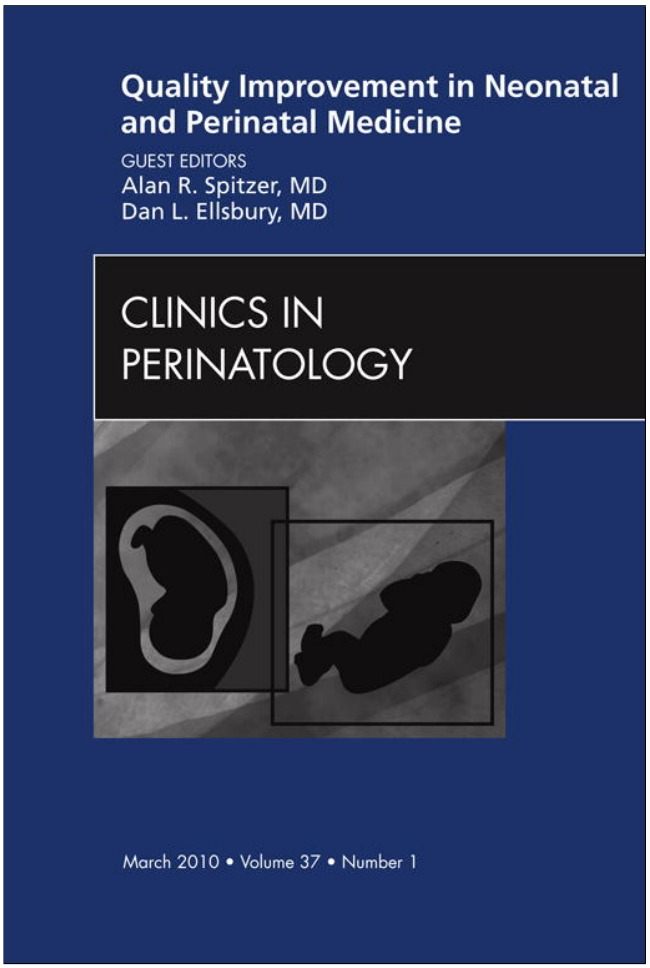

Navigating in the Turbulent Sea of Data: The Quality Measurement Journey

Robert C. Lloyd, PhD

KEYWORDS

- Quality measurement • Statistical process control
- Improvement sequence

WHERE AWAY AND WHY ALONE?

In 1892, Captain Eben Pierce offered his friend Joshua Slocum (1844–1909) a ship that "wants some repairs." Slocum went to Fairhaven, Massachusetts, to find that the ship was a rotting, old, 37-foot, oyster sloop propped up in a field. It was known as the *Spray*. Slocum spent 13 months repairing this vessel and on April 24, 1895, at the age of 51 years, he cast off from Gloucester, Massachusetts, in the *Spray*. As he was about to set off on his voyage a group of people called out to him, "Where away and why alone?"

Slocum covered 46,000 miles during his solo journey and landed back in Newport, Rhode Island, on June 27, 1898. His account of this journey, *Sailing alone around the world,* was published by the Century Co in 1900.[1] On November 14, 1909, at the age of 65 years, he set out from Martha's Vineyard on another lone voyage to South America, but was never heard from again.

Like Joshua Slocum, we are also on a journey. We are not battling 30-foot waves, howling winds, or pirates. But we are facing pressures and challenges that test our knowledge, experience, and our abilities. The primary question is this: Do you have a plan to guide your quality journey? Or are you adrift in a turbulent sea of data, hoping that your numbers meet the internal and external demands that are constantly testing your navigational skills? Or are you headed in the wrong direction and feeling a little like Joshua Slocum, adrift alone in a turbulent sea? "Where away and why alone?"

WHY ARE YOU MEASURING?

In 1997, Solberg and colleagues[2] described what they called the 3 faces of performance measurement. They wrote:

Institute for Healthcare Improvement, 20 University Road, 7th Floor, Cambridge, MA 02138, USA
E-mail address: rlloyd@IHI.org

Clin Perinatol 37 (2010) 101–122
doi:10.1016/j.clp.2010.01.006 **perinatology.theclinics.com**
0095-5108/10/$ – see front matter

We are increasingly realizing not only how critical measurement is to the quality improvement we seek but also how counterproductive it can be to mix measurement for accountability or research with measurement for improvement.

The investigators describe in detail various characteristics of performance measurement for accountability (what many today call data for judgment), research, and improvement. These characteristics are summarized in **Table 1**. The authors' distinctions between the various aspects of the measurement journey help us quickly realize that not all measurement is the same. Yet many health care professionals do not think about why they are actually measuring. You will hear managers or frontline workers say, for example, "Look, we need to submit some data on our progress related to ventilator-associated pneumonias in the neonatal intensive care unit, so find some recent numbers and send them in." Frequently this means the data submitted may not be the most recent data, defined in the same way they were defined when they were first submitted or stratified according to the same criteria used the previous year. Furthermore, the data may be presented in a manner that works when accountability questions are driving the inquiry, but they may be inadequate for questions related to quality and safety or conducting randomized control trials (RCTs).

Brook and colleagues[3] have also helped to clarify the performance measurement journey. They point out that research (ie, RCTs) designed to determine the efficacy

Table 1
The 3 faces of performance measurement

Aspect	Improvement	Accountability	Research
Aim	Improvement of care	Comparison, choice, reassurance, spur for change	New knowledge
Methods			
• Test observability	Test observable	No test, evaluate current performance	Test blinded or controlled
• Bias	Accept consistent bias	Measure and adjust to reduce bias	Design to eliminate bias
• Sample size	Just enough data, small sequential samples	Obtain 100% of available, relevant data	Just in case data
• Flexibility of hypothesis	Hypothesis flexible, changes as learning takes place	No hypothesis	Fixed hypothesis
• Testing strategy	Sequential tests	No tests	One large test
• Determining if a change is an improvement	Run charts or Shewhart control charts	No change focus	Hypothesis, statistical tests (*t* test, F test, χ^2), *P* values
• Confidentiality of the data	Data used only by those involved with improvement	Data available for public consumption and review	Research subjects identities protected

of a drug, procedure, or treatment is designed to answer questions about efficacy. Quality improvement research, on the other hand, is directed at improving the efficiency or effectiveness of processes and their related outcomes.

Anyone engaged in performance measurement needs to be clear about the reasons for collecting and analyzing data. As shown in **Table 1**, each of the 3 faces uses different methods and different statistical techniques to derive conclusions from the data. If an organization is genuinely interested in leading the way for quality and safety then it needs to be clear about the reasons for measurement. All too often organizations say they are focused on quality and safety. Then they discover that their approach to performance measurement is based primarily on data for accountability or judgment. This observation is not to suggest that 1 of the 3 faces is more correct than the other. All 3 faces of performance measurement can be useful. A problem arises, however, when organizations attempt to mix the 3 faces. This error is what Solberg and colleagues[2] indicate leads to the development of counterproductive performance measurement systems.

