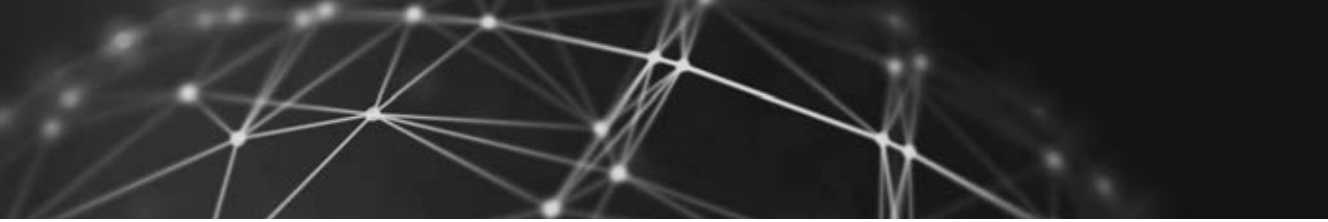

CHAPTER 4

Source: Lloyd. R. *Quality Health Care: A Guide to Developing and Using Indicators* 2nd Edition, Jones & Bartlett Learning, 2019.

Milestones in the Quality Measurement Journey

Listening to the voice of the customer (VOC) provides the starting point. Once you understand the wants, needs, and expectations of your internal and external customers, which are usually expressed as concepts (e.g., "I want better health service," "Why don't you have shorter waiting time?," or "Communication between the staff needs to get better"), it is up to you to translate these concepts into indicators that can be measured and tracked to determine whether your processes are capable of meeting the VOC expectations. Unfortunately, in health care, it is often the case that indicators are selected not because the providers of a service actually took time to listen to the VOC, but rather because (1) they made a priori decisions that they know what is best for the customers, (2) they have been given measures by external oversight or regulatory bodies that require certain measures be submitted to them, or (3) they take the measurement journey shortcut by selecting indicators that they have "always collected" and assume that these are good enough for the purposes at hand. This chapter has been designed to provide you with a roadmap for selecting and building indicators that will help you move from concepts to quantifiable measures and connect the VOC with the voice of the process (VOP).

Developing a Measurement Philosophy

The search for a few good indicators begins by having a clear understanding of why you are engaged in measuring performance in the first place. Historically, healthcare providers have collected and analyzed data strictly for internal purposes that were directed at improving clinical and operational effectiveness and efficiency. Over the years, however, the growing external demand for data has shifted much of the focus away from an internal need to understand the effectiveness and efficiency of the organization's processes to one of addressing external demands that lead to judgment. In other words, the business community, regulatory bodies, government officials, the media,

and consumers are all interested in answering a very simple question: "Which provider is the best?"

In an effort to answer this question, many initiatives, projects, and pieces of legislation have been developed over the years. The goal of these efforts has been to develop "report cards" or "score cards" on healthcare providers that can be used by various groups and consumers to make decisions about their healthcare choices. Regardless of the country or the approach to funding health care, however, there seems to be no quick or easy answer to the simple question of "Which provider is the best?"

What has been the typical response to this question is that external groups voluntarily ask for or mandate certain performance indicators from providers. These numbers are then combined with those from other providers, risk adjustments may be applied to the data to account for severity of the patient populations, and finally, reports are released to the public. These releases usually stimulate the following chain of events:

- Local and/or national media become interested.
- Investigative reporters are sent out to discover why *Your Hospital* has a higher coronary artery bypass graph (CABG) mortality percentage than *My Hospital* and why both are higher than the average for the county, region, province, or country in which they are located.
- The reporters present their findings in the next day's newspaper or on the 6 o'clock news, which usually focuses on the providers at the top and bottom of the list.
- *Your Hospital* and *My Hospital* convene internal meetings to develop strategies (rationales) for countering why their numbers are higher or lower than the average.
- Consumers become confused and/or cynical because the data do not necessarily reflect their experiences (e.g., "My father went to *Your Hospital* for his heart operation and everything was fine" or "My father went to *Your Hospital* and nearly died").

Whatever your view on the public release of data, it is quite obvious that the demand for data on provider performance and greater transparency will increase over the coming years. The simple question is "Are you prepared for it?" Healthcare organizations that have a measurement strategy and a proactive plan for investigating their own results will be in a much better position to deal with external scrutiny than those that sit back and hope that the local or national news service does not show up outside their hospital. (Refer to Chapter 2 for more on provider performance and transparency.)

Even though there is a renewed interest in the public release of provider data, the more important reason for knowing your data better than anyone else is that it is the right thing to do and it makes business sense. The complexity of today's healthcare delivery system requires that leaders have a clear understanding of their processes and the related outcomes. In order to meet operational and financial objectives, patient safety goals, and customer service expectations, healthcare providers should consider developing what Caldwell (1995) refers to as a "strategic measurement deployment matrix." Such a matrix combines strategic vision with tactical measures. It allows an organization to determine if the things they are working on are really connected to what the organization is supposed to be achieving.

The first step, therefore, on the quality measurement journey (QMJ) is achieved by having some sense of why you are measuring and your approach to measurement. Is measurement a part of the organization's day-to-day functioning? Or is it something that is done periodically in order to prepare reports for board meetings or external oversight bodies? A good place to start is to develop the organization's measurement philosophy and share it with staff, patients, and caregivers. This does not need to be a long document. Something as simple as this could serve as a starting point:

> Responsible leadership demands that we know our data better than anyone else. It further requires that we have processes in place to accurately and consistently obtain a balanced set of measures that monitor clinical outcomes, functional

status, customer satisfaction, process effectiveness, and resource utilization. Finally, we are committed to using these data to develop improvement strategies and then take **ACTION** to make these strategies a reality.

An organization needs to have a serious dialogue about its measurement philosophy and why it is measuring (i.e., for improvement, for judgment, or for research). Included in this ongoing dialogue should be specific discussions about the role of indicators and how they will (and will not) be used. Without a measurement philosophy, your efforts to identify key indicators and collect and analyze data will be nothing more than a random walk.

Ideally, indicators should be designed to improve quality by:

- Moving us away from anecdotes and focusing on objective data
- Enhancing our understanding of the variation that exists in a process
- Monitoring a process over time
- Seeing the effects of changes made to a process
- Providing a common frame of reference
- Providing a more accurate basis for prediction

Unfortunately, many organizations run into serious roadblocks when they attempt to select indicators and use them to improve quality.

Measurement Roadblocks

Many things impede good measurement practice. Based on my 40 plus years of working in the quality measurement arena, I believe there are five major roadblocks that people usually encounter in their QMJ:

Roadblock #1: Measurement Is Threatening

This is probably the largest roadblock we face with healthcare measurement. There are many examples of how data have been used both internally and externally to (figuratively) "beat people up." We often hear coworkers say that they did not want to take the monthly numbers to the boss because he or she "won't like these." Organizations have long memories when it comes to the use of data. Seasoned employees quickly tell new workers what happens when the numbers do not meet management's expectations. Quickly, the new workers hear the story about how Gwenn, nurse manager of 3 East, did not get her patient satisfaction scores up by the end of the year and now Gwenn is no longer with the organization. What the new workers didn't know, however, is that Gwenn left because her husband was transferred to another city. But her leaving and the decline of her unit's patient satisfaction scores do provide the basis for a compelling yet causally incorrect story. As time passes, this story becomes legendary, gets embellished a little, and becomes part of the organization's folklore. "Remember what happened to Gwenn" becomes the standard response whenever someone's patient satisfaction scores are below the expected targets.

What I find absolutely fascinating, however, is the fact that people actually like to measure things, including their own performance. There seems to be a natural curiosity in human beings about measurement. When my daughter Devon was 9 years old, for example, she loved to measure things. When I was in the garage one day doing a project she came up to me and said, "Measure me, Daddy." I took my tape measure and proceeded to measure her height. She acknowledged the measurement and went about her business. Ten minutes later, she returned and stated, "Measure me, Daddy." I said, "Devon, I don't think you have grown much in the last 10 minutes." But she insisted and seemed to find the actual act of measurement not only enlightening but also entertaining. The next time I observed her in the garage, she was using the tape to measure her bike, her doll, and the dog (or at least trying to measure the dog). Even adults love to measure what they do. I have a number of friends who participate in triathlons. They are very meticulous about measuring and monitoring their training

regimens. I have seen similar behavior from people involved with bowling, cycling, and golf.

How do we drive what seems to be an almost natural curiosity about measurement out of people when they get into work situations? The answer to me seems rather simple. Organizations frequently use data to instill a sense of fear in the employees. Once data are used for judgment and fear then the data are not for learning and improvement but rather for intimidation and control. It is not surprising, therefore, that the workers rapidly conclude, "Why should I participate in a measurement system that will be used against me?" Several years ago I experienced this attitude when I was facilitating a team that was attempting to reduce call button response time. During a meeting that was intended to identify a measurement plan, one team member blurted out, "Why don't you go measure 4 West? I know they are worse they we are." When measurement becomes threatening, the workers will conclude that measurement should be for someone else, not for them. The truth of the matter is that the primary audience for measurement is the manager of the department or unit and the workers. These are the people who own the process and who should be responsible for its performance. If the organization does not have a philosophy of measurement and a set of related tactics for deploying measurement throughout the organization, then measurement will generally become a threat. A strategic focus on measurement as described by Caldwell (1995) will do wonders to overcome this roadblock.

Roadblock #2: The Desire for Precision

Health care is not classified as a science. The federal government actually classifies healthcare jobs as service jobs, along with car repair, lawn service, and beauty shops. Sure, we use science and technology to accomplish what we do, but by and large health care is considered a service. It is interesting, therefore, that many people in our profession use the illusion of precision as a convenient excuse for not measuring. I have heard, for example, the following responses many times when I asked a team if they had finished their measurement plan:

- "We think it will take a little longer to make sure the survey is right."
- "The log sheet does not seem to capture all the elements we think we need to collect."
- "Why don't you check with us in a couple of weeks? We might have a better plan in place at that time."

The key point is that quality measurement does not have to be as precise as many people seem to think. We are not conducting research to win the Nobel Prize in physiology or medicine. We are trying to understand the variation that lives within our processes in order to make things more effective and more efficient for those we serve. Therefore, the concept of measurement that is "good enough" needs to be our guiding principle. The basic purpose of quality measurement is to inform the team or organization about its general direction and whether it is moving toward its goals and objectives. You do not need p-values at the 0.01 or 0.05 level of significance to tell you this. As one chief executive officer (CEO) told me, "If it passes the sniff test, that's good enough for me." Furthermore, we are not trying to conduct research that is designed around the randomized control trial (RCT) approach. RCTs are essential to test theories and build new knowledge. This is how medical science has advanced. But when we are engaged in quality improvement (QI) we are designing analytic rather than enumerative studies as was discussed in Chapter 2. Do not make your measurement efforts so precise and pure that you never proceed to the most important question: "Are we making a difference?" In short, if an organization spends its time developing academically or scientifically precise measures, it will probably never get started on its QMJ. The desire for precision will be a convenient detour in your QMJ and an excuse for avoiding the measurement mandate.

This detour was demonstrated very nicely to me by a group of physicians during an evening meeting designed to discuss their hospital's QI plan. The manager of quality was doing a very good job of presenting the plan and the related indicators. Then she got to the project on deep vein thrombosis (DVT). She described the indicator (percentage of patients evaluated for DVT risk) and then showed the historic baseline and the results for the last 8 months. Instead of discussing why the hospital's performance on this indicator was declining, the physicians became embroiled in a debate over the number of charts being reviewed and whether the sample of patient charts had sufficient "power" to be statistically significant. The detour they took was based on not understanding sampling methods for QI projects. The sample pulled for the improvement project (20 charts per month through a stratified random process) was good enough for the purposes at hand (i.e., determining how well the hospital was evaluating the risk for a DVT). As I sat and watched this discussion unfold, I realized that it was a perfect example of Roadblock #2. They were questioning the method and the data instead of discussing the processes by which they evaluate a patient's potential for a DVT. Precision was creating a roadblock for improvement.

Roadblock #3: Using Standards as Performance Objectives

Standards basically set limits on performance. In fact, standards are usually considered minimal acceptable levels of performance. Excellence is a very different concept. For example, when you go to a restaurant you have certain standards you expect to have without paying. You expect to have a table, a chair, eating utensils, a napkin, salt and pepper, and a water glass. What if, however, the waiter showed you to an open area of the restaurant that had none of these expected standard components and told you that the table would cost $20, a chair $15, utensils $10, salt and pepper $2 each, a napkin for $4, and a water glass will cost you $5? You went to the restaurant with expectations that certain minimal standards would be met before you ordered your meal. Not finding them you'd probably leave.

What are the minimal acceptable standards in a hospital or medical setting? In the United States, standards for healthcare organizations are set by a variety of governmental and nongovernmental bodies. The Joint Commission (JC), which accredits hospitals and other healthcare providers, is a dominant player in this field. The JC sets standards and regularly through announced and unannounced visits provides assessments on whether or not the facility "met the standards." Once a standard is achieved, however, complacency often sets in and people say, "What more do you want from us? We met the standard." I have heard many healthcare professionals claim that they did not have to get any better because they were already at the JC standard. I guess this means they believe that their performance is acceptable and in need of no further improvement. If standards serve as the goal for the quality journey, then it will be a limited journey.

What worked to satisfy customers or meet the prescribed standards today may not be acceptable tomorrow. For example, assume that you met the JC standards during your last survey review. What are you going to do when the Centers for Medicare and Medicaid Services (CMS) starts releasing hospital data showing that your facility is "significantly" above expected mortality percentages for the treatment of heart attack patients? Your insistence that you met the JC standards will carry little weight at this point. The concepts of baseline, target, and goal provide a much better frame of reference than standards. Compliance with standards and the desire to perform only at this level, therefore, guarantee that an organization is not really committed to QI. Improvement is a never-ending pursuit of excellence. Meeting standards is acceptance of current performance and a willingness to say, "We're good enough."

Roadblock #4: Limited Knowledge of Statistical Process Control

This roadblock relates to the use of statistical techniques, such as Shewhart control charts, to (1) understand the variation that lives in a process and (2) determine whether interventions have actually made a difference in the performance of the process. Most healthcare professionals have had at least one course in statistics at some point in their careers. Yet exposure to basic statistics is not sufficient for those who plan to manage, coach, or lead improvement efforts.

Statistical process control (SPC) is a separate and distinct body of knowledge from what many refer to as "traditional" or enumerative statistical methods. Individuals who attempt to apply statistical notions (such as testing the null hypothesis and using p-values to determine statistical significance) to their QI efforts will quickly make the wrong decisions and then become disillusioned.[1] The reason for this disillusionment is simple: they are using statistical techniques and methods that are designed to answer questions about efficacy instead of techniques designed to answer questions about effectiveness and efficiency (Brooke, Kamberg, & McGlynn, 1996).

I have been teaching SPC methods to healthcare professionals for more than 30 years. During this time, there has been an increase in not only the level of knowledge that healthcare professionals have about SPC but also its application to healthcare issues. Organizations like the American Society for Quality (ASQ) and the Institute for Healthcare Improvement (IHI) have made major contributions to spreading statistical thinking and the use of SPC methods. But we are still at the beginning stages of this journey when compared to the use of SPC in manufacturing and industry. Use of SPC in these sectors can be traced back to the mid-1920s when Dr. Walter Shewhart first formalized the theories and methods behind the control chart. In 1931, Shewhart published *Economic Control of Quality of Manufactured Product*, which stands even today as the landmark reference on SPC. The good news is that healthcare professionals are becoming more aware of what SPC can do to assist them in their quality journey. The bad news is that we still have a long way to go before statistical thinking is commonplace throughout the healthcare industry.

Roadblock #5: Numerical Illiteracy

Having skills in the use of SPC is not enough to produce world-class quality. SPC provides a wonderful foundation, but the real test comes in applying SPC knowledge to overcome the fifth and final roadblock—numerical illiteracy. Wheeler (1993, p. vi) describes numerical illiteracy as follows: "Numerical illiteracy is not a failure with arithmetic, but it is instead a failure to know how to use the basic tools of arithmetic to understand data. Numerical illiteracy is not addressed by traditional courses in primary or secondary schools, nor is it addressed by advanced courses in mathematics. This is why even highly educated individuals can be numerically illiterate."

