before an error occurs.

The Six Sigma approach also includes measuring process capability—whether a process is capable of producing the desired output—and benchmarking it against other similar processes in other organizations. Quality function deployment is used to match customer requirements (voice of the customer) with process capabilities given that trade-offs must be made. Poka-yoke is employed selectively to mistake-proof parts of a process.

A primary function of Six Sigma programs is to eliminate sources of artificial variance in processes and systems. Natural variance occurs in any system, such as heat, temperature, and patients getting sick or breaking a leg. Artificial variance is created by the people in the system and is completely

in their control. Six Sigma programs identify and eliminate those sources of artificial variance. For example, scheduling systems, overtime allocations, and business office processing systems can all be changed by people in the system. The secret to a successful Six Sigma program is removing all the artificial variance and focusing on creating value for customers. Effective Six Sigma systems strategically employ Lean concepts to achieve this goal.

Lean

Process improvement projects focused on eliminating waste and improving flow in the system or process can use many of the tools that are part of the Lean approach (chapter 10). The kaizen philosophy, which is the basis for Lean, includes the following steps:

1. *Specify value*. Identify activities that provide value from the customer's perspective.
2. *Map and improve the value stream*. Determine the sequence of activities or the current state of the process and the desired future state. Eliminate non-value-added steps and other waste.
3. *Enable flow*. Allow the process to flow as smoothly and quickly as possible.
4. *Enable pull*. Allow the customer to pull products or services.
5. *Perfect*. Repeat the cycle to ensure a focus on continuous improvement.

An important part of Lean is value stream mapping, which is used to define the process and determine where waste is occurring. Takt time measures the time needed for the process to occur. It is based on customer demand and can be used to synchronize flow in a process. Standardized work, an important part of the Lean approach, is written documentation of the precise way in which every step in a process should be performed and helps ensure that activities are completed the same way every time in an efficient manner.

Other Lean tools include the five Ss (a technique to organize the workplace) and spaghetti diagrams (a mapping technique to show the movement of customers, patients, workers, equipment, jobs, etc.). Leveling workload (heijunka) so that the system or process flows without interruption can be used to improve the value stream. Kaizen blitzes or events are Lean tools used to improve the process quickly when project management is not needed (chapter 10).

Process Improvement Project: Vincent Valley Hospital and Health System Emergency Department

To demonstrate the power of many of the process improvement tools described in this book, an extensive patient flow process improvement project at VVH is examined.

VVH has identified patient flow in the ED as an important area on which to focus process improvement efforts. The goal of the project is to reduce total patient time in the ED (both waiting and care delivery) while maintaining or improving financial performance.

The first step for VVH leadership is to charter a multidepartmental team using the project management methods described in chapter 5. The head nurse for emergency services has been appointed project leader. The team feels VVH should take a number of steps to improve patient flow in the ED and splits the systems improvement project into three major phases. First, team members will perform simple data collection and basic process improvement to identify low-hanging fruit and make obvious, straightforward changes.

Once the team feels comfortable with its understanding of the basics of patient flow in the department, it will work to understand the elements of the system more fully by collecting detailed data. Then, value stream mapping and the theory of constraints will be used to identify opportunities for improvement. Root-cause analysis will be employed on poorly performing processes and tasks; resulting changes will be adopted and their effects measured.

The third phase of the project will be the use of simulation. Because the team, by this stage in the improvement effort, will have complete knowledge of patient flow in the system, it will be able to develop and test a simulation model with confidence. Once the simulation is validated, the team will continuously test process improvements in the simulation model and implement them in the ED.

The specific high-level tasks in this project are as follows.

Phase I

1. Observe patient flow and develop a detailed process map.
2. Measure high-level patient flow metrics for one week:
   - Patients arriving per hour
   - Patients departing per hour to inpatient
   - Patients departing per hour to home
   - Number of patients in the ED, including the waiting area and exam rooms
3. With the process map and data in hand, use simple process improvement techniques to make changes in the process, then measure the results.

Phase II

4. Set up a measurement system for each individual process, and take measurements over one week.

5. Use value stream mapping and the theory of constraints to analyze patient flow and make improvements, then measure the effects of the changes.

Phase III

6. Collect data needed to build a realistic simulation model.
7. Develop the simulation model and validate it against real data.
8. Use the simulation model to conduct virtual experiments on process improvements. Implement promising improvements, and measure the results of the changes.

Phase I

VVH process improvement project team members observe patient flow and record the needed data. With the information collected, the team creates a detailed process map. Team members measure the following high-level operating statistics related to patient flow:

- Patients arriving per hour = 10
- Patients departing per hour to inpatient = 2
- Patients triaged to routine emergency care per hour = 8
- Patients departing per hour to home = 8
- Average number of patients in various parts of the system (sampled every 10 minutes) = 20
- Average number of patients in ED exam rooms = 4

Using Little's law, the average time in the ED (throughput time) is calculated as

$$\begin{aligned} \text{Throughput time} &= T \\ &= I/\lambda \\ &= \frac{24 \text{ patients}}{8 \text{ patients/hour}} \\ &= 3 \text{ hours.} \end{aligned}$$

Hence, each patient spends an average of 3 hours, or 180 minutes, in the ED.