THE QUALITY MEASUREMENT JOURNEY

Aim

The milestones in the quality measurement journey (QMJ) are outlined by Lloyd[4] and summarized in **Fig. 1**. The first milestone in this journey requires clarity about the aim of measurement. Measurement should be directly and overtly connected to the organization's mission, aims, and objectives. One can easily determine how connected a team is to the organization's strategic objectives. The next time you are involved with a pediatrics improvement team, just pose the following question: "Can anyone tell me how this team's work fits with the organization's strategic objectives?" After a period of silence, some brave soul might respond, "I have no idea. We were told by our boss to improve this process." If the employees of an organization do not understand and internalize how their work fits into the organization's overall strategy for quality and safety, they will end up going through the motions and think they are "doing quality." They will fail to connect their work to the organization's purpose and objectives, and they will go through the motions but never connect the dots. Aims help answer the question "Why are you measuring?"

Concepts

Concepts, the next milestone in the QMJ, stem from clarity around the high-level aims. Yet the concepts do not represent measurement. They are essentially an intermediate step designed to help a team set the boundaries for measurement and data collection.

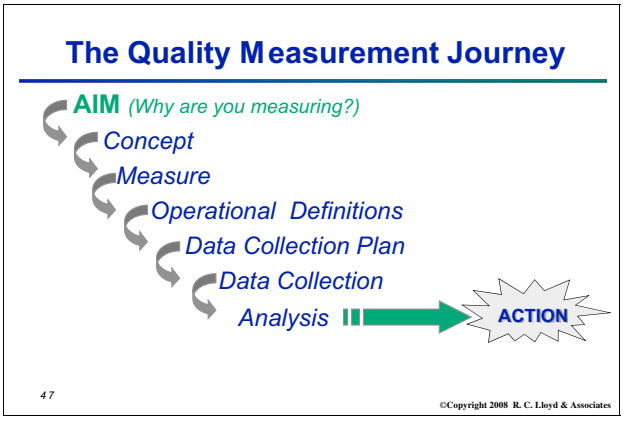

Fig. 1. Milestones in the QMJ. (*Courtesy of* R.C. Lloyd & Associates.)

For example, in **Fig. 2** the aim is to have freedom from harm. This type of statement will be found frequently in an organization's mission statement. From this aim emerge various concepts that address different aspects of harm. In **Fig. 2**, the example is reducing neonatal unplanned extubations of endotracheal tubes (ETTs). We have become more specific by saying that we want to reduce unplanned extubation as a form of harm but this is still not measurement. Reducing unplanned extubation is a desired outcome. It is not until you move to identifying a specific way to measure unplanned extubation that you can take the first steps toward reducing it.

Measures

There are numerous options to consider as we move from a concept to a measure. The first one is deciding which measure to select out of all the potential measures.[5] Using the concept of unplanned extubations we might consider the following measures:

- We could count merely the number of unplanned extubations in a defined period of time (eg, during a shift, during a week, or for the entire month). What does this give us? Is a count of the number of unplanned extubations the most appropriate way to measure the concept? This month we had 21 unplanned extubations. Last month we had 13. What does this tell us? It becomes even more challenging if you want to compare your performance to that of another hospital in your area or system. Hospital A's neonatal intensive care unit (NICU) and B's NICU each had 19 unplanned extubations this month. Which one is better? You really cannot decide which is better or worse in this situation because you have no context for the absolute numbers. If you are told, however, that hospital A is a large urban teaching hospital with 50 isolettes and hospital B is a community hospital with only 10 isolettes, you now have a context and would most likely say that it is not fair to compare the two because of differences in size, volume, location, and so forth.
- Next, we could consider computing the percentage of neonates who have an unplanned extubation. In this case, we would need to define a denominator (ie, all neonates who could possibly have an unplanned extubation). The numerator would then be all the neonates who did experience an unplanned extubation during their stay in the NICU aggregated for the defined period of time (eg, a week or a month). With these 2 numbers we could compute the percentage of neonates with an unplanned extubation during the defined time period. Because an unplanned extubation could happen more than once during a stay in an NICU, however, the percentage would not capture the multiple unplanned

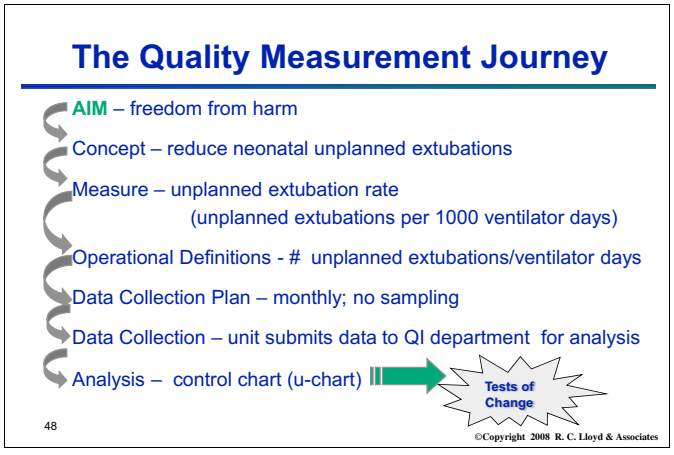

Fig. 2. Example of a QMJ. Reducing undesired or accidental extubations in neonates.

extubations. A percentage is based on a binomial distribution. Measuring unplanned extubations with a percentage, therefore, means that the team is not concerned with the specific number of times a baby experienced an unplanned extubation, but rather if the patient had an unplanned extubation once or more. The question is simply, ''Did this baby experience an unplanned extubation, yes or no?''

- This question leads us to the third option for measuring an unplanned extubation, a rate. Like a percentage a rate is calculated by having a numerator and a denominator but they are different from the ones defined for a percentage. An unplanned extubation rate would have as the numerator the total number of unplanned extubations, including multiples for 1 baby, during a defined period of time (eg, a shift, a day, a week, or a month). The denominator would then be the total number of ventilator days in the defined period of time. These calculations would produce an unplanned extubation rate (eg, 18 unplanned extubations per 1000 ventilator days). A rate-based statistic has a different measure in the numerator and the denominator (eg, extubations over days). A percentage has the same measure in the numerator and denominator but they are merely different classes of the same variable (neonates experiencing an unplanned extubation over total neonates on an ETT). Examples of potential measures for various health care concepts can be found in Lloyd.[4]

Operational Definitions

Once a team has decided what to measure, they can proceed to the next milestone in the QMJ, namely building operational definitions. This task is 1 of the most interesting stops along your journey because it addresses the lack of precision in human language. According to Deming,[6] ''An operational definition puts communicable meaning to a concept. Adjectives like good, reliable, uniform, round, tired, safe, unsafe, unemployed have no communicable meaning until they are expressed in operational terms of sampling, test, and criterion. The concept of a definition is ineffable: It cannot be communicated to someone else. An operational definition is one that reasonable men can agree on.''