What is needed to overcome numerical illiteracy is what the Statistics Division of ASQ calls "statistical thinking." The vision of the Statistics Division is that statistical thinking will be found in all aspects of organizational behavior and performance. **FIGURE 4-1** depicts

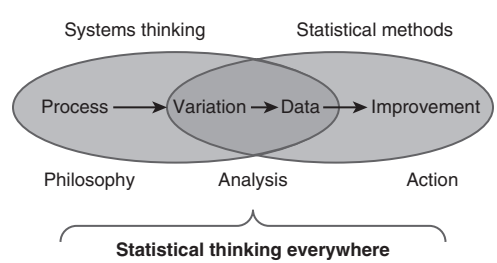

FIGURE 4-1 Vision of the ASQ Statistics Division

Reprinted with permission of ASQ.

this vision. Statistical thinking encompasses five key components:

- Systems thinking
- Statistical methods
- Philosophy (of measurement)
- Analysis (and interpretation)
- Action

As you can see from Figure 4-1, knowledge of statistical methods is only one aspect of statistical thinking. Statistical thinking is a much broader notion that has the ability not only to overcome the numerical illiteracy roadblock but also to provide a clear roadmap for the entire quality journey.

Deming's views on the value of statistical thinking are well known and have been clearly detailed in his writings (1950, 1960, 1975, 1992, 1994). Mann (1989) provides an excellent overview of how Deming and his colleagues regarded the role of statistical thinking with examples of how Deming influenced statistical thinking in this country and in Japan. In Chapter 3, titled "Statistical Methods for Tapping into the Information Flow Generated by a Process," Mann uses the following quotation from Deming to clarify the difference between using common sense and statistical thinking to make decisions: "There are many hazards to the use of common sense. Common sense cannot be measured. You have to be able to define and measure what is significant. Without statistical methods you don't know what the numbers mean" (Mann, 1989, p. 62).

Along this same line, Mann references the following point made by William Conway, the former CEO of the Nashua Corporation: "He pointed out [during a panel discussion] that one of the greatest handicaps of people who are trying to improve productivity and quality is that they attempt to deal with these matters in generalities. The use of statistics is a way of getting into specifics that will allow managers and workers to make decisions based on facts rather than speculation and hunches" (Mann, 1989, p. 62). In short, statistical thinking is a way to approach all aspects of work. It is a way of thinking about numbers and how they can be used to make improvements. Statistical thinking is the primary way to immunize yourself against numerical illiteracy.

The five roadblocks described in this section are not insurmountable. The first step in overcoming them is merely to be aware that they exist. Once they are acknowledged and understood, then it is time to take steps to immunize yourself against their proliferation. The rest of this text is directed toward this goal.

Milestones in the Quality Measurement Journey

Any successful journey begins with a plan, a good roadmap and a clear understanding of the key the milestones along the way. Developing good indicators is not all that different from planning a good road trip. The roadmap we use to guide our QMJ is shown in **FIGURE 4-2**. A completed QMJ is shown in **FIGURE 4-3**. The details on each milestone are presented next.

Welcome to Conceptland

There are two major segments of the QMJ. In the first segment of the journey, you will come upon two of the milestones: Aim and Concepts. Specifically, the team will need to develop an aim for the improvement work and then identify the relevant concepts that characterize or capture this aim. But know full well that these milestones take you to "Conceptland" not "Measurementland." Many people seem to live permanently in Conceptland. This is not a bad place to visit but if you never leave this land your QMJ will come to an abrupt and unproductive end.

The first milestone, therefore, is to establish an aim for your improvement work. What does

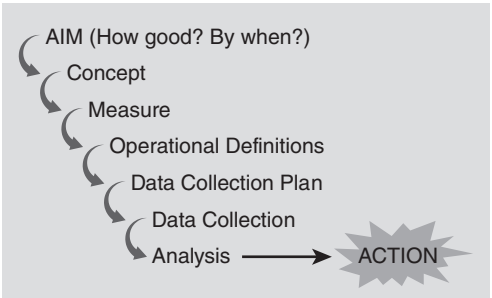

FIGURE 4-2 Milestones in the quality measurement journey

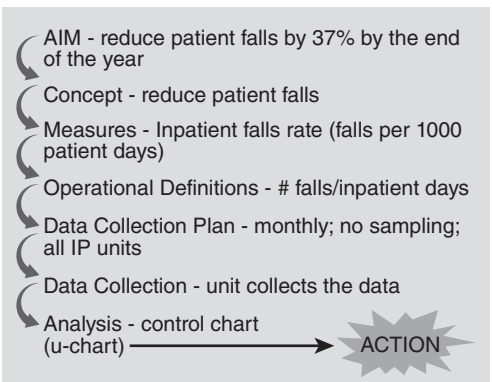

FIGURE 4-3 A completed quality measurement journey

the improvement team want to accomplish? How good do they want to be? By when do they plan to accomplish this outcome?[2] An aim statement is like a compass, it sets the direction for the QMJ and points you toward your destination: Measurementland. But Aims and Concepts are not indicators.

In Figure 4-3, we see that the team's aim is to reduce inpatient harm by 37% by the end of the calendar year. Is this an indicator? No. It is a desired end state or a vision of what could be. So, we ask the team to be more specific in order to measure the dimension of "patient harm." Then they respond, "OK, a key dimension of inpatient harm is falls. So we need to reduce inpatient falls." Is reducing inpatient falls an indicator? No. It is a concept that captures one dimension of harm. We still do not have specific quantifiable indicators that allow us to measure inpatient falls. Stating an aim or even the concepts that further define the components or aspects of an aim causes teams to live in Conceptland.[3] For example, when I ask teams or managers what they plan to measure they say things like, "We need to improve patient satisfaction," "We need to reduce medication errors," or "We need be more efficient." Again these are visions of what might be. They are noble and good but the statements are visions or desired end states. The problem is that these are not even aim statements because the concepts they are referencing (i.e., patient satisfaction, medication errors, or efficiency—concepts not indicators) do not have a specific reference as to how good they want performance to be and by when they expect to achieve this result. They are visions of what might be.

This does not mean that these two milestones are not important. These types of statements are essential in order to get a team pointed in the right direction for the start of their QMJ but such statements only provide a vague sense that we need to go "that way."[4]

In Figure 4-2 and 4-3, therefore, you can actually draw a line after the second milestone (Concept), which serves as a frontier or barrier that separates the two milestones in Conceptland from the remaining five milestones in, you guessed it. . . "Measurementland."

You Are Now Entering Measurementland

Because many people (e.g., board members, nonexecutives, senior leaders, the press, political leaders, and even patients) either live in or frequently visit Conceptland, an organization needs to have individuals with skills that can move teams and leaders beyond visions, aim statements, and concepts to address the milestones marking the road through Measurementland. These individuals need to have skills in building

quantifiable indicators (e.g., a count, a percentage, a rate, a score, an index, days between an event or cases between events) that are accepted as reasonable ways to capture the concepts of interest, build data collection plans, and have knowledge of applying SPC methods to the collected data.

Indicator Milestones

Before you actually start your measurement journey and reach the individual milestones in the QMJ, however, you need to make two brief stops in order to prepare for your journey. The first stop is where you need to decide upon the types of indicators that appropriately capture the team's aim and related concepts that need to be tracked. The second stop is where you will need to select specific indicators within the various types you have identified. Let's start by considering the various types of indicators that could be used to capture a team's aim and related concepts.

Besides the seminal work of Florence Nightingale and Ernest Codman (see Chapter 1 for details) Avedis Donabedian is another physician leader who contributed significantly to the field of indicator development. Donabedian provided the first contemporary framework for developing what I consider to be a balanced set of indicator types related to the delivery of medical care. In his classic two-volume work *Explorations in Quality Assessment and Monitoring* (1980, 1982), Donabedian described, in considerable detail, three key points in the delivery of medical services:

- Structures (the tools, resources, and organizational components)
- Processes (activities that connect patients, physicians, and staff)
- Outcomes (results)

He then suggested that measures should be developed to capture these three dimensions of medical service. Even though Donabedian provided a simple model for organizing indicators, like Codman, he, too, was a little ahead of his time. Most healthcare professionals during the early 1980s did not readily embrace Donabedian's model for evaluating medical quality or his suggestions for building indicators that represent structures, processes, and outcomes.

Kaplan and Norton (1992, 1993, 1996) made major contributions in this area by describing the components of what they call a "balanced scorecard." Even though their work has been directed more toward for-profit companies, the basic message they present is applicable to the healthcare industry. Specifically, they argue that "no single measure can provide a clear performance target or focus attention on the critical areas of the business" (1992, p. 71). According to Kaplan and Norton, an organization should monitor a set of "balanced" indicators that represent the key strategic areas in the organization's business plan. A well-selected and organized set of indicators should also place strategy and vision, not control, at the center of the organization (1992, 79). The key word for me when I first read the work of Kaplan and Norton was "balanced." Health care has a long and rich history when it comes to tracking data. What we have not done particularly well, however, is to make sure that the data we do collect is tied to our strategic objectives and represents, therefore, a balanced set of measures that cut across the full range and scope of the clinical, operational, and customer-focused services being delivered.

The Joint Commission, a U.S.-based and international accreditation body for healthcare providers, in 1993 identified nine dimensions of clinical performance that could be used to categorize indicators:

- Appropriateness
- Availability
- Continuity
- Effectiveness
- Efficacy
- Efficiency
- Respect and caring
- Safety
- Time lines

The Institute of Medicine's (IOM) report *Crossing the Quality Chasm* (2001) played a major role in identifying six aims that many organizations use to organize their indicators:

- Safe
- Effective
- Patient centered
- Timely
- Efficient
- Equitable

Another very useful way in which to think about categories or types of indicators is the value compass (Nelson, Batalden, & Godfrey 2007). The authors propose two forms of the value compass: one for clinical systems and the other for the patient. The clinical system value compass proposes organizing indicators around functional outcomes, clinical outcomes, customer satisfaction, and costs or resource issues. On the patient side the four dimensions are similar but have slight modifications to accommodate the VOC: functional status, expectations (of the patient), clinical status, and costs or resource issues.

At the IHI we teach teams to consider three types of indicators: outcome, process, and balancing. This is what we refer to as a family of measures that capture three distinct and critical aspects of any improvement effort:

- **Outcome Indicators:** These indicators should reflect and capture the VOC. How is the process or system performing in light of the stated aim? What are the results? How close are the observed outcomes to the specified targets or goals? How satisfied are the individuals who receive the outcome(s) of the process?
- **Process Indicators:** These should reflect the processes and their related indicators that drive the outcomes. How much variation is there in the process? Are the parts or steps in the process or system performing as planned? Are the process indicators you select causally connected to the outcomes?
- **Balancing Indicators:** These indicators help you look at a process or system from different directions or dimensions. Balancing indicators help you think about unanticipated consequences or other factors that might influence the outcome. Indicators of this type will help you determine if (1) you have improved one aspect of the system but made something else worse or (2) witnessed an improvement in the outcome that was not causally related to anything that the team actually did to change the process.

TABLE 4-1 provides examples of outcome, process, and balancing indicators for a family practice clinic.

In Table 4-1, the topic of interest is focused on the patient experience. Two concepts are being addressed: (1) waiting time and (2) patient satisfaction. Two outcome indicators have been identified: (1) the total length of time (in minutes) for a scheduled appointment at the clinic (note that it is only for scheduled appointments); and (2) the percentage of patients marking Strongly Agree to the single question "Would you recommend our clinic to family and friends."

In terms of process indicators, the team decided to track four dimension of the care process. Two of the indicators relate to different components of waiting (i.e., check-in time to being seen by the doctor and time spent waiting for ancillary service). The third indicator focuses on the discharge process and whether the patient received appropriate discharge instructions related to the reason for the visit. The fourth and final process indicator is qualitative in nature (patient and staff comments on the flow of the process).

The final column in Table 4-1 addresses balancing indicators of which there are four. Remember that in specifying balancing indicators we are attempting to understand whether our improvement efforts are creating any unintended consequences. For example, consider the first balancing indicator volume of patients. What if the volume of patients coming to the clinic or scheduling appointments for a particular month declined? What impact might a declining number

TABLE 4-1 Outcome, process, and balancing indicators for a family practice clinic

Topic	Outcome Measures	Process Measure	Balancing Measures
Improve waiting time and patient satisfaction in the family practice clinic	Total length of stay (in minutes) for a scheduled appointment at the clinic Percentage of patients marking Strongly Agree to the question "Would you recommend our clinic to family and friends?"	Time from check-in until seeing the doctor Patient/staff comments on the flow Percentage of patients receiving discharge material Wait time for ancillary services (lab, x-ray, ultrasound) during a visit	Volume of patients Percentage of patients leaving without being seen by a doctor Staff satisfaction Financials

of visits or scheduled appointments have on wait times? We would most likely see a drop in wait times to see the doctor because there are fewer people in the pipeline and thus the backups and delays would be reduced. So, when you show up for your appointment you get to see the doctor in less time than your previous visit simply because the volume of patients coming to the clinic has been reduced. Chances are that patients would also be more satisfied with the process (one of the outcome indicators) because they waited less time to see the doctor.

Similarly, if the percentage of patients leaving without being seen by the doctor (the second balancing indicator) goes up, chances are that those who do not leave will be seen faster and therefore also have higher satisfaction levels. In both these situations, the team has done nothing to intentionally improve the process. Other factors (i.e., reduced volume and an increase in the percentage of patients leaving without being seen) created a false impression that things have gotten better. Now consider staff satisfaction as a balancing indicator. What if the improvement team did make changes in the clinic's process (the details of which are not important at this point) that actually reduced wait time and improved patient satisfaction? But when you assess staff satisfaction you discover that it is going down. When you talk to the staff you get comments like this: "Sure the changes you made to the process have improved things for the patients but they have made operating conditions for the staff more complicated. If this continues I know of a couple staff members who are considering leaving the clinic." In this case, what have you gained? You have improved the status of one group (i.e., the patients) and compromised another group (i.e., the staff). Deming referred to this as suboptimizing the system. Balancing indicators help prevent suboptimization and make sure you are considering unintended consequences of your efforts.

Irrespective of the various types of indicators that could be identified, the key point is that a balanced approach to the types of indicators is far superior to a narrow focus. A singular or narrow focus on one or even two types of indicators will lead to shallow knowledge and

ultimately suboptimal performance of improvement teams. A balanced approach to indicator development does not mean, however, that you have to measure 30 or 40 indicators. Focusing on the vital few (with emphasis placed on the word "few") is preferable to assembling an unmanageable array of indicators that require a small army to collect, analyze, and interpret. More will be said on this point in Chapter 5 when we look at the development of strategic dashboards.

Selecting a Specific Indicator

Once you have decided which types of indicators are most appropriate, the next step is to select the specific indicator(s) that will be measured within each type. Although this seems like a straightforward activity, I have found it surprising how many teams struggle with this task. *An indicator is a specific quantifiable aspect of an outcome or a process.* Yet all too often teams confuse themselves by wandering around in Conceptland and never move on to the detailed markers indicating that they have actually entered Measurementland. For example, in Figure 4-3 the concept of interest is to reduce inpatient falls. There are two critical aspects of moving from this concept of reducing inpatient falls to actually measuring whether falls have been reduced. First, the team needs to decide on the specific quantifiable indicator(s) that will represent inpatient falls. Second, the word "reduce" is not relevant to the selection and development of the indicator(s) that will be used to measure the concept of inpatient falls. Whether the indicator demonstrates a reduction or an increase or stays at the current level is irrelevant to specifying the indicator. We will find out if the indicator is moving in the desired direction once we collect data and move to the analysis milestone in the QMJ. Until then, the targeted direction for the indicator (i.e., an increase in indicator X or a decrease in indicator Y) has no added value to the naming and specification of the indicators. Too often, however, teams focus almost exclusively on the direction of change, the target, the expected goal, or the desired end state and end up developing confusing indicators.

In terms of our inpatient falls example from Figure 4-3, the critical question is what specific indicators do you propose to develop that capture the concept of inpatient falls? The following specific indicators could be used:

- The number of inpatient falls (e.g., a simple count of the number of inpatient falls each day or week)
- The percentage of inpatients who fell once or more while they were in the hospital
- The falls rate, which includes multiple falls by the same patient during their admission and is defined as the number of falls per 1,000 inpatient days[5]
- Days between inpatient falls

Each of these indicators identifies a specific way to look at the inpatient falls concept. Each indicator has value and the team will have to decide from the various ways to measure inpatient falls which one or two indicators will best serve as the outcome indicators. The team will also need to identify a list of indicators for processes related to the falls prevention process and they should consider selecting one or two balancing indicators to provide insights on the issue of suboptimization.