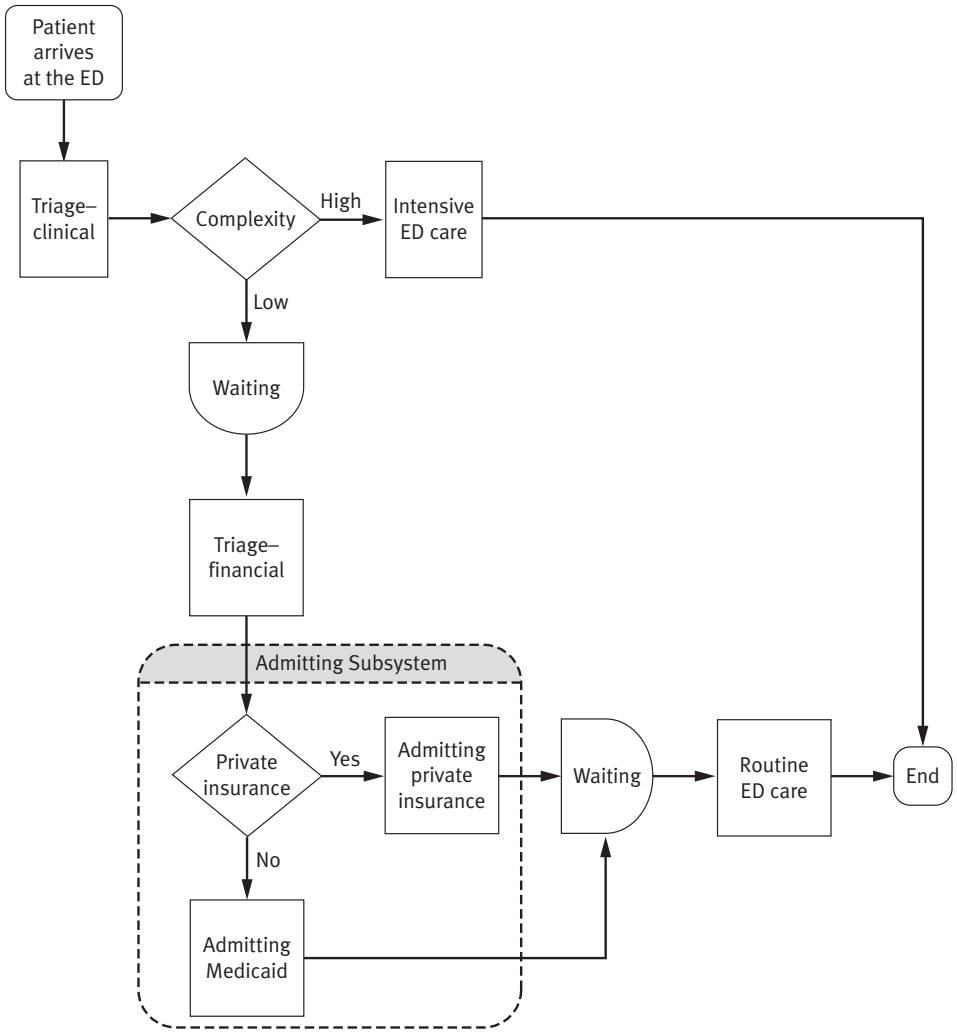
However, Little's law only gives the average time in the department at steady state. Therefore, the team measures total time in the system for a sample of routine patients and determines an average of 165 minutes. It also observes that the number of patients in the waiting room varies from 0 to 20 and the actual time to move through the process varies from one hour to more than five hours.

Initially, the team focuses on the ED admitting subsystem as an opportunity for immediate improvement. Exhibit 11.8 shows the complete ED system, with the admitting subsystem highlighted.

The team develops the following description of the admitting process from its documentation of patient flow:

> Patients who did not have an acute clinical problem were asked if they had health insurance. If they did not have health insurance, they were sent to the admitting clerk who specializes in Medicaid (to enroll them in a Medicaid program). If they had health insurance, they were sent to the other clerk, who specializes in private insurance. If a patient had been sent to the wrong clerk by triage, he was sent to the other clerk.

EXHIBIT 11.8
VVH Emergency Department (ED) Admitting Subsystem



Note: Created with Microsoft Visio.

The team determines that one process improvement change could be to cross-train the admitting clerks on both private insurance and Medicaid eligibility. This training would provide for load balancing, as patients would automatically go to the free clerk. In addition, this system improvement would eliminate triage staff errors in sending patients to the wrong clerk, hence providing quality at the source.

Phase II

Phase I produced some gains in reducing patient time in the ED. However, the team feels more detailed data are needed to improve further. As a first step in collecting these data, the team measures various parameters of the department's processes. Initially, it focuses on the period from 2:00 p.m. to 2:00 a.m., Monday through Thursday, as this is the busy period in the ED and demand seems relatively stable during these times.