Operational definitions are not universal truths. They are merely ways to describe, in quantifiable terms, what to measure and the specific steps needed to measure it consistently. A good operational definition has the following characteristics:

- It gives communicable meaning to a concept or idea
- It is clear and unambiguous
- It specifies the measurement method, procedures, and equipment (when appropriate)
- It provides decision-making criteria when necessary
- It enables consistency in data collection.

Again, using the concept of an unplanned extubation, it is necessary to ask, ''What is the operational definition of an unplanned extubation?'' All unplanned extubations are not the same. There could be a partial extubation or a complete extubation. What is the difference between a partial extubation and a complete extubation? What if the tape holding the ETT came loose and the tubing sags a little on 1 side? Is this a partial extubation? Do we all agree on the characteristics of a partial versus a complete extubation? If we sent out 3 people to collect data on unplanned extubations would they all define a partial extubation in the same way? Would the data be valid and reliable? Could we combine the data from the 3 people and have confidence

that we were comparing apples with apples? If our operational definition of a partial extubation met the 5 criteria listed earlier for a good operational definition, our data would most likely be consistent from person to person. If, on the other hand, the 3 people did not use consistent operational definitions, you would end up with fruit salad rather than apples compared with apples. Additional detail on the critical role of operational definitions plus examples can be found in Lloyd[4] and Provost and Murray.[7]

Data Collection

After reaching consensus on the operational definitions for your measures the next milestone in the QMJ (see **Fig. 1**) is to develop a data collection plan and then go out and gather the data. These 2 steps in the QMJ frequently run into roadblocks because team members or improvement advisors are not well trained in the methods and tools of data collection. The major speed bump at this point in your journey, however, is that most people wait until it is time to collect the data before they start thinking about it. A well-developed data collection plan saves you time, effort, and money. A few key questions to consider at this junction are as follows[4]:

- What is the rationale for collecting these data rather than other types of data?
- Will the data add value to your quality improvement efforts?
- Have you discussed the effects of stratification on the measures?
- How often (frequency) and for how long (duration) will you collect the data?
- Will you use sampling? If so, what sampling design have you chosen?
- How will you collect the data? (Will you use data sheets, surveys, focus group discussions, telephone interviews, or some combination of these methods?)
- Who will go out and collect the data? (Most teams ignore this question.)
- What costs (monetary and time costs) will be incurred by collecting these data?
- Will collecting these data have negative effects on patients or employees?
- Do your data collection efforts need to be taken to your organization's institutional review board for approval?
- Do you already have a baseline?
- Do you have targets and goals for the measures?
- How will the data be coded, edited, and verified?
- Will you tabulate and analyze these data by hand or by computer?
- How will these data be used to make a difference?

Besides having a serious dialog about these questions, there are 2 key skills needed during this part of your journey. The first is stratification and the second is sampling.

Stratification is the separation and classification of data into reasonably homogeneous categories. The objective of stratification is to create groupings that are as mutually exclusive as possible. Such groupings are intended to minimize variation between groups and maximize variation within a group of similar patients, procedures, or events. Stratification is also used to uncover patterns that may be suppressed when all of the data are aggregated. Stratification allows understanding of differences in the data that might be caused by:

- Day of the week (Mondays are different from Wednesdays)
- Time of day (turnaround time [TAT] is longer between 9 AM and 10 AM than it is between 3 PM and 4 PM)
- Time of year (we treat more patients with influenza in January than June)
- Shift (the process is different during day shift than during night shift)
- Type of order (short turnaround time [STAT] vs routine)

- Weight of the baby
- Type of machines or equipment.

Stratification is more of a logical issue than a statistical one. It requires talking with people who have subject matter expertise, knowing how the process works, and where pockets of variation may exist.

Returning to our example of unplanned extubation we might ask the following stratification questions:

- Does it matter if the baby is secreting fluids that could affect the tape being used to hold the ETT in place? If so, then we might stratify by mild, moderate, or copious amounts of fluid.
- Does the activity level of the baby affect unplanned extubation? If the answer is yes, then we might consider stratifying by mild, moderate, or high levels of activity, or use an activity index.
- What if a hydrocolloid dressing was placed across the neonate's philtrum before taping the ETT to the infant? Does a hydrocolloid dressing make a difference in unplanned extubations?
- Does the type of tape used to hold the ETT in place matter? If it does, then should we stratify by the type of tape (brand A vs brand B)?
- Finally, does it matter if we apply the tape to the baby's face in an H or Y pattern? If the NICU staff believe that the taping pattern makes a difference, then we should stratify on this characteristic also.

Stratification is critical especially if you think that certain factors may differ depending on the characteristic (or stratum) being used in the measurement. Once the data have been collected, it is usually too late or too time consuming to try to separate the stratification issues that may arise. Further details and examples of stratification can be found in Lloyd[4] and Provost and Murray.[7]

Sampling is the second key skill needed during the data collection stage of your journey. Sampling is an efficient and effective way to gather data when you: (1) do not need all the available data, and (2) do not have unlimited resources (time, effort, and money). First, consider the volume issue. Each day a typical hospital processes hundreds of complete blood counts (CBCs). If you are interested in TAT for CBCs, you do not need to analyze all 293 tests done on Monday each day to get a good picture of the TAT for that day. When you have these many data (ie, 293 tests during 1 day) you might consider stratification into day, afternoon, and night shifts; then stratify further to sort out STAT and routine test requests for each shift. We could then select a stratified random sample from each shift that also lets us know how STAT and routine TATs varied within the shift. In this case, a sample of 15 CBCs would be sufficient to analyze the variation on each day. Analyzing all 293 TATs is not necessary. A well-designed sampling strategy will work well.

The second reason to sample is to conserve resources. Imagine that you wanted to collect data that required 3 staff nurses to record 4 different measures on each baby in the NICU. This effort represents an expensive proposition. Rather than collect the 4 measures on all babies, you might consider developing a sampling plan to select 3 to 5 babies a day, or select a random day of the week on which to gather the data. Sampling provides a parsimonious approach to data collection. The critical question is how to draw appropriate samples.