TABLE 4-2 provides examples of concepts and the specific indicators that could be used to measure each concept. The decision as to which indicator is selected (from this list or a new list of indicators that a team might develop) depends on the questions that a QI team is trying to answer, the availability of data, and ultimately the team's aim. If you phrase the question in terms of the absolute volume of an activity, you might be interested in tracking a simple count of the number of events (e.g., the number of inpatient falls). If, on the other hand, you were interested in a relative measure, then you would be better off measuring falls as a percentage or possibly as a rate. When it comes

TABLE 4-2 Moving from a concept to a specific indicator

Concept	Potential Indicators for This Concept
Patient falls	■ The number of patient falls ■ The percentage of patient falls ■ The patient falls rate ■ The number of days between inpatient falls
Cesarean sections	■ The number of cesarean sections ■ The percentage of cesarean sections ■ The cesarean section rate
Care of surgical patients	■ The percentage of post-op deaths (sorted by American Society of Anesthesiologists class) ■ The number of days between the occurrence of post-op deaths ■ The percentage of unexpected returns to surgery ■ The number of successful cases before there was a return to surgery within 24 hours
Care of coronary artery bypass graft (CABG) patients	■ Intubation time post CABG ■ The percentage of prolonged post-op CABG intubations ■ The percentage of CABG patients with a hospital acquired infection ■ The percentage of CABG patients returning to surgery within 24 hours
Patient scheduling	■ The average number of days between a call for an appointment and the actual appointment date ■ The percentage of appointments made within 3 days of the call for an appointment ■ The number of appointments scheduled each day ■ The number of days between a call for an appointment and the first available appointment
Employee retention	■ Total number of full-time equivalents (FTEs) ■ Percentage of employee turnover ■ Employee turnover rate ■ Average number of years employed by the organization ■ The percentage of new hires who leave during the first year
Employee evaluations	■ The number of evaluations completed ■ The percentage of evaluations completed on time ■ Variance from due date of a completed evaluation
Care of emergency patients	■ The number of unplanned returns to the emergency department (ED) within 24 hours ■ The percentage of ED patients admitted as inpatients ■ The percentage of ED transfers to other facilities ■ The patient wait time in the ED

(continues)

TABLE 4-2 Moving from a concept to a specific indicator *(continued)*

Concept	Potential Indicators for This Concept
Implementation of a restraint protocol	■ The number of patients who had restraints applied ■ The percentage of patients placed in restraints ■ The restraint usage rate
Documentation of histories and physicals (H&Ps)	■ Transcription turnaround time ■ The time from patient admission to the physician-dictated H&P ■ The percentage of incomplete H&Ps
Medication usage	■ The total number of medication orders placed each day ■ The number of medication orders that had one or more errors ■ The time it takes to deliver a med order to the unit once the order is received in the pharmacy ■ The medication error rate ■ The number of wasted IVs
Customer satisfaction	■ The number of patient complaints ■ The percentage of patients providing positive responses to a survey ■ The percentage of patients who indicated that they would recommend the facility to a family member or friend ■ The percentile ranking for employee satisfaction in a national database ■ The percentage of physicians indicating that your hospital is an "excellent" facility
Home care visits	■ The number of home care visits ■ The average time spent during a home care visit ■ The percentage of time spent traveling during each home care visit ■ The number of visits each days for each home care nurse ■ The number of bottles of home oxygen delivered
Pastoral care	■ The number of patient encounters by the pastoral care staff ■ The number of minutes spent during a patient encounter ■ The percentage of inpatient admissions that have properly documented the patient's religious preference ■ The number of requests from nursing units for assistance
Delivery of oncology services	■ The percentage of outpatient oncology patients who have to be admitted ■ An individual patient's platelet counts ■ The total inpatient cost to treat a cancer patient ■ Mood scale index scores for cancer patients
Successful quality improvement (QI) training	■ The number of participants attending a QI class ■ The percentage of cancellations ■ The percentage of no-shows ■ The information recall scores at 30, 60, and 90 days

TABLE 4-2 Moving from a concept to a specific indicator *(continued)*

Concept	Potential Indicators for This Concept
Ventilator management	■ The number of patients on a ventilator ■ The percentage of patients placed on a ventilator ■ The number of days on a ventilator ■ The ventilator-associated pneumonia rate
Electronic access to information	■ The percentage of med orders submitted via the computerized physician order entry (CPOE) system ■ The minutes of system downtime ■ The percentage of physicians who regularly use online protocols ■ The number of visits (hits) to the organization's website
Outpatient testing	■ The total number of outpatient visits and therapy ■ The wait time to have a blood draw (or any other procedure) ■ The percentage of outpatient procedures with a complication ■ The complication rate for outpatient procedures ■ The time it takes to complete a colonoscopy procedure
Lab production	■ Lab turnaround time ■ The total number of lab orders ■ The percentage of inaccurate lab orders ■ The percentage of stat lab orders exceeding target ■ The percentage of stat lab orders

to indicator selection, there are more options than most people realize. It is also important to realize that there are no universally accepted "best" indicators of healthcare performance. A concept may be the same (e.g., inpatient falls) or even types of measures (e.g., outcome, process, or balancing) across different systems, regions, provinces, or even countries but the specific indicators and the subsequent milestones that mark the QMJ can be very different.

TABLE 4-3 provides a worksheet to help you move from concepts to indicators. In the left column of this worksheet, list the concepts you are interested in measuring. The next column should then list the specific quantifiable indicator(s) (e.g., count, percentage, rate, score, index, days between, cases between) you think will best capture each concept of interest. Finally, indicate whether each listed measure is an outcome, process, or balancing measure. **TABLE 4-4** provides an example of a completed indicator worksheet. A key point related to this worksheet is that you do not need to have a lengthy summary of each of your indicators. You can take this completed worksheet to a management meeting and say, "Here are the indicators for our improvement team." It is clear and yet specifies the key components of how you have moved from a concept to specific indicators plus identifying the types of indicators you will be tracking.

Summary conclusions about moving from a concept to a quantifiable indicator are provided in **BOX 4-1**.

TABLE 4-3 Organizing your indicators worksheet

Topic for Improvement: _____

Concept	Potential Indicators	Outcome	Process	Balancing

TABLE 4-4 Example of a completed organizing your indicators worksheet

Topic for Improvement: Inpatient Falls Process

Concept	Potential Indicators	Outcome	Process	Balancing
Patient harm	Inpatient falls rate	✓		
Patient harm	Number of falls	✓		
Compliance	Percentage of inpatients assessed for falls		✓	
Staff education	Percentage of staff fully trained in falls assessment protocol		✓	
Assessment time	The additional time it takes to conduct a proper falls assessment			✓

BOX 4-1 Conclusions about moving from a concept to an indicator

1. Moving from a concept to an indicator requires focused work to create agreement about adjectives, such as recovery, major, timely, complete, accurate, or excellent.
2. A concept may need more than one indicator and, therefore, the development of more than one operational definition.
3. The transition from concept to an indicator doesn't just happen; it requires both technical and clinical decision making to be blended with pragmatism and acceptance of the imperfections of the measures.
4. There is no such thing as a fact! (W. E. Deming)

Developing Operational Definitions

The real work of indicator development begins after you have selected and named a specific indicator. Now it is time to develop an operational definition. I find the specification of operational definitions to be one of the more interesting and intriguing aspects of indicator development. Every day we are challenged to think about operational definitions. They are not only essential to good measurement but also critical to successful communication between individuals. For example, if you tell your teenage son or daughter to be "home early" from a party, you will quickly understand the necessity of establishing a clear operational definition.

An operational definition is a description, in quantifiable terms, of what to measure and the specific steps needed to measure it consistently. A good operational definition:

- Gives communicable meaning to a concept or idea
- Is clear and unambiguous
- Specifies the measurement method, procedures, and equipment (when appropriate)
- Provides decision-making criteria when necessary
- Enables consistency in data collection

Some groups are better at developing operational definitions than others. For example, political leaders typically shy away from clear and unambiguous operational definitions so they can change their positions or follow a different approach. But in this age of instant information, social media, and the ability to record statements easily and quickly politicians are starting to be more concerned about the terms and definitions they use. For example, consider the following list of terms that are used frequently during political campaigns:

- A "fair tax"
- A "tax loophole"
- We need to "jump start" the economy
- The "rich" need to give more to the "poor"
- The "middle class" needs tax relief
- We need to get this country "moving" again
- The "small farmer" needs economic support

All of these terms require clear operational definitions if there is to be a consistent understanding of what they mean and how we would measure them. In the political arena, however, the desire is to frequently have a certain amount of ambiguity surrounding concepts and terms so that the person presenting the idea cannot be held to a single position or definition.[6]

On a more personal note, I had a great example of an operational definition when my daughter Devon was 9 years old. Devon and her friend Janine called up to me and asked, "When are you going to take us for the ice cream you promised?" My answer would have made any politician proud. I responded confidently, "Soon." That appeased them for about 15 minutes.

Then they called up again, and this time when I answered, "Soon," they demanded to know how many minutes made up "soon." Even a 9-year-old child understands the need for a clear operational definition.

One of the more interesting problems with an operational definition involved the September 23, 1999 incident with the Mars space probe. European scientists used metric measurements and calibrations (newton-seconds) to guide the spacecraft. The probe was built by Lockheed Martin and their engineers used decimal referents (pounds/foot-seconds) to calibrate the maneuvering of the probe. When the probe went around the far side of Mars and was ordered to go down toward the surface for a closer look and flyby, it was essentially receiving two different sets of operational definitions. It followed the programming commands but because of the differences in the operational definitions used by the builders of the spacecraft and those maneuvering it the probe took a trajectory that took it entirely too close to the planet, causing it to burn up in the Martian atmosphere. The difference between metric and decimal units of measurement created an inconsistent operational definition of the term "unit of distance." As a result, a $125 million project became a NASA embarrassment.

A more recent example of confusing operational definitions can be found in the ongoing debate over what is a "healthy or natural" food. The U.S. Food and Drug Administration (FDA) has been debating these terms for years. Food scientists maintain that a majority of our food products are not natural and therefore not healthy because they have been processed in one form or another and that they are no longer a "product of the earth." A similar debate has been occurring for over 15 years on the definition of an "organic fish." I thought all fish were organic because I have never been served a mechanical fish but I must be missing something in this debate. It turns out that the debate hinges on what the fish in a fish farm (not wild fish but farm-raised fish) are fed. If they are fed pellets that are made of other fish they are considered "organic." But, if they are fed pellets that consist of vitamin-enriched corn, wheat, and other non-fish protein then they are defined as "nonorganic." Seriously, you cannot make up stuff that is better than what you discover in real life.

One final personal story about operational definitions before returning to healthcare examples. This is absolutely one of my all-time favorites. My wife Gwenn was a nationally ranked official for women's field hockey in the United States. She was doing one of the final games for the National Collegiate Athletic Association (NCAA) championship in Iowa City, Iowa in the middle of November. It was very cold. So cold in fact, that they had to place the hockey balls in a small warming device so they would not crack or break when hit. The game was tied when one of the forward players hit a wicked shot that was headed toward the opponent's goal cage. The telltale sound of a hockey ball hitting the metal backplate of the cage was heard by all. The attacking team believed they had just scored a winning goal. But the defending team quickly pointed out that only half the ball was in the cage. The other half of the ball was still out on the playing pitch. So now what? Is it a goal or not a goal? What is the operational definition of a goal? Time out was called. The officials met at the center of the field along with the timer, the backup umpires, and an NCAA judge. This is a very important game so a decision has to be made. Goal or no goal? Gwenn, who is the lead official, is asked to render an opinion. She honestly says, "I have no idea. This has never happened before." So, they turn to the official rule book to see if there is an operational definition of "a goal." After a couple of minutes that seemed to last hours, Gwenn announces that there is an answer in the rule book. It clearly states that a goal occurs "when the entire ball passes the plane of the goal line." With half the ball still lying on the pitch and the other half in the cage the answer is easy. . .no goal. On the ride home I was mulling over the operational definition of a goal and asked Gwenn a question "What if upon being hit the ball did break in half but both halves went across the goal line and into

the cage? Would this now be considered a goal? The entire ball was in fact in the cage." There was silence for a few seconds then she said, "I'm not even going to address that question." But I still wonder whether it would be a goal.

Every day healthcare professionals must deal with operational definitions. There are many healthcare terms that beg for more precise operational definitions. How does your organization define the following terms?

- A patient fall
- A restraint
- A good outcome for the patient
- A medication error
- A complete and thorough physical exam
- A good employee performance review
- Surgical start time
- An accurate patient bill
- A readmission
- A successful surgical outcome
- An organization that supports its workers
- A late food tray
- A clean patient room
- Healthcare disparities
- A quick admission
- A blameless culture for reporting errors

Consider one of these terms that has intrigued me for years—a patient fall. One of the first definitions I heard for a patient fall was "a sudden and rapid movement from one plane to another." This sounds like something you try to do at a busy airport rather than the definition of a negative patient outcome. It is not very precise and leaves a lot to the imagination. I have frequently heard nurses talk about two basic types of falls: partial falls and assisted falls. *Partial falls* usually occur when the patient attempts to get out of bed and discovers that he or she does not have an adequate amount of strength to permit ambulation. In this case, the patient might stagger a little, slump back onto the bed, try to stand again, and attempt to make it to the chair by the window but ends up collapsing to the floor. As I have explored this scenario with nurses and asked them if this constituted a partial fall, I get mixed responses. One meeting I especially remember produced two very different views of a partial fall. After describing the conditions of a partial fall, half of the nurses indicated that they would classify the situation as a partial fall because the patient did bounce around a little before ending up on the floor. Their reasoning was that the patient bounced around a little, came in contact with some furniture, and eventually ended up on the floor. The other nurses in the group reserved their opinion until they found out the answer to one question: "Did the patient's knee hit the floor first?" If the answer was "Yes," then they agreed it would not be a partial fall. If the answer to this question was "No," however, this group of nurses believed that this was a partial fall. The knee touching the floor was the primary determinant of a "partial fall."

Assisted falls are even more interesting than partial falls. When I first heard this term I envisioned nurses getting so fed up with a patient that they gave him a gentle nudge and "assisted" him in falling. As I learned more about this topic, however, I came to realize that an assisted fall fortunately has nothing to do with the nurses causing the fall. It does, however, have a very distinctive operational definition. Here is the scenario. A patient decides to go for a walk tethered to his IV pole. The patient takes a few steps then announces to the nurse that she does not feel very well and that things are starting to move in circles. As the patient begins to sway the nurse moves into position, grabs the patient, and assists her to the floor. But is this really a fall? It seems to me it is more like a recline or possibly a lay-down. Most nurses I have worked with agree that being present when a patient is starting to go down and intervening to help break the patient's fall constitutes an "assisted fall." But there is certainly not universal agreement on the precise operational definition of an assisted fall.

You can see the problem that all this poses for measurement. If you are part of a multihospital system, a region, province, or a country or plan on comparing hospital outcomes across providers, then you should make sure that each provider being compared is defining the

indicator of interest in the same way. Without such consistency you will end up with apples and oranges at best and more likely apples and carburetors. The pieces will not be comparable, which means that ultimately the conclusions that are derived from the data are not accurate. All good measurement begins and ends with operational definitions.

An example of an operational definition for the percentage of medication errors is summarized as follows.

Indicator Name: Percentage of medication errors

Numerator: Number of outpatient medication orders with one or more errors. An error is defined as wrong med, wrong dose, wrong route, or wrong patient

Denominator: Number of outpatient medication orders received by the family practice clinic pharmacy

Data Collection:

- This indicator applies to all patients seen at the clinic
- The data will be stratified by type of order (new versus refill) and patient age
- The data will be tracked daily and grouped by week
- The data will be pulled from the pharmacy computer and the computerized physician order entry (CPOE) systems
- Initially, all medication orders will be reviewed. A stratified proportional random sample will be considered once the variation in the process is fully understood and the volume of orders is analyzed.

A second example of an operational definition provides the details for a perioperative nasal swabbing indicator.