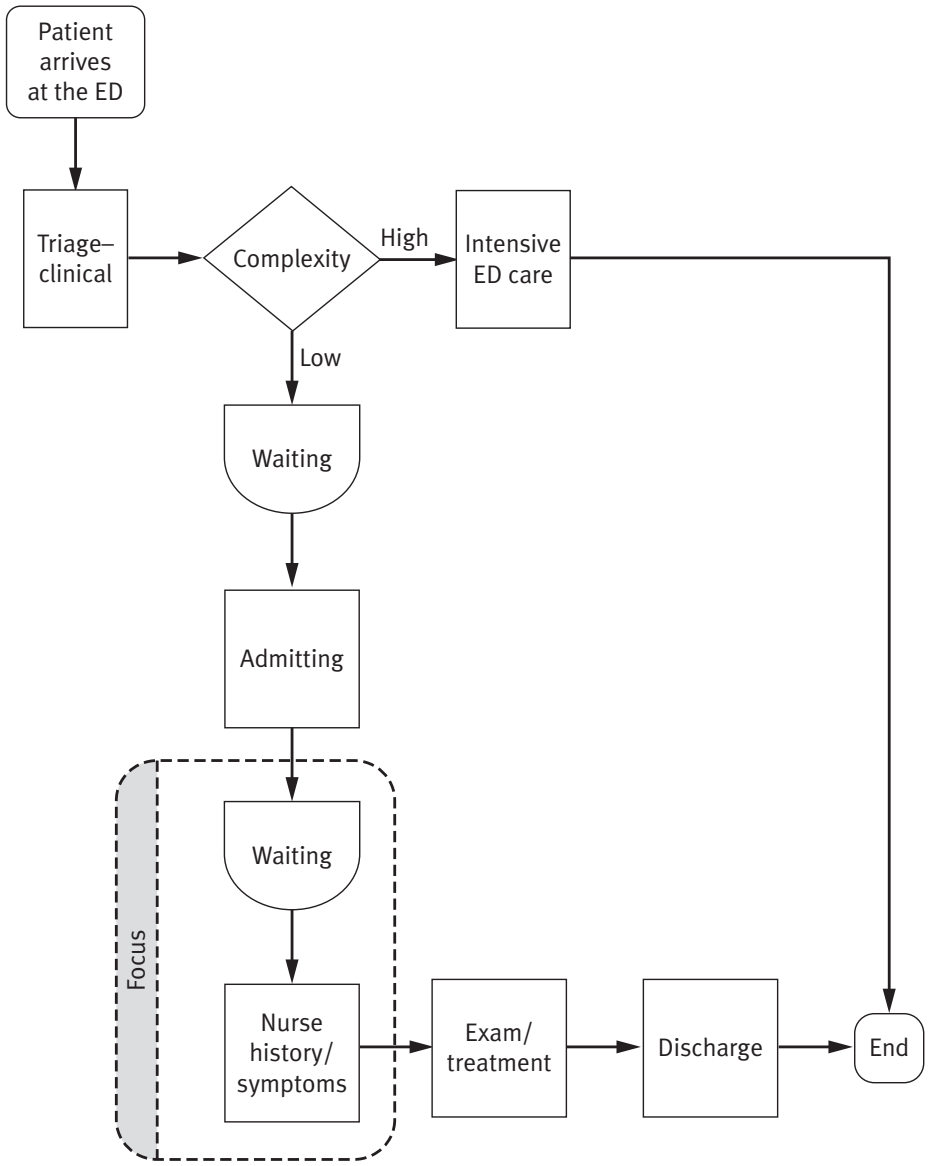
The team draws a more detailed process map (exhibit 11.9) and performs value stream mapping of this process (exhibit 11.10). First, team members evaluate each step in the process to determine if it is value-added, non-value-added, or non-value-added but necessary. Then, they measure the time a patient spends at each step in the process. The team finds that after a patient has given his insurance information, he spends an average of 30 minutes of non-value-added time in the waiting room before a nurse is available to take his history and record the presenting complaint, a process that takes an average of 20 minutes to complete. The percentage of value-added time for these two steps is

$$(\text{Value-added time} \div \text{Total time}) \times 100 = [20 \text{ minutes} \div (30 \text{ minutes} + 20 \text{ minutes})] \times 100 = 40\%.$$

The team believes the waiting room process can be improved through automation. Patients are handed a tablet personal computer in the waiting area and asked to enter their symptoms and history via a series of branched questions. The results are sent via a wireless network to VVH's electronic health record (EHR). This step takes patients an average of 20 minutes to complete. Staff know which patients have completed the electronic interview by checking the EHR and can prioritize which patient is to be seen next. This new procedure also reduces the time the nurse spends with the patient to 10 minutes because it enables the nurse to verify, rather than record, presenting symptoms and patient history. The percentage of value-added time for the new procedure is

$$\begin{aligned}&(\text{Value-added time} \div \text{Total time}) \times 100\\ &= [(\text{Patient history time} + \text{Nurse history time}) \div (\text{Patient history time} + \text{Wait time} + \text{Nurse history time})] \times 100\\ &= [(20 \text{ minutes} + 10 \text{ minutes}) \div (20 \text{ minutes} + 10 \text{ minutes} + 10 \text{ minutes})] \times 100\\ &= 75\%.\end{aligned}$$

EXHIBIT 11.9
VVH Emergency Department (ED) Process Map: Focus on Waiting and History



Note: Created with Microsoft Visio.