There are 2 basic major types of sampling: probability and nonprobability. The details on the advantages and disadvantages of the various sampling approaches

can be found in Lloyd[4] and Provost and Murray.[7] Also you can find practical discussions of sampling methods in any basic text (old or new) on statistical methods or research designs.

Probability sampling methods are based on a simple principle: within a known population of size n, there is a fixed probability of selecting any single element (n_i). The selection of this observation (and the remaking members of the sample) must be determined by objective statistical means if the process is to be truly random (not affected by judgment, purposeful intent, or convenience). There are 4 basic approaches to probability sampling:

- Systematic sampling, which is achieved by numbering or ordering each element in the population and then selecting every *k*th element. The key point that most people ignore when pulling a systematic sample is that the starting point for selecting every *k*th element should be generated through a random process. For example, if you were evaluating how long it takes to get a newborn baby from the delivery room to the NICU, and you wanted to draw a systematic sample, you would pick a random number between 1 and 10 (eg, 7) and then start observing the time of every *k*th baby after the seventh one. If you said, ''Let's start at the first baby and then take every 10th baby to check the time it takes from delivery to the NICU'' you would potentially be introducing bias. A random starting point is critical to making systematic sampling a form of probability sampling.
- Simple random sampling is accomplished by giving every element in the population an equal and independent chance of being included in the sample. A random number table or a random number generator in a computer program is usually used to develop a random selection process.
- Stratified random sampling results when stratification is applied to a population; then a random process is used to pull samples from within each stratum. The CBC example presented earlier provides an illustration of this approach.
- Stratified proportional random sampling is more complicated because it requires figuring out what proportion each stratum represents in the total population, then replicating this proportion in the sample that is randomly pulled from each stratum. To successfully use this approach, you need to have sufficiently large populations that can be divided into smaller stratification levels, yet still have enough data from which to draw an appropriate sample. For example, if you stratify all deliveries by age, race, and prior deliveries within the last 30 days, you may have a category of Hispanic women more than 40 years old who had a previous cesarean section that contains only 2 patients. In this case, you have stratified by so many levels that you have reduced the number of patients to a point that sampling does not make sense.

Nonprobability sampling methods are usually used when the researcher is not interested in being able to generalize the findings to a larger population. The basic objective of nonprobability sampling is to select a sample that the researchers believe is typical of the larger population. A chief criticism of these approaches to sampling is that there is no way to factually measure how representative the sample is of the total population under consideration. Samples pulled through nonprobability designs are assumed to be good enough for the people drawing the sample, but the finding should not be generalized to larger populations.

- Convenience sampling is the classic man-on-the-street interview approach to sampling. In this case, a reporter may select 10 people standing on the train platform (who look interesting or approachable) and ask them what they think of the

national health care debate and the public option. Although these interviews may provide interesting sound bites, they should not be used to arrive at a conclusion that this is how the general public feels about the issue.

- Quota sampling is frequently used with convenience sampling. When this approach is used, the reporter knows that they need to get a total of 2 sound bites (the quota) for the producer to use. So the reporter focuses on obtaining these 2 interviews as the quota. This technique is used frequently in health care settings, when a quota of *n* charts or *m* patient interviews is set as the desired amount of data. There are steps that can be taken in developing quota samples[8] to ensure reasonably robust data. Most of the time these steps are not followed, and the quota sample represents a weak and biased approach to sampling.
- Judgment sampling is frequently used in quality improvement initiatives. Judgment sampling relies on the knowledge of subject matter experts. These individuals can tell you when the performance of a process varies and when this variation should be observed. For example, if the admitting clerk tells you that patients bunch up between 08:30 and 09:30 AM, and that this is a different process than what she observes between 15:00 and 16:00 PM, then we should consider sampling differently during these 2 time periods. Similarly, if a staff nurse tell you that ''Things get crazy around here at 11:00 due to discharge timing,'' we would want to create a sampling plan for ''crazy time'' and ''noncrazy time.'' The critical point for judgment sampling is that the person offering the judgment needs to be a subject matter expert on the process and how it works. Otherwise, bias increases dramatically in this form of sampling.

Building knowledge in sampling methods is 1 of the best things that someone can do to enhance data collection processes. Good sampling techniques help to ensure the validity and reliability of the data that are taken to the next milestone in your QMJ analysis.

Analysis

How you analyze your data depends on a critical question: Will you approach data analysis from a static or dynamic perspective? Deming[9] labeled these 2 approaches as enumerative (static) and analytical (dynamic). He pointed out that quality improvement studies are best approached from an analytical perspective. Yet, most health care professions have received statistical training that is grounded solely in static approaches to data analysis.

Static approaches are designed to summarize a characteristic of the data with a single measure that is fixed at a single point in time. The descriptive statistics used include measures of central tendency (mean, median, and mode) and measures of dispersion (minimum, maximum, range, and standard deviation). Once the descriptive statistics have been computed the next step in the static journey is to compare 2 or more data points to find out if they are statistically different. In this example, techniques such as χ^2, Student *t* test, analysis of variance, or correlation/regression analyses are used to determine if 1 data point is different from another. Statistical tests of significance, usually determined by a *P* value, are the standard method to verify differences.

The analytical approach to data analysis stands in contrast to the static approach. Analytical methods are based on statistical process control (SPC) methods. This branch of applied statistics was developed by Dr Walter Shewhart in the early 1920s while he was working at Western Electric Co.[10] The primary SPC tools are the run and control charts. Statistical analysis conducted with SPC methods looks at variation in a process or outcome measures over time, not at a fixed point in time, or compares 2 data points and asks if they are statistically different.

Because variation exists in all processes (eg, consider morning commute time), the use of run charts and Shewhart control charts allows the researcher to analyze data as a continuous stream that has a rhythm and pattern. Statistical tests are used to detect whether the process performance reflects what Shewhart classified as common cause variation or special cause variation. Decisions about improvement strategies and their effects are based on understanding the type of variation that lives in the process, not on whether 1 data point is different from another. SPC charts, therefore, are more like the patterns of vital signs seen on telemetry monitors in the NICU.

Run Charts

A run chart provides a running record of a process over time. It offers a dynamic display of the data and can be used on virtually any type of data (eg, counts of events, percentages, rates, or physiologic measures). **Fig. 3** shows the layout for a typical run chart. The measure of interest is always plotted along the y-axis, whereas the x-axis is reserved for the subgroup or unit of time used to organize the data. Day, week, month, shift, or even patient are typical units that are placed on the x-axis. Because run charts require no complex statistical calculations, such as sigma limits, they can be understood easily by everyone. The major drawback in using run charts, however, is that they can detect some but not all special causes in the data.