Indicator: Percentage of patients undergoing hip and knee replacement surgery during the measurement period who have had preoperative nasal swabs to screen for *Staphylococcus aureus* carriage

Goal: 95%

Frequency of Data Collection: Monthly

Numerator Definition: Number of patients undergoing hip or knee replacement surgery who have had a nasal swab specimen processed to screen for *Staphylococcus aureus* carriage prior to surgery

Denominator Definition: Number of patients undergoing elective hip or knee replacement surgery

Numerator and Denominator Exclusions:

- Patients who are less than 18 years of age
- Patients who had a principal or admission diagnosis suggestive of preoperative infectious diseases
- Patients with physician-documented infection prior to surgical procedures
- Patients undergoing nonelective hip or knee replacement surgery

Definition of Terms: Hip or knee replacement surgery includes operations involving placement of a nonhuman-derived device into the hip or knee joint space. ICD-9 Codes include 00.70-00.73, 00.85-00.87, 81.51-81.53, 00.80-00.84, 81.54, and 81.55.

Calculate as: (numerator/denominator * 100, with only 1 decimal place)

Summary conclusions about developing operational definitions are provided in **BOX 4-2**.

BOX 4-2 Conclusions about developing operational definitions

1. Operational definitions are not universal truths!
2. Operational definitions require agreements on terms, measurement methods, and decision criteria.
3. Operational definitions need to be reviewed periodically to make sure everyone is still using the same definitions and that the conditions surrounding each measure have not changed.

Developing Data Collection Plans

I have separated this milestone from the actual collection of data because I do not believe that as an industry we have devoted enough time to thinking about the numerous factors that influence the success or failure of our data collection efforts. Most people want to move directly from "I have an indicator" to "Let's go get some data" without spending much time thinking about how to actually collect the data. From my perspective, planning for data collection should occupy about 80% of your data collection time and the actual act of collecting the numbers should consume about 20% of your time.

Data collection is not unlike other aspects of life that require planning. Whether the activity is painting a house, planting a garden, or going on a major vacation, preparation is key. If you do not spend enough time preparing a wooden house, the paint will not last as long as you would like it to. Similarly, if you do not take time to properly prepare the soil in your garden, the seeds and young plants will not get off to a very good start. Finally, a major vacation (e.g., a cruise or a bed-and-breakfast tour of Ireland) usually requires more time to plan than the time you actually spend on the holiday itself. The act of data collection is very similar. Inadequately prepared data collection plans will usually produce unacceptable results. The data will be challenged, questioned, and/or seen as being rather useless.

There are several important data collection issues that require some elaboration, most notably stratification and sampling. Stratification is one of the best things a team can discuss when building indicators, yet it is frequently overlooked. Stratification is more of a logical issue than a statistical one. It essentially consists of the separation and classification of data into categories or homogeneous buckets that reflect common characteristics. The objective of stratification is to create strata or categories within the data that are mutually exclusive and allow you to discover patterns that would not otherwise be observed if the data were all aggregated. The overall strategy is to minimize the variation within a stratification category in order to compare the variation between categories. By doing this you can increase your knowledge about the possible influence that the stratification levels might have on the outcome indicator. Frequently used stratification levels include:

- Age
- Gender
- Socioeconomic status
- Prior admission for the same diagnosis
- Day of the week
- Time of day
- Month of the year
- Shift (day, afternoon, night)
- Type of order (stat versus routine)
- Type of ambulatory procedure
- Type of surgery
- Machine (such as ventilators or lab equipment)
- Severity of the patients
- Tenure of the staff

If you do not think about the factors that might influence the outcome of your data before you collect the data, you run the risk of having to try to tease out the stratification effect manually after the data have been collected. At this point not only is it too late to effectively address the stratification question, but you will also have to engage in rework and wasted time to even attempt to untangle the stratification questions.

FIGURES 4-4 and **4-5** provide examples of stratification problems. In the first example (Figure 4-4), the indicator of interest is turnaround time (TAT) in the lab (the particular test does not matter at this point). The data reveal that the process displays extremes because the team did not separate the TATs for the day and evening shifts. They merely collected data and combined the two shifts, which are obviously different. In this case, the average TAT will fall exactly in the middle of the two extremes of data. The average and even the standard deviation of the TAT are meaningless statistics for data like this. The mean and

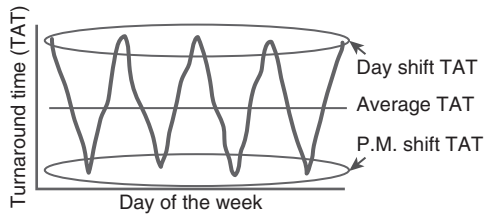

FIGURE 4-4 A stratification problem with turnaround time

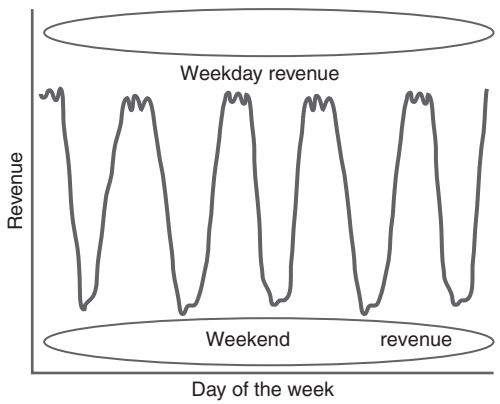

FIGURE 4-5 A stratification problem with tracking revenue

standard deviation can be calculated of course but they are mathematical artifacts that are the result of two distributions of data—one high and one low. These data should be separated and two charts should be made—one chart for the day shift TAT and another for the evening shift. Stratification is an essential aspect of data collection. If you do not spend some time discussing the implications of stratification, you will end up thinking that your data are worse (or better) than they should be.

In Figure 4-5, the indicator of interest is revenue by day of the week. There are several data points in a row that are at relatively high and roughly at the same level. Then there is a sharp drop in the data for two data points. This pattern demonstrates a clear problem with stratification. In this case, the revenue generated Monday through Friday (the higher data points) is markedly different from that generated on Saturday and Sunday (the lower data points). This hospital is clearly not a 7-day-a-week hospital. If you were to calculate the average revenue generated per day for this hospital you would get a misleading number. Although the overall mean would be skewed toward the weekday revenue side of the chart (because of the higher volume generated during this time as well as more days), it would not reflect the average generated during the weekdays. For this example, someone should have said before the data were collected, "Because we do not generate the same amount of revenue on the weekend as we do on weekdays, we should stratify the data into two categories—weekday revenue and weekend revenue—and analyze the data separately."

Sampling is the second key component of a data collection plan. Not every data effort will require sampling. If a process does not generate a lot of data, then you will probably analyze all the occurrences. This happens most often when the indicator is a percentage. For example, when we compute the percentage of primary C-sections for the month we typically do not use a sampling plan. We usually take all the C-sections for the month and divide this numerator by the total number of deliveries (the denominator) for the month. When a process generates considerable data, however (e.g., lab TAT for blood tests or all admissions during the month), a sampling plan is usually appropriate. From my perspective, building knowledge of sampling methods is one of the most important things you can do to establish efficient and effective data collection strategies.

Like stratification, sampling deals more with logic than statistics. Individuals trained in the social sciences are typically exposed to extensive training in sampling principles and concepts. Unfortunately, most healthcare professionals are given only a cursory foundation in this subject. The irony with this situation is that sampling is actually quite easy. Healthcare

professionals would grasp sampling principles quickly if they were exposed to them throughout their formal training.

Try this simple test to demonstrate this point. The next time you are with a group of healthcare professionals, ask them, "Have any of you ever drawn a random sample?"[7] Rather quickly you will receive a bunch of positive nods. When you ask one of the people who was nodding rather energetically how they actually drew the random sample, they will usually announce rather proudly that they "picked every 10th chart." Selecting every 10th chart is a form of random sampling known as *systematic sampling* (described later), but it can introduce considerable bias if the steps involved in drawing a systematic sample are not followed.

The purpose of sampling is to be able to draw a limited number of observations and to be reasonably comfortable that they represent the larger population from which they were drawn. If you had all the time and money in the world you would never draw a sample. You would always do a complete enumeration of all cases. But time and resources are limited, so we draw samples.[8] Whenever you draw a sample, however, the key question is, "How much data do I need?" One of my professors in graduate school, Dr. Bob Bealer, had a great answer to this question. When asked by one of my fellow doctoral students how much data we should collect for our dissertation research, he merely answered, "As much as you must and as little as you dare." At the time I thought this was a clever and rather professorial response. But after spending many years trying to help healthcare professionals develop reasonable sampling strategies, I have come to realize that this was very practical advice. For example, if I wanted to check your weight by weighing you on only one day of the year, would you say this is a representative sample of your true weight? As an aside, I should tell you that the day I have selected to weigh you is Thanksgiving Day (A U.S. holiday celebrated on the third Thursday of November) after you eat. Most people would say, "No way do I want you to use my Thanksgiving Day weight." Their initial reaction would probably be correct. On the average, for example, adults consume upwards of 5,000 calories on Thanksgiving Day. So most people would probably say, "If you are going to weigh me once, check me in the spring when I am trying to get back into my shorts or bathing suit." To be even more reasonable (reliable), I might weigh you every couple of weeks as they do in many weight control programs. In this way, I would obtain a more representative sample of your weight as it fluctuates over time. Remember, as much as you must and as little as you dare.[9]

What happens if you draw a sample and it is not representative of the population from which it was drawn? **FIGURE 4-6** shows the relationship between three samples and a population. The larger curve represents the total population of interest (e.g., all asthma patients returning to the emergency department [ED] within 24 hours). Curve A identifies a properly pulled sample of patients. The shape and location of this sample are very similar to the population. Curve C, on the other hand, represents a sample that was drawn with a negative bias. In this case, you could get the false impression that your results were much worse than they really were just because you pulled a sample that came from the negative end of the population curve. Similarly, Curve B depicts a positive sampling bias, which leads you to an overly optimistic conclusion. A well-designed sampling plan will not only produce data that are representative of the population but also save time and money for those collecting the data.

There are many ways to draw a sample. The key question you have to ask yourself whenever you want to draw a sample is, "How representative and precise do I need to be with this sample?" For example, if you have received numerous calls and complaint letters about the wait time in outpatient testing and therapy, you basically have two sampling options: (1) develop a statistically based sample that allows you to generalize to your total outpatient population,

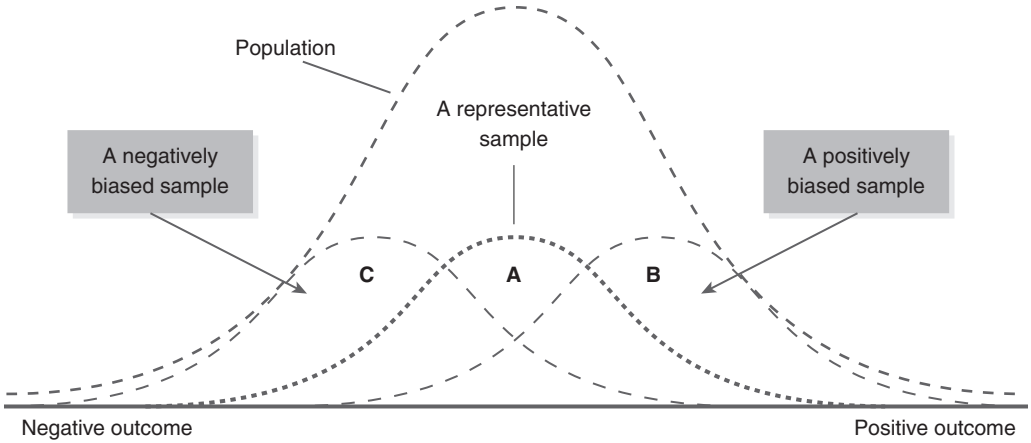

FIGURE 4-6 The relationship between a sample and the population

or (2) go out on any given day, grab a convenient handful of willing patients, and ask them how they like your outpatient testing and therapy services. If the level of precision you need to answer this question is low, then option 2 is appropriate. If, on the other hand, you need to be very sure (statistically sure) that there is a problem in outpatient testing and therapy, then you need to formulate a more scientific approach to sampling.

Ishikawa, in his classic work *Guide to Quality Control* (1982), identifies four conditions for developing a sampling plan:

- Accuracy
- Reliability
- Speed
- Economy

These four criteria should serve as a fundamental checklist for building sampling designs. Not every sample will maximize all four criteria. There are times when accuracy will be the primary objective of sampling (e.g., when designing a randomized clinical trial). At other times reliability will become more important (e.g., when you are establishing a sampling plan for patient satisfaction and you want to be able to draw reliable samples each month or quarter). Speed may be essential when you have to sample a number of blood specimens to determine whether there is a contamination problem. Finally, the economics of sampling will usually pose a challenge for everyone. Each time you draw a sample, whether it is a sample of medical records or a sample of patients, there are economic factors involved with the pull of data. Complicated sampling plans require more time, effort, and money. In the end, however, it all comes down to a fairly simple question—how can you pull an accurate, reliable, fast, and inexpensive sample? Obviously it is difficult to obtain a sample that meets all four criteria simultaneously. Sampling, therefore, really consists of a series of compromises. It basically gets us back to Professor Bealer's words of wisdom, "As much as you must and as little as you dare."

Sampling methods are basically divided into two major categories—probability and nonprobability. Any standard research methods or

statistics book will provide a review of sampling methods. I would encourage you to obtain several books on this topic to see how different writers classify and describe the various methods. Do not worry about the age of the book. Most of my books on sampling, for example, are 20 to 30 years old. Even when I pick up a new book on sampling, the terms remain virtually the same as those I find in my older books.

The terms and approaches to sampling have remained rather constant since the late 1930s.[10] I do not intend to replicate in this text what can be found in many good references (Babbie, 1979; Campbell, 1974; Daniel and Terrell, 1989; Duncan, 1986; Gonick and Smith, 1993; Hess, Riedel, & Fitzpatrick, 1975; Ishikawa, 1982; Miller, 1964; Selltiz, Jahoda, Deutsch, & Cook, 1959; Weiss, 1968; Western Electric Co, 1985). I provide, however, a brief review of the major sampling methods and let the reader explore the details.

Probability Sampling

Probability sampling is designed to provide the highest possible level of predictability and confidence in the sampled data at the most economical cost to the researcher. Although most people have some notion of what a probability or random sampling entails, many are unclear on the specific aspects of actually designing and selecting the sample. At the very foundation of probability sampling is trust: trust in statistical probability and the fact that when you draw a random sample you do not throw it away merely because it does not conform to your personal belief about what the data are supposed to tell you. I have drawn many random samples for people over the years. On numerous occasions I have been questioned about the "accuracy" of the samples because the individuals who requested the samples did not like the results. In their minds, they thought they should be allowed to pick and choose what should be included (and excluded) in the sample. If they were to do this it would not make the sample a probability sample. Whenever judgment, purposeful intent, or convenience enter into the sampling plan, you have moved from probability sampling to nonprobability sampling, which is addressed in the next section.

Campbell (1974, p. 143) identifies three characteristics of probability sampling:

1. A specific statistical design is followed.
2. The selection of items from the population is determined solely according to known probabilities by means of a random mechanism, usually using a table of random digits.
3. The sampling error—that is, the difference between results obtained from a sample survey and that which would have been obtained from a census of the entire population conducted using the same procedures as in the sample—can be estimated and, as a result, the precision of the sample result can be evaluated.

There are numerous ways to draw a probability sample. They are all essentially variations on the simple random sample.

Simple Random Sampling

A random sample is one that is drawn in such a way that it gives every element in the population an equal and independent chance of being included in the sample. This is usually accomplished by using a random number table (usually found in the back of any good statistics book) or a computer-based random number generator (found in all statistical software programs and in many spreadsheet packages). Step-by-step procedures for drawing a random sample can be found in *Probability Sampling of Hospitals and Patients* (Hess et al., 1975) and in *Flaws and Fallacies in Statistical Thinking* (Campbell, 1974). Even though this method is referred to as a "simple" random sample, the term "simple" can be a little misleading. The mechanics of drawing a random sample may

not feel simple to those who have to number all the elements in the population and learn how to apply a random number table or a computerized random number generator.