The average throughput time for a patient in the ED is reduced by 10 minutes. The average time for patients to flow through the department (throughput time) prior to this improvement was 155 minutes. Because this step is on the critical path of the complete routine care ED process, throughput time for noncomplex patients is reduced to 145 minutes, a 7 percent productivity gain. An analyst from the VVH finance department (a member of the project team) is able to demonstrate that the capital and software costs for the

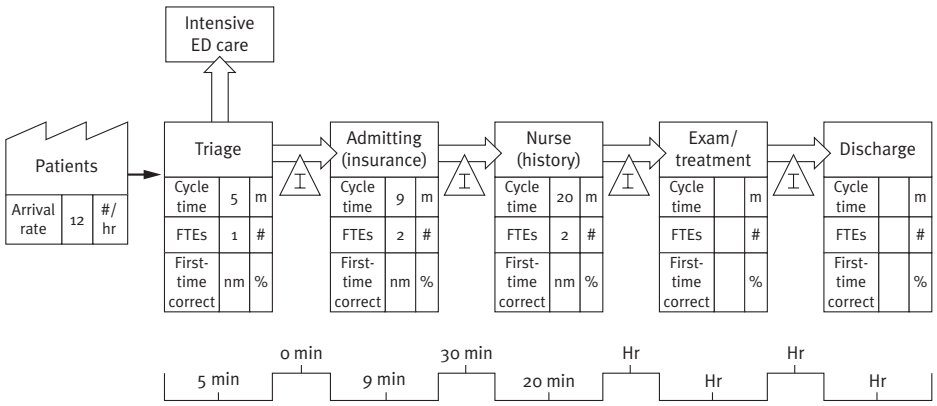


EXHIBIT 11.10 VVH Emergency Department (ED) Value Stream Map: Focus on Waiting and History

Note: Created with eVSM software, a Microsoft Visio add-on from GumshoeKI, Inc. FTE = full-time equivalent; nm = number of patients in this step of the process.

tablet computers will be recovered within 12 months by the improvement in patient flow.

This phase of the project used three of the basic process improvement tools discussed in this chapter:

- Have the customer (patient) do it.
- Provide quality at the source.
- Gain information feedback and real-time control.

Although the process improvements already undertaken have had a visible impact on flow in the ED, the team believes more improvements are possible. Bottlenecks plague the process, as evidenced by two waiting lines, or queues: (1) the waiting room queue, where patients wait before being moved to an exam room, and (2) the most visible queue for routine patients, the discharge area, where patients occasionally must stand because all of the area's chairs are occupied. In the discharge area, patients wait a significant amount of time for final instructions and prescriptions.

The theory of constraints suggests that the bottleneck be identified and optimized. However, alleviating or eliminating the patient examination and treatment or discharge bottlenecks would require significant changes in a long-standing process. Because this process improvement step seems to have the probability of a high payoff but would be a significant departure from existing practice, the team moves to phase III of the project and uses simulation to model different options to improve patient flow in the examination/treatment and discharge processes.

Phase III

First, the team reviews the basic terminology of simulation.

- An *entity* is what flows through a system. Here, the entity is the patient. However, in other systems, the entity can be materials (e.g., blood sample, drug) or information (e.g., diagnosis, billing code). Entities usually have attributes that affect their flow through the system (e.g., male/female, acute/chronic condition).
- Each individual *process* in the system transforms (adds value to) the entity being processed. Each process takes time and consumes resources, such as staff, equipment, supplies, and information.
- *Time* and *resource* use can be defined as an exact value (e.g., ten minutes) or a probability distribution (e.g., normal—mean, standard deviation). Most healthcare tasks and processes do not require the same amount of time each time they are performed—they require a variable amount of time. These variable usage rates are best described as probability distributions. (Chapter 7 discusses probability distributions in detail.)
- The geographic location of a process is called a *station*. Entities flow from one process to the next via *routes*. The routes can branch out on the basis of *decision points* in the process map.
- Finally, because a process may not be able to handle all incoming entities in a timely fashion, *queues* occur at each process and can be measured and modeled.

The team next develops a process map and simulation model for routine patient flow (exhibit 11.11) in the ED using Arena simulation software (see the companion website for links to videos detailing this model and its operation). The team focuses on routine patients rather than those requiring intensive emergency care because of the high proportion of routine patients seen in the department. Routine patients are checked in and their self-recorded history and presenting complaint(s) verified by a nurse. Then, patients move to an exam/treatment room and, finally, to the discharge area. Of the ten patients who arrive at the ED per hour, eight follow this process.


On the web at
ache.org/books/OpsManagement3

Next, to build a simulation model that accurately reflects this process, the team needs to determine the probability distributions of treatment time, admitting time, nurse history time, discharge time, and arrival rate for routine patients. To determine these probability distributions, team members collect data on time of arrival in the department and time to perform each step in the routine patient care process.