The first step in analyzing a run chart is to understand what is meant by a run. A run is defined as 1 or more consecutive data points on the same side of the median. When you are counting runs, you should ignore points that decrease directly on the median. **Fig. 4** shows the number of runs on the chart shown initially in **Fig. 3**. An alternative way to count the number of runs is to examine the number of times the sequence of data points crosses the median and add 1. If you count the number of circled runs, or if you add 1 to the number of times the data cross the median, you get the same number: 14. So, in **Fig. 4**, there are 14 runs.

Once the number of runs is identified, you can then decide if the chart indicates the presence of common cause (random variation) or special cause (nonrandom variation). Four simple run chart rules are used to detect the 2 types of variation. The tests include:

- A shift in the process (6 or more consecutive data points above or below the median)

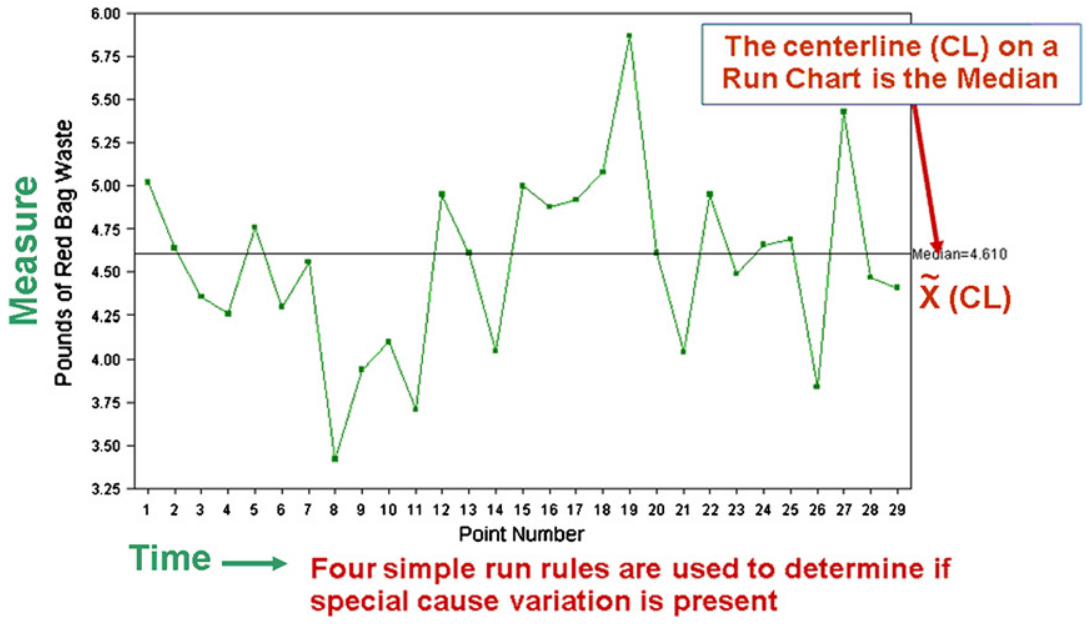

Fig. 3. Elements of a run chart.

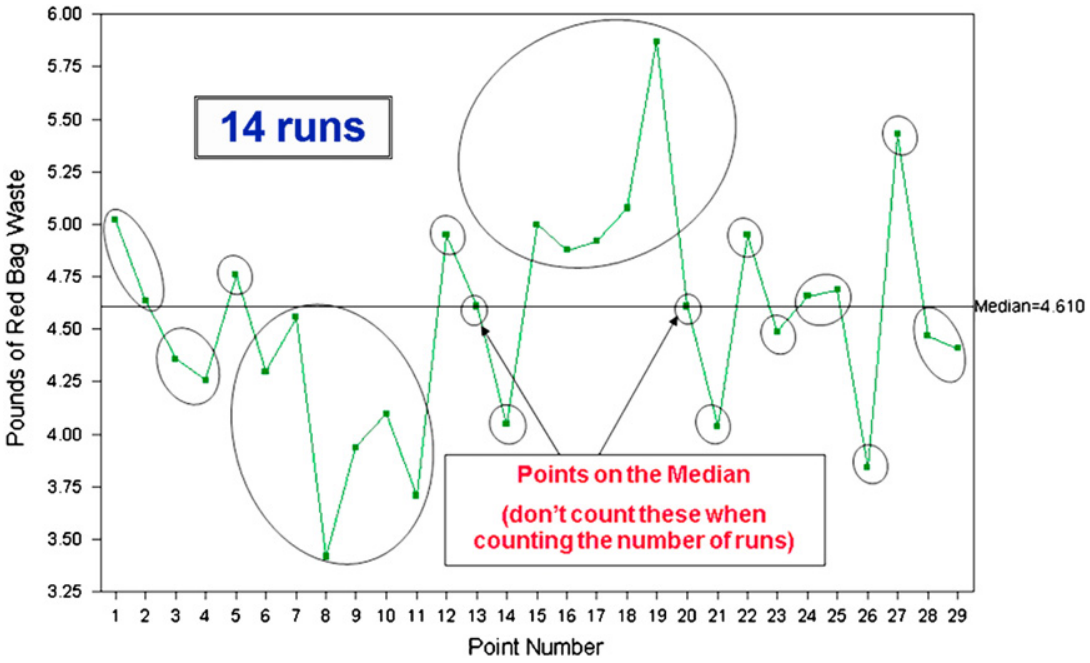

Fig. 4. Determining the number of runs.

- A trend (5 or more consecutive data points constantly going up or down)
- Too many or too few runs (determined by using a table that shows the number of runs expected for a given data set)
- An astronomical data point (this is a judgment call to decide if there is 1 or more data points in the set that seem to have an extreme variation).

Fig. 5 provides a visual display of these 4 run chart rules. The run chart rules are applied to the chart shown in **Fig. 6**.

The box next to **Fig. 6** shows how the run chart would be analyzed. There are 29 total data points on the chart. Two of the data points are on the median so they are not counted. This assessment leaves 27 useful observations (data points not on the median). When you look up 27 useful observations in a table,[11] you will see that the