As an alternative, you can simply write the names or numbers of the population elements on separate pieces of paper, place them in a bowl, and draw out the sample. I did this to develop a sampling plan for a medical group. They wanted to sample the wait times of patients in one of their clinics, but they did not have the resources to sample every day of the week. They initially said they would pull a sample of patients every Monday. I advised them that this could produce biased results, because Mondays are typically busier than other days of the week. So I wrote the days of the week (excluding the weekend) on five pieces of paper (all of the same size), placed them in a bowl, and then drew out a day of the week. The first day I pulled was a Wednesday. This meant that during the first week, a sample of patients would be pulled on Wednesday and their wait times to see the physician would be recorded. I placed the slip of paper with Wednesday written on it back into the bowl and drew another piece of paper. The second piece of paper had Friday written on it, which would be the sampling day during week 2. I replaced the piece of paper and repeated this process 23 more times to obtain a total of 25 sample days. To pull the sample of patients on a given day, I first asked the staff to run a report showing the actual volume of patient visits by day for the last 6 months. From this report we determined the minimum and maximum number of visits as well as the mean, median, and standard deviation (to see if the data approximate a normal distribution). We determined that the average number of visits each day was 74 with a minimum of 63 and a maximum of 86. I advised them to place the numbers 1 through 86 on pieces of paper (again all the same size) and place them all in the bowl. We then proceeded to draw a random sample of eight patients from the bowl for each sampling day. The first day to be sampled, for example, was a Wednesday. On this day the following eight patients would have their wait times tracked: 43rd, 15th, 63rd, 2nd, 47th, 23rd, 18th, and 4th. Because the clinic basically knew how many patients they had scheduled for that day, they could identify those patients from the charts that had been pulled ahead of time. This allowed the staff to be prepared to track the various steps in the process for these patients (i.e., the time from check-in to being called by the nurse, time with the nurse, wait time to see the doctor, time with the doctor, and finally checkout time). In this example, two random samples were selected, one for the day of the week and another for the patients to be tracked within a selected day. By using the pieces of paper and a bowl, we were able to apply the principles and precision of probability sampling and avoid some of the complexity associated with using random number tables or computer-generated random samples.

Stratified Random Sampling

This method of sampling is not an alternative to simple random sampling but rather a variation on a theme. Simple random sampling assumes that the composition of the total population is unknown. A random selection process is seen, therefore, as the best way to obtain a "representative" sample. The problem is that the very nature of a random selection process could produce a sample that is not truly representative of the characteristics of the total population. This is where stratification comes into the picture. By stratifying the population into relatively homogeneous strata or categories before the sample is selected, you increase the representativeness of the sample and decrease the sampling error. Once the stratification levels have been identified, a random selection process is applied within each stratum. For example, you might stratify a hospital's patients into medical and surgical strata and then sample randomly within each group. This would help to ensure that one group was not over- or underrepresented in the sample. A key point to remember when setting up a stratified random sample, however, is that it requires the knowledge of people who actually work in the process. As subject matter experts,

they can tell you what key stratification categories are relevant. An external statistician who is very skilled in sampling methods, for example, will not have knowledge of the local characteristics that affect the decisions about proper stratification. Bring the subject matter experts together with a skilled statistician and you will be able to set up a good sampling strategy.

Stratified Proportional Random Sampling

In this case, we are going to use the approach outlined for stratified random sampling, but we are going to add another twist. We are going to determine the proportion that each stratum represents in the population and then replicate this proportion in the sample. For example, if we knew that medical services represented 50% of the hospital's business, surgical services represented 30%, and emergency services represented 20%, then we would draw 50% of the sample from medical units, 30% from surgery, and 20% from the ED. This would produce a sample that not only was representative but also proportionally representative. This would further increase the precision of the sample and reduce the sampling error. The stratified proportional random sample is one of the more sophisticated sampling designs. It does require knowledge of the population being sampled, however, as well as having a sufficiently large enough population that as you stratify the population into a variety of categories you will have sufficient numbers in each category to be eligible for sampling. Note that a stratified proportional random sample can be more costly in terms of both money and time.

Systematic Sampling

Systematic sampling offers one of the easiest ways to draw a sample. It consists of numbering or ordering each element in the population and then selecting every *k*th observation after you have selected a random place to start, which should be equal to or less than *k* but greater than zero. For example, if you had a list of 500 medical records and you wanted to pull a sample of 50, you would pull every 10th record. To determine the starting place for the sample, you would pick a random number between 1 and 10. For argument's sake, imagine that when we do this we select the number 6. So to start our systematic sample we would go to the 6th medical record on our list, pick it, and then proceed to select every 10th record after this starting point. Technically, this is known as a systematic sample with a random start (Babbie, 1979, p. 178). The most frequent ways to organize the elements are either alphabetically or chronologically. There are two major advantages of systematic sampling: it is simple and you have to generate only the first random number. This sampling method is what many healthcare professionals think of as a random sample. Although it is a form of random sampling, it does have certain limitations. The major problem with systematic sampling is that you are eliminating chunks of data that could provide knowledge about the process. If, for example, you are selecting every 10th record, you have automatically eliminated from further consideration records 1 through 9. You pick the 10th record, then skip 11 through 19 and pick number 20. The records in between the ones you select will never have a chance of being included in your analysis. If there is something that occurs regularly in the data or something that causes your data to be organized into bunches of, say, seven or eight, then these records would be automatically eliminated from consideration. The other problem I have observed with this form of sampling in healthcare settings is that the people drawing the sample do not base the start on a random process. They merely pick a convenient place to start and then start applying the sampling interval they have selected. This introduces bias and greatly increases the sampling error.

Cluster Sampling

In cluster sampling, the population is divided into mutually exclusive and exhaustive clusters, then a simple random sample is drawn within each cluster. On the surface this approach does

not seem very different from stratification. Cluster sampling differs from stratified random sampling in that cluster sampling seeks to create "bunches" within the population. Sampling in this way is almost always less expensive than simple random sampling (which is not as focused). The other key distinctions between stratification and cluster sampling include the following: (1) with stratified sampling, a sample of elements is selected from within each stratum or category; and (2) with cluster sampling, a sample of stratum is selected. Because the cluster sample is selecting a sample of stratum or categories, it is desirable to have each cluster be a small collection of the population. Cluster samples, therefore, should establish groupings that are as heterogeneous as possible. Stratified samples, on the other hand, attempt to create homogeneous categories (e.g., all medical and all surgical patients).

Another distinction with cluster sampling is that it is typically done with fairly large populations. This method could be applied, for example, to a large system that has 15 to 20 hospitals. Each hospital could be considered a cluster, or they could be grouped into regional clusters. A cluster sample also could be drawn in a large metropolitan area. Instead of looking at individual hospitals or hospital systems, you could divide the metropolitan area into neighborhoods or regions (the clusters) and then sample patients within these regions. In Chicago where I live, for example, it would be possible to divide the metropolitan area into north, south, west, and urban core clusters (east would not work because Lake Michigan is located to the east of Chicago). If we did this we would not be so concerned with the individual hospitals and their organizational affiliations but rather with bundling people together into common geographic areas. With large populations, therefore, cluster sampling can be a very economical approach to sampling. Cluster sampling would not apply to the unit or department level because the population of interest (e.g., all hip or knee replacement patients) is not large enough to permit clusters to be created.

Nonprobability Sampling

Nonprobability sampling is typically used when the researcher is not worried about estimating the reliability and precision of the sample or of generalizing the results to a larger population. This is not to say, however, that nonprobability samples do not serve a useful purpose. More specifically, nonprobability approaches to sampling can be used when:

- Probability samples are either too expensive to collect or too complicated for the question being asked
- There is no need to draw inferences or generalize to larger populations
- There is no need to estimate the probability that each element has of being included in the sample
- There is no need to have assurance that every element (e.g., patient) had an equal opportunity to be included in the sample
- The objective is to conduct an exploratory or descriptive study on an issue or process that has not been studied in detail
- You are testing a potential improvement strategy and want to run a quick pilot study (i.e., sending up a trial balloon to see if it has any hope of succeeding)
- Mechanical selection of the sample is not required; personal judgment and subjective choice are sufficient

The major forms of nonprobability sampling are convenience sampling, quota sampling, and judgment sampling. The basic objective with all of these methods is to select a sample that the researchers believe is "typical" of the larger population. The problem is that there is no way to actually measure how typical or representative a nonprobability sample is with respect to the population it supposedly is representing. In short, nonprobability samples can be considered "good enough samples" (i.e., they are good enough for the people pulling the sample).

Convenience Sampling

As the name implies, convenience sampling is designed to obtain a handful of observations that are readily available and convenient to gather. Convenience sampling is also referred to as "chunk" sampling (Hess et al., 1975, p. 8) or accidental sampling (Maddox, 1981, p. 3; Selltiz et al., 1959, p. 516). A classic example of convenience sampling is found in the "man on the street" interview conducted by TV stations. The local TV channel parks its action-cam van along a busy downtown street at lunchtime. The investigative reporter positions herself strategically and begins to scan the people who walk by. She knows that she needs to get at least four good comments from local citizens (her quota sample), so she eliminates anyone from consideration who looks like they would be (1) uncooperative, (2) argumentative, or (3) too chatty without any substantive sound bites. Then she sees a likely candidate and strikes: "Hi, I'm from Channel 5 News and I'd like to know how you feel about . . . (fill in the blank)." Okay, one down and three more to go (to meet the quota). So the search continues. There is no science behind this type of sampling. It produces a biased sample that is essentially a collection of anecdotes that cannot be generalized to larger populations. In technical terms, this is what is referred to as a convenient quota sample (i.e., I need a quota of four people and I'm willing to take anyone who is convenient and agrees to talk). In the healthcare setting, convenience sampling is used frequently, possibly too often. I have seen it used to pull a convenient sample of medical records, obtain patient satisfaction input (go grab a few people waiting in the ED and ask them how they feel about their wait time), or select a "typical" day to study call button response time. The primary question that someone should ask when a convenience sample is drawn is, "How important is it to know whether the sample of elements we just selected are representative of the larger population?" If the consequences of being wrong do not matter, then the convenience sample might be good enough.

Quota Sampling

Quota sampling was developed in the late 1930s and used extensively by the Gallup organization to gain great recognition as well as ridicule. (See Chapter 3 for additional details on how Gallup benefited from quota sampling in 1936 and then was criticized in 1948 for its failure to predict accurately.) If you ask healthcare professionals to describe quota sampling, they will probably tell you that it is merely a simple way to determine the total minimum number of elements needed in a sample (e.g., we need a quota of 5% of the medical records) or the total minimum amount of data that the team can afford to gather. These two factors, although part of quota sampling, are only part of the picture. Babbie (1979, p. 196) nicely describes the steps involved in developing a quota sample:

1. Develop a matrix describing the characteristics of the target population. This may entail knowing the proportion of males and females; various age, racial, and ethnic proportions; as well as the educational and income levels of the population.
2. Once the matrix has been created and a relative proportion assigned to each cell in the matrix, you collect data from persons having all the characteristics of a given cell.
3. All persons in a given cell are then assigned a weight appropriate to their proportion of the total.
4. When all the sample elements are so weighted, the overall data should provide a reasonable representation of the total population.

Theoretically, an accurate quota sampling design should produce results that are reasonably representative of the larger population. Quota sampling has several inherent problems, however, that are related primarily to how the cells in the quota matrix are actually populated. If, for example, the individuals collecting the quota samples are not particularly vigilant and honest about filling their

quotas, the results will be biased. Remember, the actual selection of the elements to fill the quota is left up to the individual gathering the data, not to random chance. If the data collectors are not diligent and/or honest about their work, they will end up obtaining their quotas in a manner that is more like a convenience sample than a true quota sample. This happens frequently when quota samples are being collected in neighborhoods. The 2000 census in Chicago provided a good example of this type of bias. The census workers were given quotas to fill on the North Shore of the city. This is a rather wealthy area where it is not uncommon to find homes that are gated and monitored by security. Many of the census workers were not given access to these homes, even though they were technically in the cell they were supposed to obtain. Apparently pressured by the requirement to meet their quotas, the census workers creatively began to substitute other residents for the ones defined by the quota sample. As a result, the cells in question (i.e., neighborhoods) were underreported and not properly representative of the area (*Chicago Tribune,* July 5, 2000, "Census Shortcuts Alleged"). Another threat to the validity of the quota sample is that the patient population characteristics might be outdated and not reflect the current patient population. The final threat involves the process by which the data collectors actually gather the data. For example, if a quota sample was established to gather data in the ED but only during the day shift, you would run the risk of missing key data points during the afternoon and evening shifts.

Judgment Sampling

I saved the discussion of judgment sampling until the end because it can be viewed in two very different lights. If you approach sampling from an academic research perspective, then judgment sampling is regarded as having a low level of precision and statistical rigor. If, on the other hand, your objective is not academic research but rather QI research, then judgment sampling provides a useful approach to sampling. The academic view that judgment sampling (also referred to as purposive sampling) has a low level of precision is based on the fact that the sample is drawn on the basis of the knowledge of the person(s) drawing the sample. No objective mechanical means are used to select the sample. The assumption is that experience, good judgment, and appropriate strategy can select a sample that is acceptable for the objectives of the researcher. An example of judgment sampling is seen every 4 years when a handful of states and communities are selected to be "pulse checks" for the U.S. presidential election. In this case, the assumption is that the people in Iowa and New Hampshire are "typical" of the rest of the nation and that the responses of these citizens provide a snapshot of how the average American views the presidential candidates. Obviously the major challenge to judgment sampling is related to the knowledge and wisdom of the person making the judgment call. If everyone believes that this person exhibits good wisdom, then they will have confidence in the sample that the person selects. If, on the other hand, people doubt the person's wisdom and knowledge, then the sample will be discredited.

Now, consider the nonacademic use of judgment sampling. Deming considered judgment sampling to be the method of choice for QI research. Langley, Nolan, Nolan, Norman, and Provost, (1996, p. 111) maintain that "A random selection of units is rarely preferred to a selection made by a subject matter expert." In QI circles, this type of sampling is also known as expert sampling or rational sampling. It essentially consists of having those who have expert knowledge of the process decide on how to arrange the data into subgroups and pull the sample. The subgroups can be elected either by random or nonrandom procedures, which is a major distinction between the QI perspective and the academic view of judgment sampling. The other important distinction about Deming's view of judgment sampling is that the samples should be selected at regular intervals over time, not at a single point in time. Most sampling designs, whether they are probability or nonprobability, are static in nature. The researcher decides on a time frame then picks as much data as possible. In contrast, Deming's view of sampling was that

it should be done in small doses (rather than large quantities) and pulled as a continuous stream of data (Deming, 1950, 1960, 1975). The primary criticism of judgment sampling is that the "expert" may not fully understand all facets of the population under investigation and may therefore select a biased sample. The second criticism is that the sampling error cannot be measured. The final challenge is that the results of a judgment sample cannot be generalized to the larger population because the sample was not selected by random methods.

A review of the probability and nonprobability sampling methods is provided in **TABLE 4-5**. Developing a working knowledge of these sampling techniques will be one of the best ways to reduce the time spent on collecting data. Done correctly, sampling will also be a way to ensure that the data you do collect are directly related to your QI efforts.