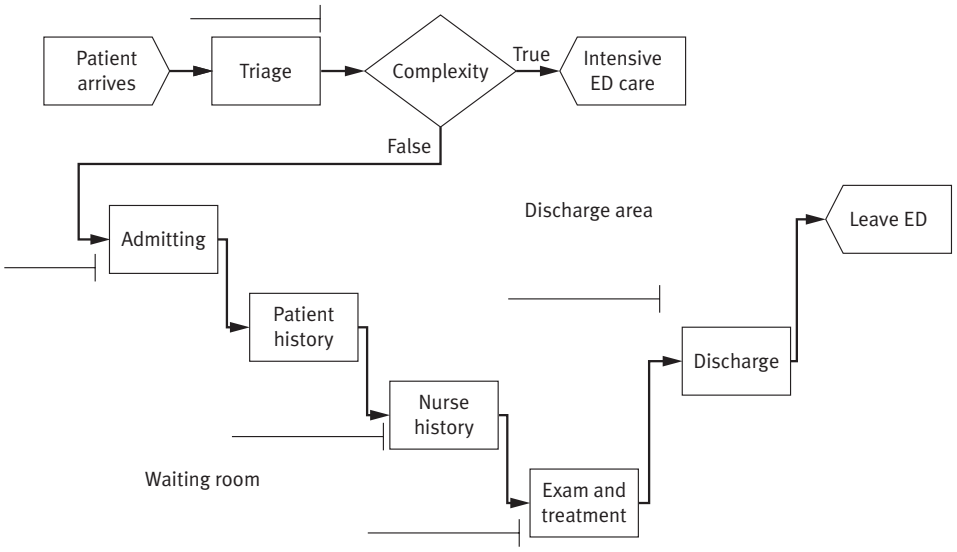


EXHIBIT 11.11 VVH Emergency Department (ED) Initial State Simulation Model

Note: Created with Arena simulation software.

Probability distributions are determined using the input analyzer function in Arena. Input Analyzer takes raw input data and finds the best-fitting probability distribution for them. Exhibit 11.12 shows the output of Input Analyzer for 500 observations of treatment time for ED patients requiring routine care. Input Analyzer suggests that the best-fitting probability distribution for these data is triangular, with a minimum of 9 minutes, mode of 33 minutes, and maximum of 51 minutes.

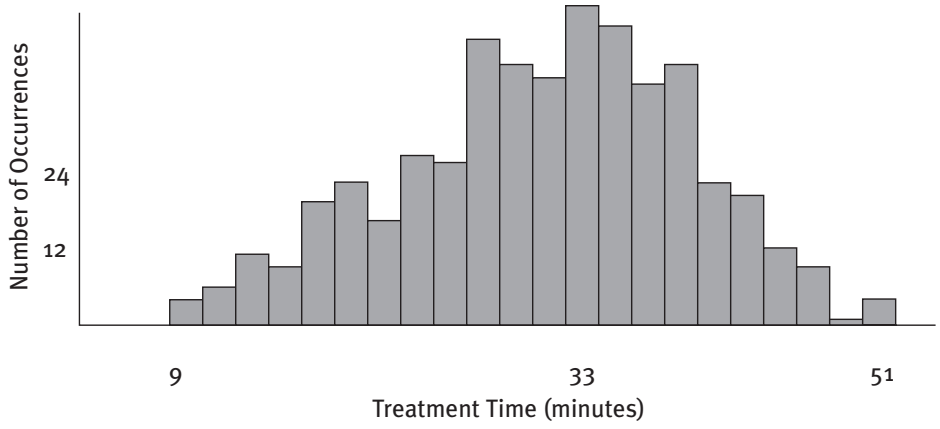


EXHIBIT 11.12 Examination and Treatment Time Probability Distribution: Routine Emergency Department Patients

The remaining data are analyzed in the same manner, and the following best-fitting probability distributions are determined:

- Emergency routine patient arrival rate—exponential (7.5 minutes between arrivals)
- Triage time—triangular (2, 5, 7 minutes)
- Admitting time—triangular (3, 8, 15 minutes)
- Patient history time—triangular (15, 20, 25 minutes)
- Nurse history time—triangular (5, 11, 15 minutes)
- Exam/treatment time—triangular (14, 36, 56 minutes)
- Discharge time—triangular (9, 19, 32 minutes)

The Arena model simulation is based on 12-hour intervals (2:00 p.m. to 2:00 a.m.) and replicated 100 times. Note that increasing the number of replications decreases the half-width and, therefore, gives tighter confidence intervals. The number of replications needed depends on the desired confidence interval for the outcome variables. However, as the model becomes more complicated, more replications take more simulation time; this model is fairly simple, so 100 replications take little time and are sufficient for this purpose.

Most simulation software, including Arena, is capable of using different arrival rate probability distributions for different times of the day and days of the week, allowing for varying demand patterns. However, the team believes that this simple model using only one arrival rate probability distribution represents the busiest time for the ED, having observed that by 2:00 p.m. on weekdays no queues are created in either the waiting room or the discharge area.

The results of the simulation are reviewed by the team and compared with actual data and observations to ensure that the model is, in fact, simulating the reality of the ED. The team is satisfied that the model accurately reflects reality.

The focus of this simulation is the queuing that occurs in both the waiting room and the discharge area and the total time in the system. Exhibit 11.13 shows the results of this base (current status) model. On average, a patient spends 2.4 hours in the ED.