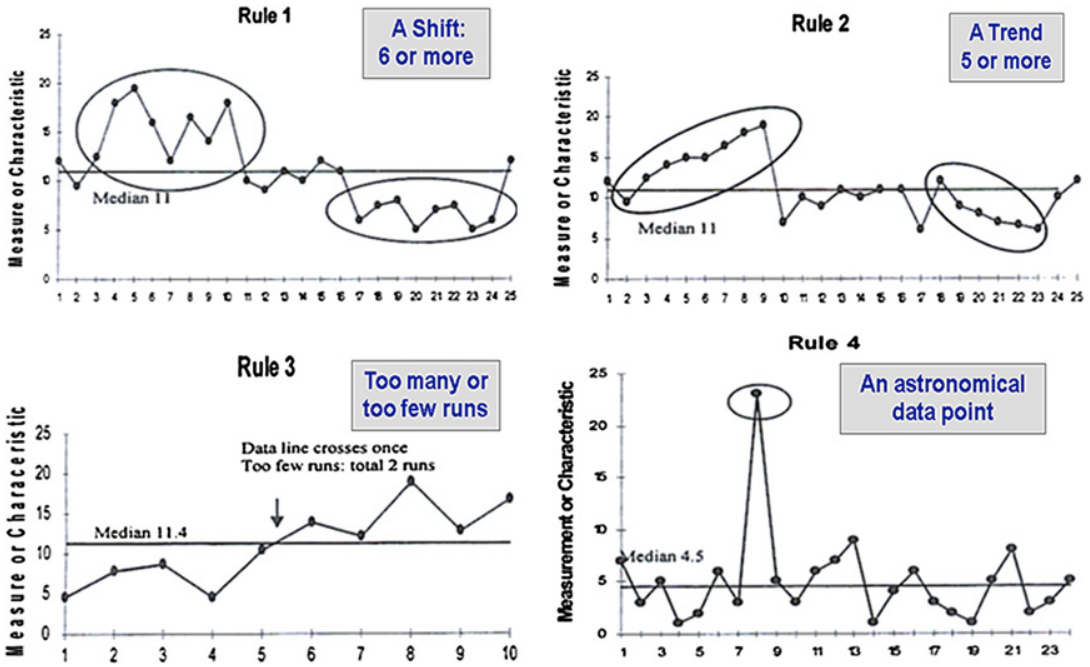

Fig. 5. The 4-run chart rules.

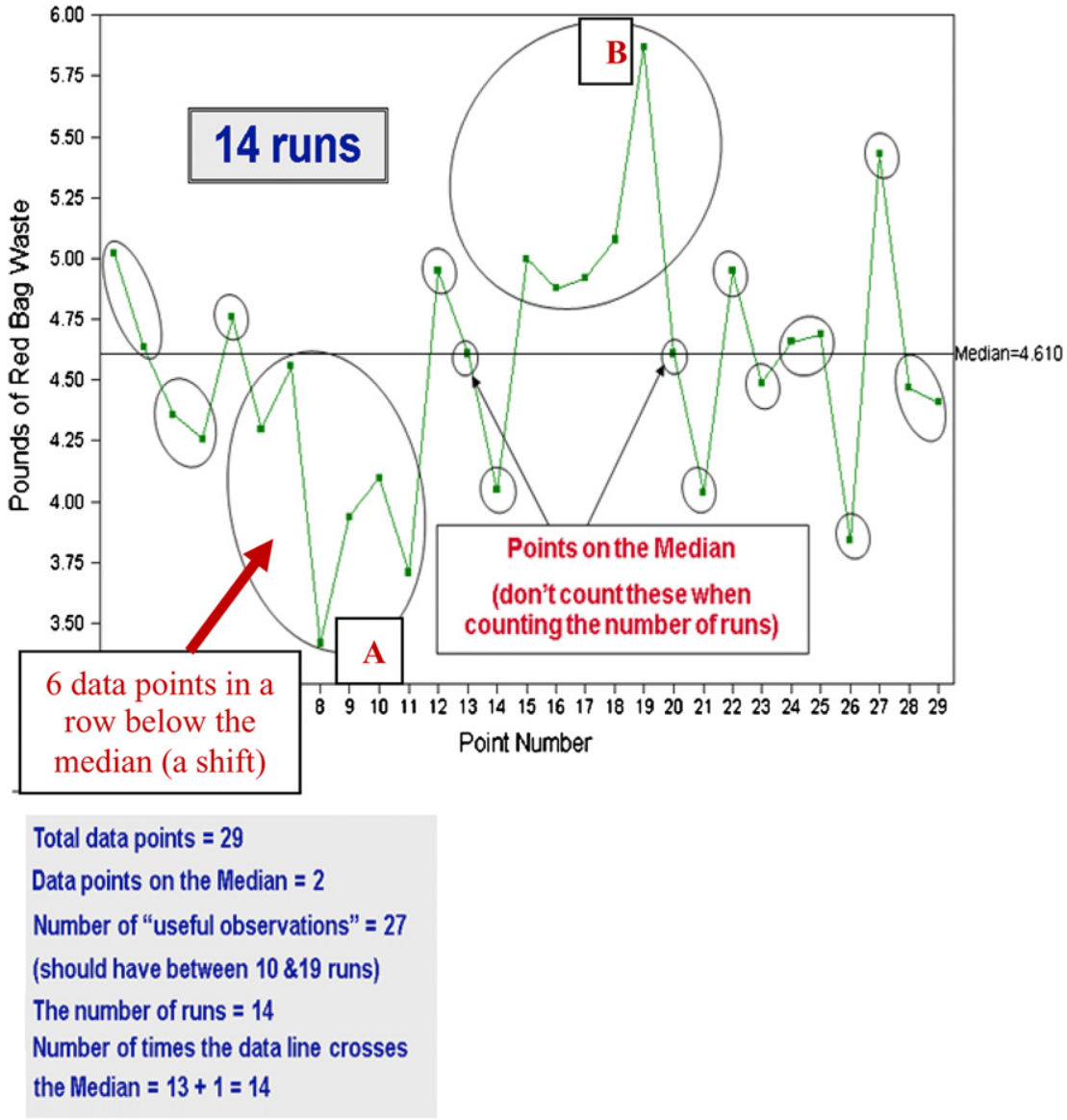

Fig. 6. Applying the run chart rules. (*Adapted from* Provost L, Murray S. The data guide. Austin [TX]: Associates in Process Improvement and Corporate Transformation Concepts; 2007. p. 3–15; with permission.)

lower number of runs for 27 data points is 10 and the upper number of runs is 19. This calculation indicates that if the data reflect random variation, there should be between 10 and 19 runs. If the number of runs was less than 10 or more than 19 it would indicate that the data set has either too little or too much variation. **Fig. 6** contains 14 runs that decrease within this range, so we know that at least for this test (too few or too many runs), the chart shows random variation (ie, nothing special is observed).

If we apply the trend test (5 data points constantly going up or down) we do not find such a pattern. We do observe a shift in the data, however. A shift is 6 or more data points on the same side of the median. The fourth run from the left contains 6 data points and indicates a statistically significant shift downward in the data (ie, a nonrandom pattern). Another way of interpreting this finding is that for this many data points (n = 27) we should not see data hanging in a run above or below the centerline. When it does (in this case 6 data points below the median), we have a signal that the process does not display random variation. The appropriate management decision in this case is to investigate why we had pounds of red bag waste significantly lower than at other points in the data collection period for 6 weeks in a row. Did we

have fewer patients? Were fewer procedures performed? Were more staff on vacation during this period? Because the goal is to reduce the amount of red bag waste, we would like the process to function at lower levels. So, what does it take to shift the entire process average (the median in this situation) to a more desirable level? This is an improvement question for a team to investigate.