Now that you are familiar with the principal approaches to stratification and sampling, it is time to start applying these techniques to your own set of indicators. **TABLE 4-6** provides

TABLE 4-5 Advantages and disadvantages of various sampling methods

Sampling Method	Description	Advantages	Disadvantages
Probability Sampling			
Simple random sample	A sample that is drawn in such a way that every member of a population has an equal chance of being included. A random number table or a random number generator is typically used to actually pull the sample.	■ Requires minimum knowledge of the population in advance ■ Free of possible classification errors ■ Easy to analyze the data and compute errors ■ Fairly inexpensive	■ Does not take advantage of the knowledge the researcher might have about the population ■ There could be over- or underrepresentation of subgroups within the population ■ Typically produces larger sampling errors for the same sample than a stratified sample
Stratified random sample	The population is divided into relevant strata before random sampling is applied to each stratum.	■ Helps to reduce the chances of over/underrepresenting subgroups within the population ■ Allows you to segment the data into "buckets" during the analysis phase ■ Create more efficient samples ■ Reduces sampling error	■ Requires knowledge of the presence of various characteristics within the population ■ Sampling costs can increase if knowledge of the population is shallow ■ If the strata are not highly homogeneous then sampling error goes up and efficiency goes down

(continues)

TABLE 4-5 Advantages and disadvantages of various sampling methods *(continued)*

Sampling Method	Description	Advantages	Disadvantages
Proportional stratified random sample	The proportion (or percentage) of a particular stratum is determined in the population and then applied to the random sample.	■ Adds even more precision than the stratified random sample ■ Increases sample representativeness ■ Creates very efficient samples ■ Reduces sampling error	■ Requires more human and financial resources than other methods ■ Requires even more information about the population than stratified random methods
Systematic sample	Select every *k*th observation from the population after a random starting point has been selected.	■ Very easy to conduct ■ Has "intuitive" appeal ■ Inexpensive to conduct	■ Can produce bias due to periodic ordering of observation, which produces exclusion of segments of the population ■ Increased probability of sampling bias
Cluster sample	Clusters or "bunches" of the population are identified, and then random sampling is applied to each cluster.	■ Can be low cost, especially if geographic clusters are used ■ If properly done, each cluster is a small model of the population ■ High level of practicality	■ Clusters need to be as heterogeneous as possible ■ Typically has lower statistical efficiency ■ Large samples are often needed to ensure precision
Nonprobability Sampling			
Convenience sample	Observations are selected based on availability and convenience. Also known as "accidental" samples.	■ Ease of obtaining a sample ■ Relatively low cost	■ Extremely low generalizability ■ No way to determine sampling bias or sampling error

TABLE 4-5 Advantages and disadvantages of various sampling methods *(continued)*

Sampling Method	Description	Advantages	Disadvantages
Quota sample	A population is divided into relevant strata. The desired proportion of samples to be obtained from each stratum is determined, and then a fixed quota within each stratum is set.	■ Stratification effect is achieved if the strata are appropriately structured ■ In theory, the quota sample should be reasonably representative of the population ■ Human and financial costs can be kept to a minimum if the strata from which the quotas are to be drawn are grouped close together (reduced the amount of travel the data collectors have to perform in order to gather the data)	■ The people assigned to collect the quotas need to be scrupulous and free from selection bias and follow the prescribed sampling design (otherwise this method becomes a convenience sample) ■ It is difficult to guarantee that the quotas were filled accurately ■ In-depth knowledge of the population is required ■ Nonrandom selection of the quotas can also introduce bias
Judgment sample	Subgroups are drawn from a process over time based on expect knowledge. The subgroup samples can be drawn either by random or nonrandom procedures.	■ Samples in a subgroup can be small (3–5) because many subgroups will be selected ■ Data collection costs can be reduced ■ Provides a dynamic picture of the data and serves as the basis for process improvement ■ Minimum stratification effect is achieved	■ Sampling bias and sampling error cannot be calculated ■ Expert knowledge of the process or population is required ■ Generalization of the judgment sample to larger populations cannot be done ■ Personal bias enters into the selection of the sample

a Data Collection Plan Worksheet designed to help you clarify the data collection plan for your improvement team. For each indicator that you identified earlier in this section (see Table 4-3, the Organizing Your Indicators Worksheet), outline the decisions your team has made related to stratification of the data, sampling (if appropriate), and frequency and duration of data collection. The frequency of data collection addresses how often you plan to collect the data. Will you, for example, collect the wait time of every patient or develop a sampling strategy?

TABLE 4-6 Data collection plan worksheet

Team Name and Improvement Topic: _____

Indicator Name	Is stratification appropriate? If Yes, list the levels of stratification	Will you use sampling? If Yes, describe the sampling method you will use	Frequency of data collection (e.g., hourly, daily, weekly?)	Duration of data collection (i.e., how long do you plan to collect the data?)

BOX 4-3 Conclusions about data collection

1. Sampling should produce representative and workable numbers for the unit of interest.
2. Customers providing feedback about their service or care they receive can be very susceptible to sampling bias (sampling and recall biases).
3. Sampling bias can be introduced if you always use the same place or time and this is not representative of the whole. This is a major problem when single point in time audits are relied on as the sampling method.
4. When conducting surveys recall bias occurs if the questions are reliant on the individual's memory.
5. The worst case scenario occurs when you have no idea where the sample came from or how representative it is of the population or organization overall.
6. Clear guidance on data collection methods, in particular sampling and stratification, are required whether you are collecting data for improvement, judgment, or research.
7. Data on "why did this happen" are critical to improvement efforts!

Will you collect the wait time of all patients but only on Mondays? These questions relate to the frequency of data collection (i.e., how often do you need to dip into the ongoing stream of data to gain an adequate understanding of the variation in the indicator?). The duration issue deals with how long you plan to collect the data. Will you do it for a week, a month, or several months? If you do not spend time discussing the frequency and duration questions, you will inevitably come to a point when someone says, "How long do I have to collect this stuff?"

Summary conclusions about key data collection issues are provided in **BOX 4-3**.

The Indicator Development Worksheet

Now that we have reviewed the individual milestones in the QMJ it is time for you to organize your indicators into a coherent roadmap. The Indicator Development Worksheet is shown in **EXHIBIT 4-1**. It provides a practical and convenient way for a team to organize the details for one specific indicator. If you can provide responses to all of the items on this worksheet you will have an indicator that, at least for the short term, will enable you to proceed with data collection and eventually analysis. The details related to each section on the Indicator Development Worksheet are provided next.

1. *What is the overall **AIM** of this improvement initiative?*

   We addressed the particulars of building an aim statement earlier in this chapter. It is important, however, to make sure that the overall aim is stated when developing specific indicators so that the team is able to clearly see how the indicator is linked to what the team is trying to accomplish. How good do you want to be? And by when do you expect to achieve the outcome? These two simple questions are essential to a team's journey. Again, it does not have to be a long and detailed statement. One or two sentences should suffice.

2. *What is the **NAME** of this **SPECIFIC INDICATOR?***

   Naming indicators is an important component of indicator development that is frequently taken for granted. Some might ask, "What's the big deal? Just give it a name." Indicator names should be objective, and they should reflect quantifiable nouns. Often, however, teams feel the need to include adjectives and adverbs as well as targets and goals in the indicator name (e.g., history and physical transcription TAT will be 12 hours or less). This produces what I call "thou shalts" (i.e., thou shalt perform this task in 12 hours or less or there will be consequences). Indicators named in this fashion identify the desired outcome. When you include the desired level of performance in the indicator name, you have basically built in a barrier or, worse yet, a threat. It sends a message to the workers that you had better do this or else. If the desired outcome is an unrealistic goal, the workers quickly figure this out. The indicator then becomes an unrealistic metric or a joke. Consider this example: I was working with a medical group on their indicators and asked a team what they intended to measure. A member of the team said that the indicator was, "No one should have to wait more than 30 minutes to see the doctor." My comment to the team was that this was not the name of an indicator but rather a threat. The indicator name should have been "wait time to see the doctor." Although it seems like a minor aspect of performance measurement, I believe that the naming of indicators sets the tone for the rest of the measurement journey. It is the point at which you leave Conceptland and actually enter into Measurementland.

3. *What **TYPE of INDICATOR** is this?*

   It is important to be clear about the types of indicators being developed. Teams need to be careful to not develop so many indicators that they become over burdened with collecting data. Most improvement teams will be tracking one to three outcome measures, four or

EXHIBIT 4-1 Indicator Development Worksheet

Team name and topic of interest: _____
Date: _____ Contact person: _____ Email _____

1. What is the overall AIM of this improvement initiative?
   (How good do you want to be? By when do you expect to achieve the outcome?)
2. What is the NAME of this SPECIFIC INDICATOR? (e.g., the number of x-ray exams performed, the percentage of x-ray reports that could not be found, the medication error rate or the days between a patient fall).
3. What TYPE OF INDICATOR is this?
   _____ Outcome _____ Process _____ Balancing
4. What is the OPERATIONAL DEFINITION for this indicator?
   Define the specific components of this indicator. Specify the numerator and denominator if it is a percentage or a rate. If it is an average, identify the calculation for deriving the average. Include any special equipment needed to capture the data. If it is a score (such as a patient satisfaction score) describe how the score is derived. When an indicator reflects concepts such as accuracy, complete, timely, or an error, describe the criteria to be used to determine "accuracy."
5. What is your DATA COLLECTION PLAN?
   - How frequently will the data be collected?
     _____ Every Patient_____ Hourly_____ Daily _____Weekly _____Monthly _____Other (please specify)
   - What are the data sources to be used for this indicator (be specific)?
   - What is to be included or excluded (e.g., only inpatients are to be included in this measure or only stat lab requests should be tracked).
   - How will these data be collected?
     _____Manually _____From a logbook _____From an automated system
   - Who will be responsible for the actual collection of the data?
     _____
   - Will you use stratification? If "Yes" specify the stratification levels you will use.
   - Will you employ sampling? If "Yes" specify your sampling strategy you plan to use.
6. Do you have BASELINE DATA for this indicator? _____Yes _____No _____Unknown
   - What is the actual baseline number? _____
   - What time period was used to collect the baseline? _____
7. Are there TARGET OR GOALS for this indicator?
   Internal target(s) or goal(s)? _____Yes _____No _____Unknown

   If "yes" please list the actual internal target or goal (e.g., the number, rate, or volume, etc., as well as the source of the target/goal)

   _____

   External target(s) or goal(s)? _____Yes _____No _____Unknown

   If "yes" please list the actual internal target or goal (e.g., the number, rate, or volume, etc., as well as the source of the target/goal)

   _____

five process measures, and one or two balancing measures (if appropriate). As mentioned earlier in the Types of Indicators section of this chapter, the objective should be to identify a reasonably small set of indicators (six to eight total) that capture the critical aspects or dimensions of the team's aim (i.e., build a balanced set of indicators).

4. *What is the **OPERATIONAL DEFINITION** for this indicator?*

   This is essentially the heart and soul of the Indicator Development Worksheet. In this section, you should provide detail on the components of the operational definition in very specific terms. If it involves a percentage, then the numerator and denominator should be described. Similarly, if the measure is a rate, then the rate-based statistic should be defined (i.e., falls per 1,000 patient days). The easiest way to do this is to take the indicator name (e.g., inpatient fall rate) and then say, "Inpatient fall rate is defined as . . .(fill in the blank)." If it is an average, identify the calculation for deriving the average. Include any special equipment needed to capture the data. If it is a score (such as a patient satisfaction score) describe how the score is derived. When an indicator reflects concepts such as accuracy, complete, timely, or an error, describe the criteria to be used to determine "accuracy." Remember to describe what is to be included (e.g., all inpatients, including pediatrics, and geriatrics and falls in the ED) and what is to be excluded (e.g., visitor falls in and out of the hospital, staff falls, and falls in the rehabilitation unit). The litmus test for a good operational definition is really quite simple. Just ask yourself, "How could someone get confused with this definition and collect wrong data?" If you have written a clear and unambiguous operational definition, then you will be able to avoid confusion during the data collection stage.

5. *What is your **DATA COLLECTION PLAN?***

   The items listed in this section require knowledge, skill, and experience with data collection practices for QI, not data collection practices for conducting randomized clinical trials (RCTs) or traditional academic research but QI research. A key difference between data for QI and data for RCTs or more traditional academic research is that QI data will typically be collected more frequently and in smaller doses than traditional research. The questions related to data collection are listed in the worksheet but a few suggestions on data collection are listed here.

   - **Collection Frequency and Duration.** Remember that frequency deals with how often you plan on collecting the data, whereas duration addresses the question of how long you will continue to collect the data. Will you collect the wait time of every patient or develop a sampling strategy? Will you collect the wait time of all patients but only on Mondays? Next focus on the duration of data collection question. How long do you plan on collecting the data? Will you do it for a week, a month, or several months? Frequency and duration are critical issues for the team to address and they need to be clarified **before** you actually start to collect the data.

- **Data Sources.** Where do you plan on getting the data? Will they be manually collected, or will they come from an automated system? If it is to be a manual process, will they come from existing log sheets or the medical record, or do you have to create a new data collection tool? If they come from an automated system, what segment of the automated system will be used (e.g., is it the registration system, the billing system, or the patient satisfaction system)? A related issue is by what method do you propose to actually gather the data? For example, if you are tracking lab TAT, will the recorded time come from the watch of the individual who is recording the log-in time, the clock on the wall by the door, or the automated time stamp produced by the computer system? If the data are to be collected manually, do you have a procedure outlining how the person recording the data is supposed to identify the particular piece of data (e.g., the log-in time to the lab of a specimen) and then enter it into the logbook? For some of you this probably seems like a very left-brained, compulsive set of questions. If someone does not attend to the details, however, they will be ignored. This is why you need both left- and right-brained people on QI teams. A team with all left- or all right-brained people usually does not achieve as much as teams that have a mix of perspectives. Sometimes you need vision and creativity, and sometimes you need structure and attention to details.

- **Person(s) Responsible for Collecting the Data.** It is not uncommon to arrive at this step in the process and realize that you have not made any provisions for actually collecting the data. Everyone seems to assume that someone else will do the work of recording wait times or extracting documentation history from the medical records. Someone has to do the data collection. Frequently, however, there is a wonderful chain reaction when it comes to this task. Physicians assume that the nurses will collect the data. The nurses assume that the unit secretaries will complete this task. The unit secretaries hope that several administrative interns will be assigned to the department for the summer and this job can be pawned off on them. If all else fails, the volunteers can be asked to do the data collection. Now think about this. If you have spent considerable time building the indicator's operational definition and data collection plan, why would you assume that some undetermined person will magically appear and solve all your data collection problems? In Greek dramas, this solution was referred to as the *deus ex machina* (the god from the machine) or the unexpected solution to a difficult problem.[11] If this aspect of the measurement, journey is not determined prior to the collection of data, I can guarantee you that (1) it will not go smoothly and (2) the data gathered will be questionable.

- **Stratification and Sampling.** Considerable detail on stratification and sampling has already been provided in this chapter. It is essential, however, that the team devote time to discussing the relevance of these two extremely important data methods to their data collection plan. Both are extremely important from a practical as well as resource perspective. As was mentioned, they are less about statistical issues than logical issues and they can be properly decided upon only by those who have subject matter experience.

6. *Do you have **BASELINE DATA** for this indicator?*

   Remember that the baseline is what the current process is producing. It is not what you want it to be or expect it to be. Baseline is a fundamental concept in medicine. We get a baseline on a patient before we start to prescribe medications or treatments. In a similar fashion, we basically want to know how the indicator we are tracking is performing and what it is capable of producing under current operating conditions. Also note that targets and goals should not be established without having a clear understanding of what the current baseline is. Frequently, however, teams set targets and goals without knowing the current performance, which typically leads to the establishment of arbitrary and capricious targets or goals that have little chance of being reached.

7. *Are there **TARGETS or GOALS** for this indicator?*

   This is where you identify what you want or expect the indicator's performance to be. Targets are usually seen as short-term objectives (several months to a year). Goals, on the other hand, are usually designed for a little longer period of time, say 2–3 years. I have seen numerous examples, however, of how organizations have confused staff by not being clear on whether the new number was a target or a goal. Frequently, people use the terms as if they were synonymous, but note that targets are typically more short term (e.g., weeks or months) whereas goals are more long term (e.g., a year or longer). The critical points are that you (1) develop targets and/or goals that are reasonable and achievable and (2) have a plan for how the targets and/or goals are to be achieved. As Deming pointed out frequently in his seminars and writings, "Goals are necessary for you and me, but numerical goals for other people, without a road map to reach the goal, have effects opposite to the effects sought" (Deming, 1992, p. 69). The simpler version of Deming's basic challenge was "By what method?" That is, by what method do you propose to achieve this target or goal?

   Another key aspect of targets and goals is whether they are established internally or externally. Increasingly, healthcare targets and goals are being established by external bodies and given to providers. In this case, however, the external bodies may not fully understand the capability of the existing processes to achieve the targets or goals. This can lead and has led to considerable stress and challenge for healthcare providers.

   The final aspect of setting targets and goals is one that has fascinated me for years. My experience has

convinced me that nearly all targets and goals are established as whole numbers that are divisible by 5. The next time you are in a meeting where targets and goals are being discussed test my theory. You will generally find that people start a discussion by saying "I think we need a 10% reduction in X." This will be followed by someone else saying, "No, I think we should shoot for a reduction of 15%." If 15% is not accepted as the target someone else will then raise the ante to 20% or even 25% . . . and so on. If you want to offer a disruptive innovation into a meeting propose that the next target or goal be 11.58% not 15 and see what happens. I did this once in a management meeting and a majority of the attendees looked at me in a very strange way. One of them even said, "Are you trying to be a wise guy? Why would we have a target of 11.58%?" I smiled at the fellow and asked him why he thought whole numbers divisible by 5 provided a realistic foundation for the target. He hesitated for a moment the responded, "Well, that is how we always set targets." QED!