The team next examines the discharge process in depth because patient waiting time is greatest there. The ED has two rooms devoted to discharge and uses two nurses to handle all discharge tasks, such as making sure prescriptions are given and home care instructions are understood. However, because of the limited number of nurses and exam rooms, queuing is inevitable. In addition, the patient treatment information must be handed off from the treatment team to the discharge nurse. The process improvement team simulates having the discharge process carried out by the examination and treatment team. Because the examination and treatment team knows the patient information, the handoff task can be eliminated. The team estimates that this change will save about five

EXHIBIT 11.13 VVH Emergency Department Initial State Simulation Model Output

Replications: 100 Time Unit: Hours

Total Time	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Routine patient	2.4207	0.08	1.7953	3.4082	1.2004	5.2448

Queue

Time

Waiting Time	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Admitting queue	0.00526930	0.00	0.00048553	0.01668610	0.00	0.2235
Discharge queue	0.3972	0.26	0.06416692	0.8865	0.00	2.0531
Exam and treatment queue	0.3382	0.38	0.04167122	1.1956	0.00	2.5777
Nurse history queue	0.01764541	0.01	0.00272715	0.05309733	0.00	0.3694
Triage queue	0.06437939	0.05	0.01703829	0.1402	0.00	0.6506

Other

Waiting Time	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Admitting queue	0.03458032	0.00	0.00267040	0.1001	0.00	2.0000
Discharge queue	2.2481	0.26	0.2888	5.1713	0.00	13.0000
Exam and treatment queue	2.1930	0.38	0.2062	9.4408	0.00	22.0000
Nurse history queue	0.1136	0.01	0.01298461	0.4069	0.00	5.0000
Triage queue	0.5394	0.05	0.1145	1.7216	0.00	10.0000

Resource

Usage

Instantaneous Utilization	Average	Half-Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Discharge nurse 1	0.8285	0.01	0.6715	0.8972	0.00	1.0000
Discharge nurse 2	0.8360	0.01	0.6673	0.9105	0.00	1.0000
Exam room 1	0.8441	0.01	0.6253	0.9497	0.00	1.0000
Exam room 2	0.8329	0.01	0.6548	0.9297	0.00	1.0000
Exam room 3	0.8182	0.02	0.5358	0.9200	0.00	1.0000
Exam room 4	0.8075	0.02	0.6135	0.9156	0.00	1.0000
Financial clerk 1	0.4615	0.01	0.3320	0.5636	0.00	1.0000
Financial clerk 2	0.4580	0.01	0.3286	0.5823	0.00	1.0000
History nurse 1	0.5294	0.01	0.3886	0.6796	0.00	1.0000
History nurse 2	0.5240	0.01	0.3937	0.7107	0.00	1.0000
Triage nurse	0.6267	0.01	0.4861	0.8373	0.00	1.0000

Note: Created with Arena simulation software.

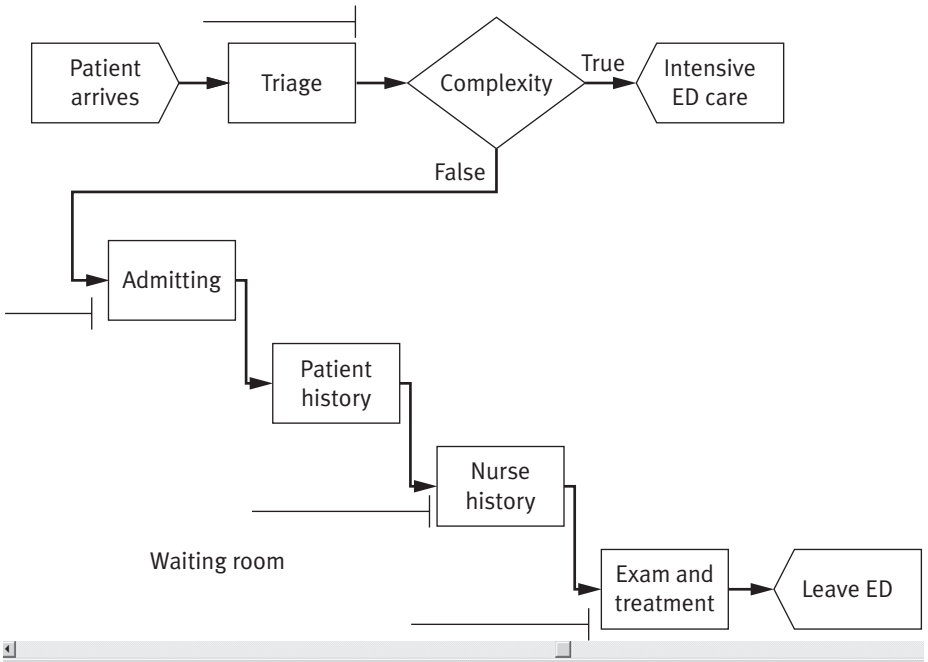
minutes. To ensure that this is the correct outcome, team members simulate the new system by eliminating discharge as a separate process.