The last run chart test determines if there are astronomical data points present. Remember that in any given data set, there will be a high and low data point. These points are not necessarily astronomical. Rule 4 in **Fig. 5** shows an astronomical data point. In **Fig. 6**, some might conclude that point A or point B is astronomical. Neither of these points is astronomical because they essentially balance each other out. If you had only point B on the chart and point A was nuzzled in the midst of the rest of the data, then point B might be an astronomical data point. Another way to look at this issue is to imagine that all the data points were pushed to the far right side of the chart to form a distribution. The data in **Fig. 6** would form an almost perfect normal distribution, with points A and B lodged in the outermost tails of the normal curve. In conclusion, the management decision with these data rests on the answers to 2 important questions: (1) Are we comfortable that, on average, about 4.6 pounds of red bag waste is produced each shift (shift is the unit of time across the x-axis)?; and (2) Are we willing to accept the variation in the process? A ''No'' response to either of these questions would indicate the need for improvement.

Shewhart Charts

Although most people refer to control charts as the primary SPC tool, the appropriate terminology is actually Shewhart charts, in honor of Dr Walter Shewhart, who developed the fundamental aspect of the charts in the early 1900s while he was working at Western Electric Co.[10] In 1931, Shewhart published his classic work, *Economic control of quality of manufactured product*. This book has served as the foundation for all subsequent work in SPC.

Shewhart charts are preferred to run charts because they:

1. Are more sensitive than run charts
   - A run chart cannot detect special causes that are a result of point-to-point variation (the median of the run chart is replaced with the mean on a Shewhart chart)
   - Tests for detecting special causes can be used with control charts, whereas the run charts are able to identify random or nonrandom patterns in the data
2. Have the added feature of control limits, which allow us to determine if the process is stable (common cause variation) or not stable (special cause variation)
3. Can be used to define process capability (which run charts cannot do)
4. Allow us to more accurately predict process behavior and future performance.

Like the run chart, Shewhart charts are plots of data arranged in chronologic order (**Fig. 7**). The mean or average is plotted through the center of the data; then the upper control limit (UCL) and lower control limit (LCL) are calculated from the inherent variation in the data. The control limits are not set by the individual constructing the chart. If appropriate, the individual making the chart can place specification limits or a target on the chart to determine how well the actual variation matches the desired performance of the process.

Shewhart was keenly interested in trying to understand the scientific basis for statistical control. As he observed the world around him, he realized that certain types of variation (common cause variation) were part of the normal function of life. At other times, however, he observed that variation was not normal and random, but a result

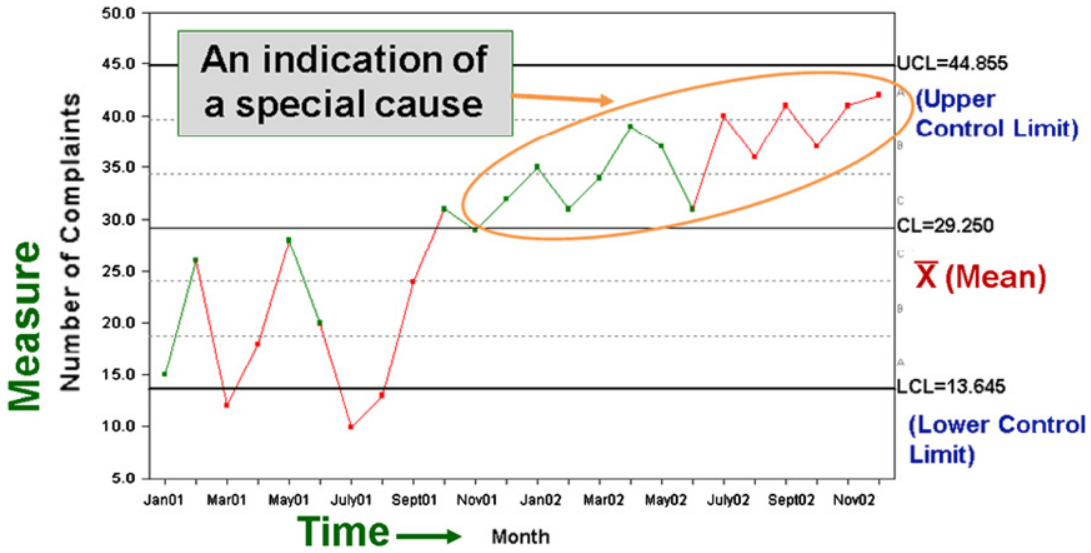

Fig. 7. Elements of a Shewhart chart.

of special or assignable causes. From Shewhart's perspective, the challenge was to distinguish 1 type of variation from the other. In 1931 he wrote:

> *A phenomenon will be said to be controlled when, through the use of past experience, we can predict, at least within limits, how the phenomenon may be expected to vary in the future. Here it is understood that prediction within limits means that we can state, at least approximately, the probability that the observed phenomenon will fall within the given limits.*

This definition provides a verbal description of the purpose of a Shewhart chart: prediction of the future. The question that most people ask at this point, however, is "Okay, I understand what Shewhart is trying to tell us, but I do not understand where these control limits come from." If you are asking this question, it is a sign that you are comfortable with the analytical concept of variation and ready to proceed with some of the more technical aspects of SPC. If, on the other hand, you would like to read more about understanding variation you may want to review Provost and Murray,[7] Lloyd,[4] Wheeler,[12] and Duncan.[13]

The technical aspects related to the Shewhart charts are numerous and too involved for the space limitations of this article. There are, however, several key points that need to be highlighted. The reader can then decide if a deeper dive into the theory and mechanics behind the Shewhart charts is required. Additional details on SPC methods can be found in Refs.[4,7,12,14–19]

The first step in applying Shewhart charts to your work is to decide if your data can be classified as variables or attributes. This consideration is not an issue with run charts because there is only 1 way to make a run chart and you can place any type of data on a run chart without distinguishing whether those data are characterized as a count, a percentage, or a rate. It does make a difference with the Shewhart charts, however, because there are different types of charts for different types of data.

Variables data (sometimes referred to as continuous data) can take on different values on a continuous scale. These data can either be whole numbers, or they can be expressed in as many decimal places as the measuring instrument can read. Examples of continuous data include time in minutes, weight in grams, length of stay, blood sugar levels, total number of procedures, or total number of discharges. Attributes data, on the other hand, are basically counts of events that can be

aggregated into discrete categories (eg, acceptable vs not acceptable, infected vs not infected, or late vs on time).