CASE STUDY #1: *Transcription Turnaround Time*

The following case study is designed to demonstrate how the milestones addressed in this chapter can be applied to an improvement challenge: reducing transcription turnaround time (TAT) for histories and physicals (H&Ps).

Situation

Imagine that you are the director of QI at a medium-sized hospital. One morning you receive a call from your friend Becky, the manager of medical transcription, who asks if she can meet with you ASAP. You sense that she is bothered by something and tell her that you will come to her office in an hour. You have known Becky for more than 8 years and are a little surprised when she dispenses with the usual pleasantries and jumps right into her concern. She tells you that several physicians have been complaining recently about transcription TAT for H&Ps. She even confides in you that she actually had a rather "energetic" exchange with the head of surgery about this issue earlier in the morning. Now Becky is asking for your help, as she explains, "to get the doctors to realize that TAT is actually very good." To "prove" her point she shows you **FIGURE 4-7**. She points out quickly that the goal is to get the transcriptions completed in 12 hours or less. Becky feels that the graph demonstrates that over the last 10 months the transcription team has been able to meet this goal 92–99% of the time. She asks, "So what is the big deal with the physicians? Why are they complaining?"

Diagnosis of the Problem

As you look at the title of Figure 4-7, the first thing that strikes you as confusing is that the indicator of interest is transcription TAT, yet Becky has presented the data as the percentage of H&P transcriptions completed within 12 hours. You also point out that she has actually violated one of the basic principles of naming an indicator—she created a "thou shalt" (H&P transcription TAT will be 12 hours or less). Her response is, "We have to have a goal; otherwise, the doctors will

(continues)

CASE STUDY #1: Transcription Turnaround Time (continued)

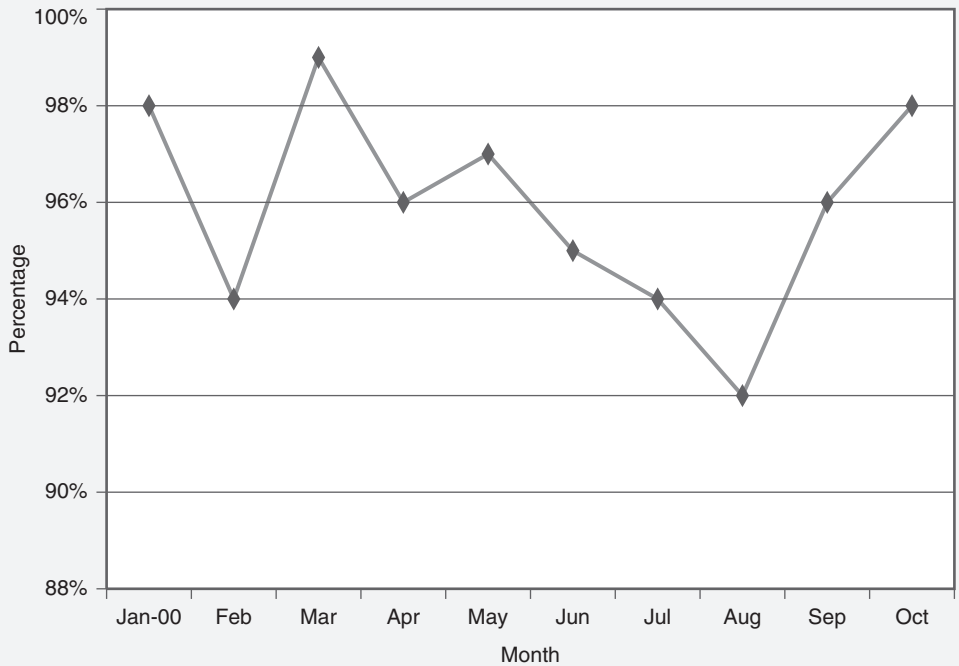

FIGURE 4-7 Percentage of history and physical (H&P) transcription turnaround times completed within 12 hours

not take us seriously." You respond by saying that goals are critical to improvement initiatives but that they should not be built into the name of the indicator. Indicator names should be objective statements about what is to be measured, not what you want it to be. You also point out that by measuring TAT as a percentage, you are losing information about the true variation in the process. For example, consider the month of August in Figure 4-7. During this month, 92% of the transcriptions were completed in 12 hours or less. What was the variation in the TATs for this month? What was the shortest TAT and what was the longest TAT in this month? You cannot answer this question because the measure is whether the transcriptions were completed in more than 12 hours or less than 12 hours. This is a yes or no type of indicator. The longest TAT might have been 13 hours or 40 hours. But because of the way in which the indicator was developed, you will never know the answer to this question about variation. You suggest that it would be useful, therefore, to look at the actual time (in hours) to transcribe H&Ps. Becky is so desperate for help that she agrees to do this even though she is not sure why.

You also point out that there is no necessary reason why this indicator is structured around monthly data. Becky has shown you only 10 months of data. This is the minimum amount of data for a run chart (i.e., 10 data points) but you are thinking that there is probably a sufficient amount of data that it could be broken down into smaller subgroups that would allow the construction of a Shewhart chart.[12] So, you ask, "Why can't you display this indicator by week or every 2 weeks?" Becky's response is one you have heard many times before: "Well, there are several reasons. This is how we have always

(continues)

CASE STUDY #1: *Transcription Turnaround Time* *(continued)*

TABLE 4-7 Summary data by month for H&P turnaround times

Month	Number of H&P Reports	Total Hours TAT*	Average TAT Hours
January	500	5,140	10.3
February	487	5,734	11.8
March	498	4,948	9.9
April	521	6,024	11.6
May	517	5,882	11.4
June	508	5,913	11.6
July	489	5,756	11.8
August	501	6,031	12.0
September	520	5,850	11.3

Goal = 12 hours *Hours are rounded up to the nearest hour

collected this data, it is how the tracking system is set up and how management expects to see the data reported." You explain that by moving to more frequent subgroups such as by week or every 2 weeks, a more detailed picture of the true variation in the process will be obtained. Becky is not quite sure what you are talking about but she is willing to bring you the detailed data.

The next day Becky comes to your office with **TABLE 4-7** and **FIGURE 4-8**. Table 4-7 shows the number of H&P reports completed each month, the total hours required to complete the transcriptions, and the average hours. Figure 4-8 presents these data as a line graph with the 12-hour goal as a reference line. Becky is quick to point out that this chart clearly "proves" her point. All of the average TATs she says, are below the goal of 12 hours. So, again, why are the doctors upset? What would be your next set of recommendations for these data?

Recommendations

As you look at Figure 4-8 you notice what you believe might be causing the physicians to complain about TAT. You see from the title of this chart that the process starts with dictation and ends when transcription is completed. You ask Becky a simple question: "What do the physicians care most about?" She looks at you in a quizzical manner and says, "I don't understand your question." You proceed to explain that the physicians probably do not care very much about the time from dictation to transcription but rather the time from dictation to posting the results on the

(continues)

CASE STUDY #1: *Transcription Turnaround Time* *(continued)*

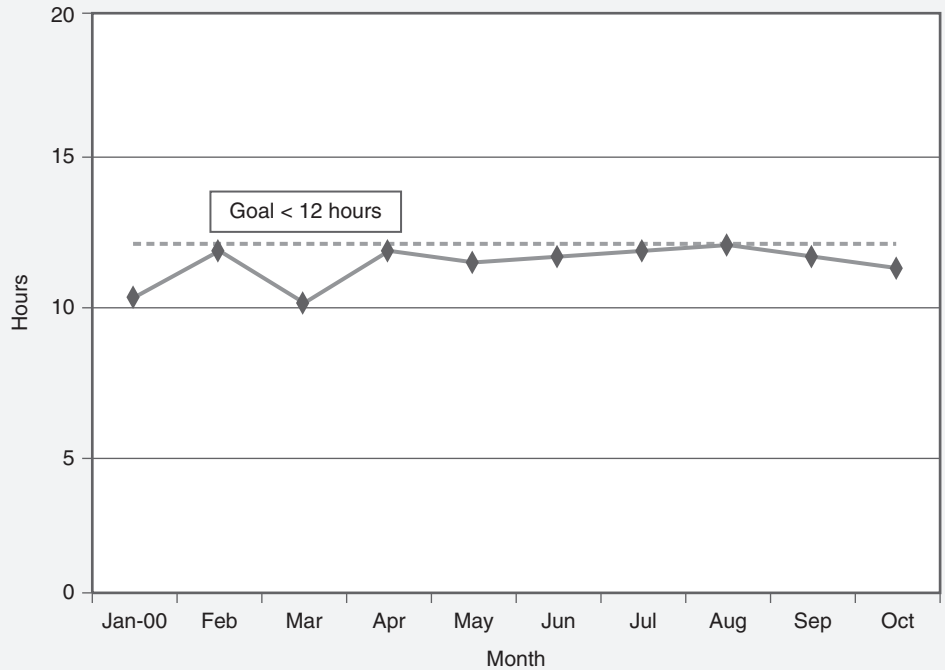

FIGURE 4-8 Average H&P transcription turnaround times from dictation to transcription

computer or finding the results in the chart. Becky is very quick to point out that "now you're being unreasonable."

She continues by explaining that they have never looked at the end piece of the process (i.e., transcription to posting of the results) because that part of the process has never been very good, and it is something they do not feel that they have much control over. She concludes her comments by stating, "We look at what we do and expect someone else to take care of the rest of the process by posting the results in a timely manner. It is not my problem that the people who are supposed to post the results don't do it according to the doctor's demands." You politely point out that the customers apparently do believe that the transcription department is responsible for making the posting of results happen in a timely manner. With reluctance Becky agrees to look at the process from dictation to charting, but she points out that there is no way that they can look at all the cases for these new starting and ending points. At this point you are glad she has raised this issue because it provides you with an opportunity to return to the topic of the monthly data collection. Your guidance is as follows:

- Upon looking at Table 4-7 you explain that having roughly 455 transcriptions each month provides ample reason to make smaller subgroups.
- With this many transcriptions each month it would be possible to make a subgroup for every 2 weeks of data (with roughly 226 transcriptions being received every 2 weeks). Or if week was

(continues)

CASE STUDY #1: *Transcription Turnaround Time* *(continued)*

selected as a subgroup then there would be about 114 transcriptions received each week. Finally with this amount of data, day could be used as the subgroup, which would provide approximately 15 transcriptions being received each day. The problem with looking at the data by day, however, would be that this would produce over 300 individual data points on a chart for the 10-month period, which is clearly not necessary.

- In discussing these options with Becky, you reach agreement that looking at the data by week makes the most sense. This will produce 40 data points on a chart, which is not excessive. You go a step further to point out to Becky that analyzing all the data in a week (approximately 114 transcriptions each week) is also not necessary. A smaller subset of transcriptions (approximately 15–20 each week) can be selected by developing a sampling strategy. Becky agrees to help you in implementing a sampling strategy on future data but asks that you use the existing data for 10 months to help her think through her current challenge of addressing her customers' dissatisfaction with the TAT.

When Becky returns to your office on Monday, she seems to be somewhere between shock and embarrassment. When you look at **FIGURE 4-9** you start to understand why she is in such a state of mind. It shows that the average TAT for the transcription process, dictation to posting on the patient's chart, is somewhere around 18 hours—much higher than the goal of 12 hours discussed earlier. When you ask Becky about these results, she says that they are very surprising. Because you know her well, however, you quickly figure out that she is not being totally honest with you. With a little probing she admits that she knew this all along and that this is why she decided to focus only

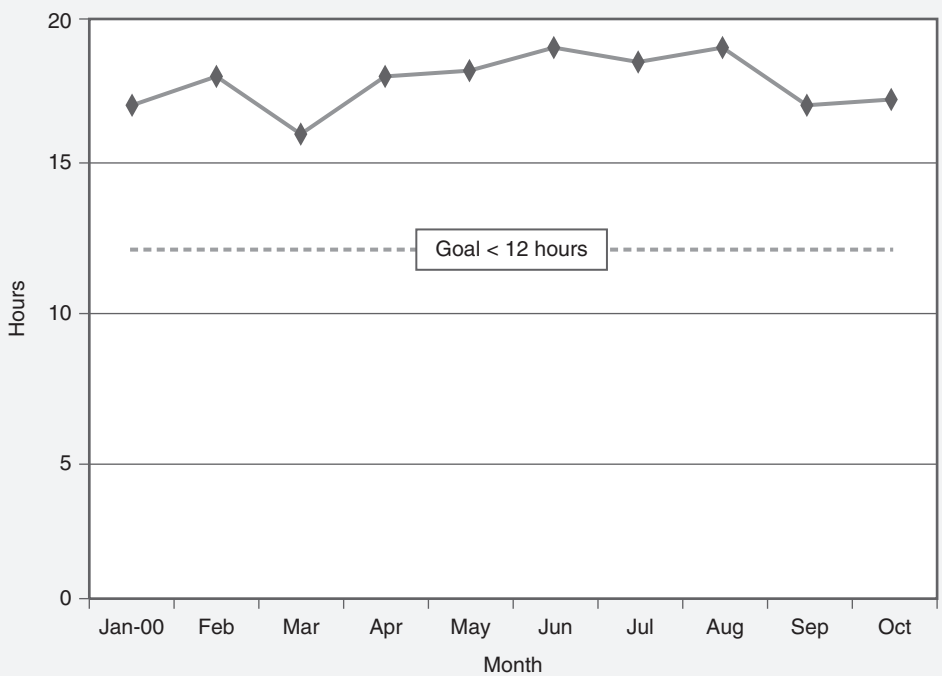

FIGURE 4-9 Average H&P transcription turnaround times from dictation to chart

(continues)

CASE STUDY #1: *Transcription Turnaround Time* *(continued)*

on the dictation-to-transcription part of the process (this is where the embarrassment starts to settle in). So, you offer another quality measurement insight. Specifically, you ask if this process needs to be stratified. Becky is not quite sure where you are going with this question. You explain that it might be possible that the transcription process varies by day of the week, time of day when the dictation is received, or possibly by type of procedure. As you explore this idea with her, she indicates that it is possible that the TAT could vary by type of patient, namely, nonsurgical versus surgical H&Ps. When she says this you notice a slight hesitancy in her voice, but you are not sure why.

To your surprise, Becky returns to your office later that same day. She now has two new charts. **FIGURE 4-10** shows the TAT for nonsurgical patients and **FIGURE 4-11** shows the results for surgical patients. As soon as you look at the charts you realize why Becky was looking a little embarrassed and sounded hesitant earlier in the day. The answer to her problem is obvious. There is not one transcription TAT process occurring here but two. There is one process for nonsurgical patients and another for surgical patients. The nonsurgical patient TATs are turned around in roughly 21 hours whereas those for the surgical patients have TATs of about 5 hours. This spread in the two processes helps to explain why the average for all patients is about 18 hours (Figure 4-9): it reflects the average of two very different processes.