Team members estimate the probability distribution of the combined exam/treatment/discharge task by first estimating the probability distribution for handoff as triangular (4, 5, 7 minutes). The team uses Input Analyzer to simulate 1,000 observations of exam/treatment time, discharge time, and handoff time using the previously determined probability distributions for each. For each observation, it adds exam/treatment time to discharge time and subtracts handoff time to find total time. Input Analyzer finds the best-fitting

probability distribution for the total time for the new process as triangular (18, 50, 82 minutes).

The team simulates the new process and finds that, under the new system, patients will spend an average of 2.95 hours in the ED—increasing the time spent there. However, it will eliminate the need for discharge rooms. The team decides to investigate the impact of converting the former discharge rooms to exam rooms and runs a new simulation incorporating this change (exhibit 11.14). The result of this simulation is shown in exhibit 11.15. Both the number of patients in the waiting room (examination and treatment queue) and the amount of time they wait are reduced substantially. The staffing levels are not changed, as the discharge nurses are now treatment nurses. Physician staffing also is not increased, as some delay inside the treatment process itself has always existed due to the need to wait for lab results, resulting in a delayed final physician diagnosis. Having more patients available for treatment fills this lab delay time for physicians to perform patient care.

EXHIBIT 11.14
VVH Emergency Department (ED) Proposed Change Simulation Model



Process - Basic Process

	Name	Type	Action	Priority	Resources	Delay Type	Units	Allocation	Minimum	Value	Maximu
1	Admitting	Standard	Seize Delay Release	Medium(2)	1 rows	Triangular	Minutes	Value Added	3	8	15
2	Nurse History	Standard	Seize Delay Release	Medium(2)	1 rows	Triangular	Minutes	Value Added	5	11	15
3	Exam Treatment and Discharge	Standard	Seize Delay Release	Medium(2)	1 rows	Triangular	Minutes	Value Added	18	50	82
4	Triage	Standard	Seize Delay Release	Medium(2)	1 rows	Triangular	Minutes	Value Added	2	5	7
5	Patient History	Standard	Delay	Medium(2)	0 rows	Triangular	Minutes	Value Added	15	20	25

Note: Created with Arena simulation software.

EXHIBIT 11.15
VVH Emergency Department (ED) Proposed Change Simulation Model Output

Category Overview

4:11:35 PM February 8, 2012

Values Across All Replications

VVH Emergency

Replications: 100 Time Unit: Hours

Entity

Time

Total Time	Average	Half Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Routine patient	1.8376	0.05	1.5459	2.8729	1.0063	4.5989

Queue

Time

Waiting Time	Average	Half Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Admitting queue	0.00519434	0.00	0.00041085	0.01364095	0.00	0.2235
Exam and treatment and discharge queue	0.2039	0.04	0.00197293	1.1105	0.00	2.2943
Nurse history queue	0.01791752	0.00	0.00244500	0.07537764	0.00	0.3417
Triage queue	0.06635691	0.01	0.01863876	0.2547	0.00	0.8065

Other

Waiting Time	Average	Half Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Admitting queue	0.03400433	0.00	0.00218978	0.0946	0.00	3.0000
Exam and treatment and discharge queue	1.3571	0.31	0.00838496	7.8288	0.00	19.0000
Nurse history queue	0.1098	0.01	0.01120623	0.5716	0.00	4.0000
Triage queue	0.5629	0.08	0.1227	2.5046	0.00	11.0000

Resource

Usage

Instantaneous Utilization	Average	Half Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Exam room 1	0.7827	0.02	0.5405	0.9303	0.00	1.0000
Exam room 2	0.7644	0.02	0.5468	0.9103	0.00	1.0000
Exam room 3	0.7626	0.02	0.5577	0.9052	0.00	1.0000
Exam room 4	0.7478	0.02	0.4984	0.8993	0.00	1.0000
Exam room 5	0.7859	0.02	0.5420	0.9313	0.00	1.0000
Exam room 6	0.8030	0.02	0.4990	0.9472	0.00	1.0000
Financial clerk 1	0.4606	0.01	0.3250	0.5985	0.00	1.0000
Financial clerk 2	0.4529	0.01	0.2968	0.6119	0.00	1.0000
History nurse 1	0.5236	0.01	0.3642	0.6766	0.00	1.0000
History nurse 2	0.5154	0.01	0.3403	0.6982	0.00	1.0000
Triage nurse	0.6226	0.02	0.4742	0.8185	0.00	1.0000

Note: Created with Arena simulation software.

EXHIBIT 11.16
Summary of VVH Emergency Department Throughput Improvement Project

Process Improvement Change	Throughput Time, Routine Patients
Baseline, before any improvement	165 minutes
Combine admitting functions	155 minutes
Patients enter their own history into computer	145 minutes
Combine discharge tasks into examination and treatment process, and convert discharge rooms to treatment rooms	110 minutes

The most significant improvement resulting from the process improvement initiative is that total patient throughput time now averages 1.84 hours (110 minutes). This 33 percent reduction in throughput time exceeds the team's goal and is celebrated by VVH's senior leadership. The summary of process improvement steps is displayed in exhibit 11.16.

Conclusion

The theory of swift, even flow provides a framework for process improvement and increased productivity. The efficiency and effectiveness of a process increase as the speed of flow through the process increases and the variability associated with that process decreases.