So why the big discrepancy? It turns out that two of the surgeons were very upset with Becky and her department about 8 months ago. They complained not only to the president of

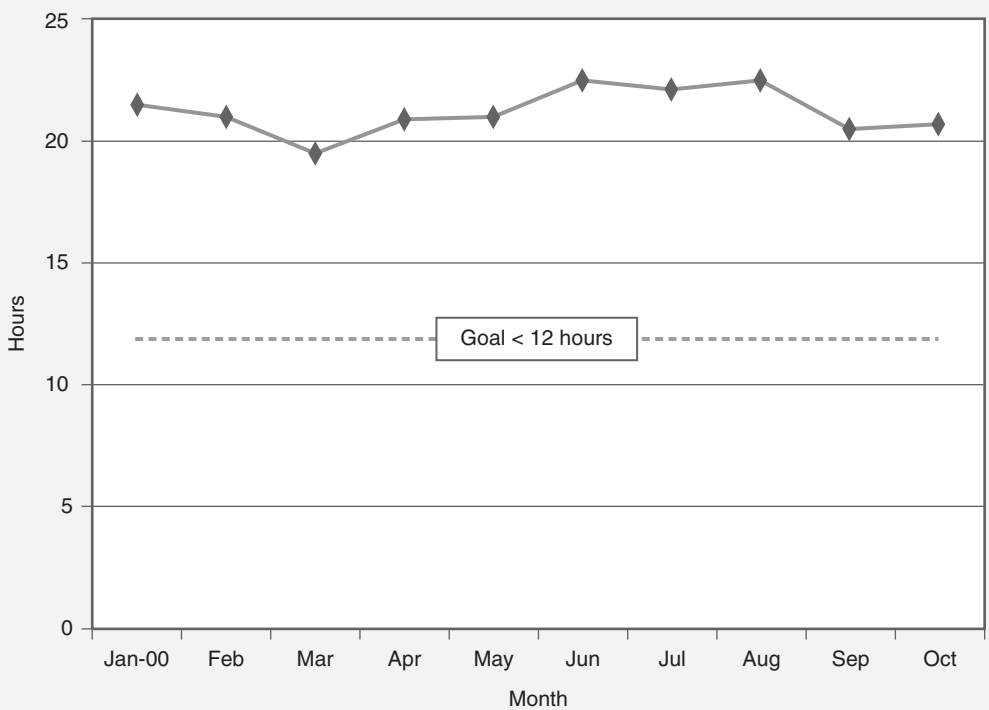

FIGURE 4-10 Average H&P transcription turnaround times for nonsurgical patients from dictation to chart

(continues)

CASE STUDY #1: *Transcription Turnaround Time* *(continued)*

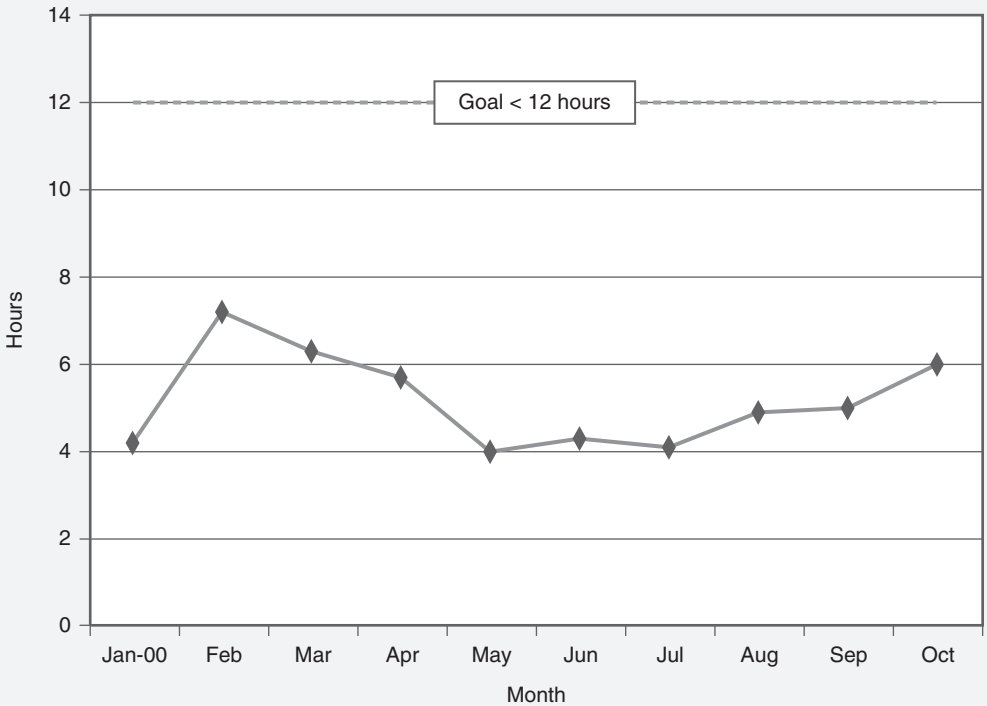

FIGURE 4-11 Average H&P transcription turnaround times for surgical patients from dictation to chart

the medical staff but also to the CEO about the TAT. Becky said she still remembers vividly the discussion she had with the CEO when she was called to her office. In order to avoid that situation again, Becky made sure her staff gave preferential treatment to any H&P related to surgical procedures. As a result, the nonsurgical H&Ps received less attention than the surgical patients, as demonstrated in Figures 4-10 and 4-11.

The Improvement Plan

Once Becky was over her cathartic release of pent-up angst, she was able to acknowledge readily that she and her department set themselves up for this outcome. Obviously, one objective will be to maintain the performance of the surgical H&P TAT process. The challenge will be to work on the nonsurgical process and take steps to reduce this time to meet the goal of 12 hours or less. You agree to assist Becky by helping her (1) create an improvement team to work on the TAT process, (2) serve as the improvement advisor for the team, (3) guide them through the next round of data collection and analysis, (4) develop a stratified proportional random sample for a weekly subgroup of 15–20 transcription times, and (5) create a Shewhart chart that is appropriate for understanding the variation in TAT (i.e., an X bar and S chart described in Chapter 9). Becky leaves your office not only feeling better about revealing the entire story behind this issue but also motivated to work with her team to improve the nonsurgical process.

Notes

1. An excellent book on this topic is *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives* by Stephen Ziliak and Deirdre McCloskey (The University of Michigan Press, 2011).
2. This is actually the first question in IHI's approach to improvement. It is part of a larger framework called the Model for Improvement (MFI), which was developed by Associates in Process Improvement (API). The MFI is detailed in *The Improvement Guide* (Langley et al., 2009). IHI has adopted the MFI as its overarching approach to improvement. More will be said about the MFI in Chapter 9.
3. The higher up individuals go in the food chain the more they seem to live in Conceptland. Board members, senior leaders, political leaders, and even the media frequently live in or have time-shares in Conceptland. This is not necessarily bad. These folks are supposed to set direction and provide visions of what can be achieved. The problem is that if an organization does not have people skilled in the science of improvement who can move the organization out of Conceptland and into Measurementland the visions will never be achieved.
4. When teams get stuck in Conceptland and have vague notions of what they want to improve it reminds me of the famous statement often credited to Horace Greeley (or as the story goes possibly to John Babson Lane Soule): "Go West young man." A general direction is offered but no specific aim or milestones are given as to where in the West you should end up. So the journey stays essentially in Conceptland.
5. Knowing the difference between a proportion, a percentage, and a rate is critical. Often in healthcare settings people refer to a rate when they are really referencing a percentage. Additional detail on these distinctions is provided in Chapter 6.
6. Increasingly, politicians are being asked to clarify what they say and what they mean. It is not uncommon, for example, for a political candidate or an incumbent to have a reporter repeat what the politician said yesterday or last week about an issue and then ask them why they are now saying something different. The evening news is notorious for playing a clip of what the president said last month and then showing that he said just the opposite today. Think how easy our nation's founders had it in this regard. It was probably very easy for them to take a position one day and another position the next. No one recorded their comments verbatim, and there were no cameras, videotapes, or recording devices. Operational definitions in George Washington's day could be very loose. Today, however, there is increasing scrutiny on the part of the public and a desire to have the politicians be more precise in their definitions of terms.
7. You will need to be clever in how you actually fit this into a conversation or a meeting. One possible opening is to couch it in light of some vague reference to the public release of healthcare data and indicate that you read in the paper that there would be a "random sample of patients pulled from all admissions." Now you can innocently ask, "Have any of you ever drawn a random sample?" and it will seem to be consistent with your setup.
8. Although it seems counterintuitive, it is possible to have too much data. There is frequently a belief (although it is generally a false belief) that if a little data is good, then a lot of data must be better. This is not always the case. For example, when a national news service conducts one of its "man on the street" surveys to test the political climate or a national public polling agency conducts a survey of American opinions, how many people do you think they include in their sample? Typically they shoot for 1,000 to no more

than 2,000 people. We have more than 280 million people in this country and they get only 1,500 respondents. Why don't they get more? After a certain point the additional data do not add anything to the statistical precision of a study. It merely wastes resources and time. A general rule of thumb is that 30–50 observations (data points, survey respondents, or numbers) will start to produce a distribution. If you stratify your respondents by age, gender, race, region of the country, urban/rural status, education, income, and religious preference, which is what the national news polls do, then you need more than 30–50 observations in order to ensure that each level of stratification has sufficient data to enable the appropriate statistical analysis. Telemarketers are experts at sampling. They can pinpoint down to the neighborhood area or census track level how many people represent the categories they need for their marketing study. A stratified proportional random sampling plan is put into place, the computer automatically dials the numbers, and your dinner is interrupted because you fit the sampling profile they need. But remember they do not need much data to complete their sampling plan.

9. While I was a doctoral student at Penn State University in the Department of Agricultural Economics and Rural Sociology, Dr. Bob Bealer used this phrase frequently. He used it when a student would ask, "How many pages do you want for this paper?" or when one of us would want to know how much data we needed to produce a "significant" result. Professor Bealer challenged us to think by using few words. He knew that the answers were rattling around somewhere within our developing brains. His skill was in providing a light for us to find the path. As much as you must and as little as you dare—it is a wonderfully simple phrase that relates to many aspects of life.

10. There are two classic stories about sampling that both the critics of sampling and its proponents have referenced for years. The first is the 1936 *Literary Digest* poll that predicted the landslide victory of Alf Landon over incumbent president Franklin D. Roosevelt. Using a mailed sample of more than 2 million voters, the *Literary Digest* predicted that Landon would win by almost 15%. The mistake they made was in selecting the list of individuals for the sample (this is referred to as the sampling frame). The sample was drawn from telephone directories and automobile registration lists. These methods had worked in the past elections quite nicely. What the *Literary Digest* pollsters forgot was that in 1936 the nation was still feeling the negative effects of the depression and the more positive impacts of Roosevelt's New Deal program. The 1936 election witnessed an unprecedented turnout of poor voters. These people were not proportionately represented in the telephone and car registration lists because they could not afford such luxuries. The other key issue that the *Literary Digest* missed was that the poor voters were primarily Democrats whereas the more wealthy voters, who could afford cars and telephones, were primarily Republicans. In this same year, however, George Gallup correctly predicted that Roosevelt would be the winner. Gallup's approach was based on using quota sampling, which ensured that samples were drawn from various segments of society (e.g., urban, rural, rich, poor, Republicans, and Democrats). As a result of this event, Gallup's credibility increased dramatically while that of the *Literary Digest* plummeted. The next major sampling fiasco occurred in 1948 when Gallup, and most other public opinion polling organizations, predicted that Thomas Dewey would be victorious over Harry Truman. What they all missed

in this case was (1) that nearly all the pollsters finished their polling too soon and missed the late surge for Truman and (2) the people who in earlier polls said they were not sure who they would vote for decided to vote predominantly for Truman. The success that Gallup had in 1936 with quota sampling proved to be disastrous 12 years later. It was after the 1948 election that academic statisticians began a serious push for using probability theory as a basis for drawing samples. Today the use of probability sampling methods remains the accepted standard for drawing the least amount of data with the highest level of predictability and confidence.

11. The other alternative to the *deus ex machina* is found in the following story:

> **The Facts of Life**
> The story that follows is about four people named Everybody, Somebody, Anybody, and Nobody. There was an important job to be done and Everybody was asked to do it. Anybody could have done it, but Nobody did it. Somebody got angry about that because it was Everybody's job. Everybody thought Anybody could do it, but Nobody realized that Everybody blamed Somebody when Nobody accused Anybody.

I am not sure of the origin of this story. My mother gave me a copy of it when I first started college. At the time I accepted it graciously and tucked it away, thinking that it was one of those things mothers give their children as they go off to college and hope that it makes them think about how their actions affect others. That was back in 1966. Today, I still have the original piece of paper she gave me with this story typed on it. Over the years it is funny how many times I have pulled out this little piece of paper or run across it in a cluttered desk drawer and realized how relevant the lines are to so many aspects of life. Many of the challenges we face with data and measurement stem from the fact that the people who own the process do not take ownership of their data and the results produced by their processes. Inevitably, when I am involved with assisting people in developing their indicators, there will be a moment when their discussion about data collection makes me think of this story.

12. The subgroup is basically how you have organized your data (e.g., daily, weekly or monthly) and appears on a chart as the label on the x or horizontal axis. More detail on selecting and using subgroups will be provided in Chapters 8 and 9.

References

Babbie, E. R. *The Practice of Social Research.* Belmont, CA: Wadsworth, 1979.

Brooke, R., C. Kamberg, and E. McGlynn. "Health System Reform and Quality." *Journal of the American Medical Association* 276, no. 6 (1996): 476–480.

Caldwell, C. *Mentoring Strategic Change in Health Care.* Milwaukee: Quality Press, 1995.

Campbell, S. *Flaws and Fallacies in Statistical Thinking.* Englewood Cliffs, NJ: Prentice-Hall, 1974.

Daniel, W., and J. Terrell. *Business Statistics.* Dallas: Houghton Mifflin, 1989.

Deming, W. E. *Some Theory of Sampling.* New York: John Wiley & Sons, 1950.

Deming, W. E. *Sample Design in Business Research.* New York: John Wiley & Sons, 1960.

Deming, W. E. "On Probability as a Basis for Action." *American Statistician* 29, no. 4 (1975): 146–152.

Deming, W. E. *Out of the Crisis.* Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study, 1992.

Deming, W. E. *The New Economics for Industry, Government, Education.* Cambridge, MA: MIT Press, 1994.

Donabedian, A. *Explorations in Quality Assessment and Monitoring.* Vol. 1: *The Definition of Quality and Approaches to Its Assessment.* Ann Arbor, MI: Health Administration Press, 1980.

Donabedian, A. *Explorations in Quality Assessment and Monitoring.* Vol. 2: *The Criteria and Standards of Quality.* Ann Arbor, MI: Health Administration Press, 1982.

Duncan, A. *Quality Control and Industrial Statistics*, 5th ed. Homewood, IL: Irwin Press, 1986.

Gonick, L., and W. Smith. *The Cartoon Guide to Statistics.* New York: Harper Perennial, 1993.

Hess, I., D. Riedel, and T. Fitzpatrick. *Probability Sampling of Hospitals and Patients.* Ann Arbor, MI: Health Administration Press, 1975.

Institute of Medicine. *Crossing the Quality Chasm.* Washington, DC: National Academy Press, 2001.

Ishikawa, K. *Guide to Quality Control.* White Plains, NY: Quality Resources, 1982.

Joint Commission on Accreditation of Healthcare Organizations. *The Measurement Mandate: On the Road to Performance Measurement in Health Care.* Oak Brook, IL: JCAHO, 1993.

Kaplan, R., and P. Norton. "The Balanced Scorecard—Measures That Drive Performance." *Harvard Business Review* (January–February 1992): 71–79.

Kaplan, R., and P. Norton. "Putting the Balanced Scorecard to Work." *Harvard Business Review* (September–October 1993): 134–147.

Kaplan, R., and P. Norton. "Using the Balanced Scorecard as a Strategic Management System." *Harvard Business Review* (January–February 1996): 75–85.

Langley, G., K. Nolan, T. Nolan, C. Norman, and L. Provost. *The Improvement Guide.* San Francisco: Jossey-Bass, 1996.

Maddox, B. *Sampling Concepts, Strategy and Techniques* (Technical Report 81-1). Harrisburg: Pennsylvania Department of Health, State Health Data Center, July 1, 1981.

Mann, N. R. *The Keys to Excellence: The Story of the Deming Philosophy*. London: Mercury Books, 1989.

Miller, D. *Handbook of Research Design and Social Measurement.* New York: David McKay Company, 1964.

Nelson, E., P. Batalden, and M. Godfrey. *Quality by Design: A Clinical Microsystem Approach*. San Francisco: Jossey-Bass, 2007.

Selltiz, C., M. Jahoda, M. Deutsch, and S. Cook. *Research Methods in Social Relations.* New York: Holt, Rinehart and Winston, 1959.

Shewhart, W. *Economic Control of Quality of Manufactured Product.* New York: D. Van Nostrand, 1931. Reprint, Milwaukee: Quality Press, 1980.

Weiss, R. *Statistics in Social Research.* New York: John Wiley & Sons, 1968.

Western Electric Co. *Statistical Quality Control Handbook*. Indianapolis, IN: AT&T Technologies, 1985.

Wheeler, D. *Understanding Variation: The Key to Managing Chaos.* Knoxville, TN: SPC Press, 1993.