The movement of patients in a healthcare facility is one of the most critical and visible processes in healthcare delivery. Reducing flow time and variation in processes results in a number of benefits, including the following:

- Patient satisfaction increases.
- Quality of clinical care improves as patients have reduced waits for diagnosis and treatment.
- Financial performance improves.

This chapter demonstrates many approaches to the challenges of reducing flow time and process variation. Starting with the straightforward process map, many improvements can be found immediately by inspection. In other cases, the powerful tool of computer-based discrete event simulation can provide a road map to sophisticated process improvements.

Ensuring quality of care is another critical focus of healthcare organizations. The process improvement tools and approaches in this chapter may be

used to reduce process variation and eliminate errors. Healthcare organizations must employ the disciplined approach described in this chapter to achieve the needed improvements in flow and quality.

Discussion Questions

1. How do you determine which process improvement tools should be used in a given situation? What is the cost and return of each approach?
2. Which process improvement tool can have the most powerful impact, and why?
3. How can barriers to process improvement, such as staff reluctance to change, lack of capital, technological barriers, or clinical practice guidelines, be overcome?
4. How can the electronic health record be used to make significant process improvements for both efficiency and quality increases?
5. Describe several places or times in your organization where people or objects (paperwork, tests, etc.) wait in line. How do the characteristics of each example differ?

Exercises

1. Access the National Guideline Clearinghouse (www.guideline.gov/) and translate one of the guidelines described into a process map. Add decision points and alternative paths to account for unusual issues that might occur in the process. (Hint: Use Microsoft Visio or another similar application to complete this exercise.)
2. Access the following process maps on the companion website:
   - Operating Suite
   - Cancer Treatment Clinic

   Use basic improvement tools, theory of constraints, Six Sigma, or Lean tools to determine possible process improvements.

   On the web at
   ache.org/books/OpsManagement3

3. The hematology lab manager has received complaints that the turnaround time for blood tests is too long. Data from the past month show that the arrival rate of blood samples to one technician in the lab is five per hour and the service rate is six per hour. Using queuing theory, and assuming that (a) both rates are exponentially distributed and (b) the lab is at steady state, determine the following measures:

- Capacity utilization of the lab
- Average number of blood samples in the lab
- Average time that a sample waits in the queue
- Average number of blood samples waiting for testing
- Average time that a blood sample spends in the lab

References

Butterfield, S. 2007. "A New Rx for Crowded Hospitals: Math." *ACP Hospitalist*. Published December. www.acphospitalist.org/archives/2007/12/math.htm#sb1.

Clark, J. J. 2005. "Unlocking Hospital Gridlock." *Healthcare Financial Management* 59 (11): 94–104.

Cooper, R. B. 1981. *Introduction to Queueing Theory,* 2nd edition. New York: North-Holland.

Deming, W. E. 1998. "The Deming Philosophy." Deming-Network. Accessed June 9, 2006. http://deming.ces.clemson.edu/pub/den/deming_philosophy.htm.

Devaraj, S., T. T. Ow, and R. Kohli. 2013. "Examining the Impact of Information Technology and Patient Flow on Healthcare Performance: A Theory of Swift and Even Flow (TSEF) Perspective." *Journal of Operations Management* 31 (4): 181–92.

Litvak, E. 2003. "Managing Patient Flow: Smoothing OR Schedule Can Ease Capacity Crunches, Researchers Say." *OR Manager* 19 (November): 1, 9–10.

McManus, M., M. Long, A. Cooper, and E. Litvak. 2004. "Queuing Theory Accurately Models the Need for Critical Care Resources." *Anesthesiology* 100 (5): 1271–76.

Rockwell Automation. 2016. Arena home page. Accessed September 21. www.arenasimulation.com/.

Rodi, S. W., M. V. Grau, and C. M. Orsini. 2006. "Evaluation of a Fast Track Unit: Alignment of Resources and Demand Results in Improved Satisfaction and Decreased Length of Stay for Emergency Department Patients." *Quality Management in Healthcare* 15 (3): 163–70.

Sayah, A., M. Lai-Becker, L. Kingsley-Rocker, T. Scott-Long, K. O'Connor, and L. F. Lobon. 2016. "Emergency Department Expansion Versus Patient Flow Improvement: Impact on Patient Experience of Care." *Journal of Emergency Medicine* 50 (2): 339–48.

Schmenner, R. W. 2004. "Service Businesses and Productivity." *Decision Sciences* 35 (3): 333–47.

———. 2001. "Looking Ahead by Looking Back: Swift, Even Flow in the History of Manufacturing." *Production and Operations Management* 10 (1): 87–96.

Schmenner, R. W., and M. L. Swink. 1998. "On Theory in Operations Management." *Journal of Operations Management* 17 (1): 97–113.

Simul8 Corporation. 2016. "Process Simulation Software." Accessed September 21. www.simul8.com/.

Further Reading

Goldratt, E. M., and J. Cox. 1986. *The Goal: A Process of Ongoing Improvement*. New York: North River Press.

Kelton, W., R. Sadowski, and N. Swets. 2009. *Simulation with Arena*. New York: McGraw-Hill.