It is helpful to distinguish 2 types of attributes data. The first type involves counting the occurrences and the nonoccurrences of an event and reporting the number or percentage of defectives. An example would be the percentage of neonates who had an unplanned extubation during their stay in the NICU. In this case, you know the occurrences (total number of unplanned extubations) and you know the nonoccurrences (total number of babies with an ETT). The ability to obtain a numerator and a denominator allows you to calculate the percentage of incomplete patient charts.

There are times, however, when you know the occurrences but you do not know the nonoccurrences. At first this may seem like an anomaly, but there are many situations in health care that have this characteristic. For example, on a given day you can count the number of patient falls but you do not know how many ''nonfalls'' there were. Similarly, you can count the number of needlesticks but you do not know how many ''non-needlesticks'' occurred. Counts of this nature are usually regarded as defects, compared with defectives. For many students of SPC this distinction between defectives and defects requires a little soaking time to fully absorb. This may be 1 of the areas that you bookmark for further study and consideration.

Once you know the type of data you have collected, it is time to decide which control chart is most appropriate for your data. There are basically 7 different control charts, as summarized in **Fig. 8**. Note that 3 of the charts relate to variables data, whereas 4 charts are appropriate for attributes data.

The Shewhart decision tree shown in **Fig. 9** provides an algorithm that many find useful when deciding which chart is most appropriate for their data. The successful use of the decision tree requires understanding the following terms: subgroup, observation, and area of opportunity. These terms are defined in **Table 2**. Note that subgroup and observation relate to all the charts, whereas the area of opportunity is pertinent to only the attributes charts.

Of these 7 charts, health care data are most often displayed on 5 of the charts. These include X bar and S chart, XmR chart (individuals chart), the p-chart (percentages or proportions), the c-chart, and the u-chart (rates). Specifically, applications and examples of the use of these charts can be found in Provost and Murray,[7] Lloyd[4]; Carey,[15] and Carey and Lloyd.[16]

Once you have selected and made the appropriate Shewhart chart, it is time to interpret the chart. This process is similar to the one we used for determining if the run chart

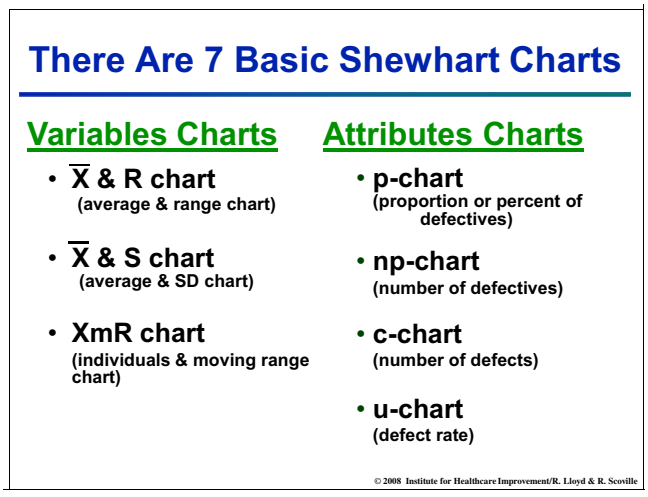

Fig. 8. The basic Shewhart charts. (*Courtesy of* Institute for Health Improvement.)

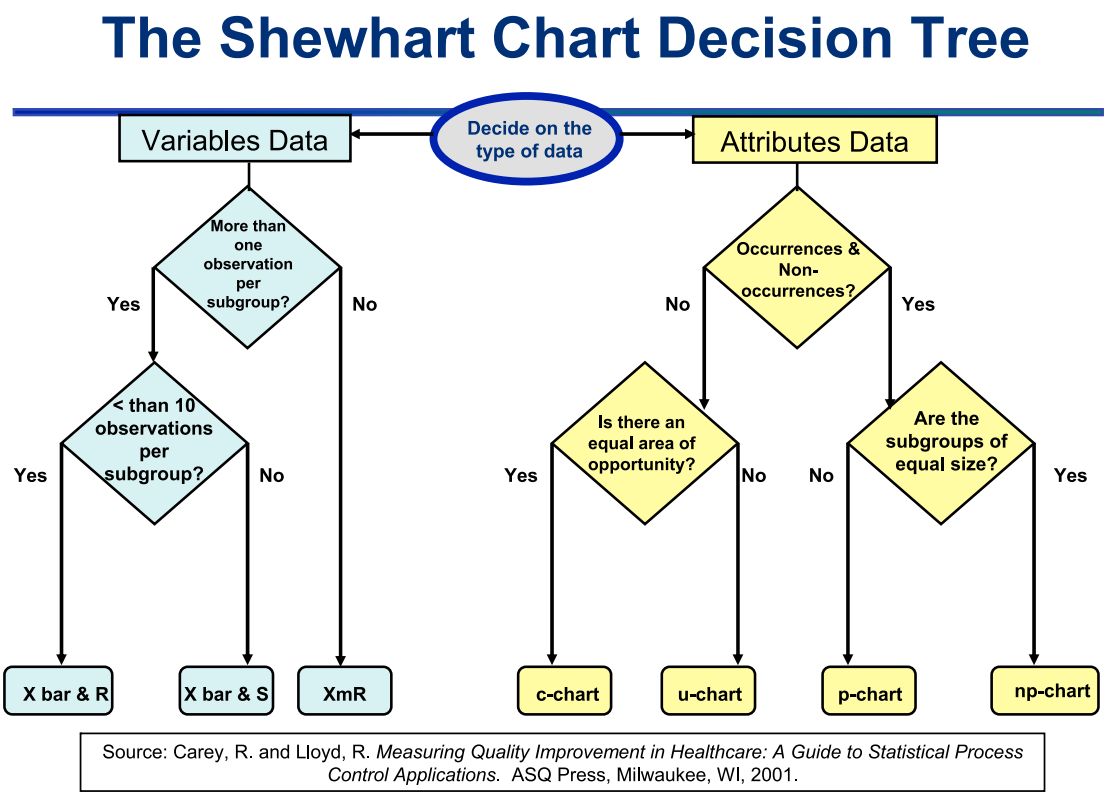

Fig. 9. The Shewhart chart decision tree. (*Courtesy of* Institute for Health Improvement.)

had random or nonrandom data patterns. Because the Shewhart charts perform at a higher level of statistical precision than the run chart, however, the rules to detect common or special causes of variation are more precise. There are rules to identify a shift on a Shewhart chart (8 data points above the centerline rather than 6 on a run chart) and a trend (6 data points constantly going up or down rather than 5 used on the run chart). There are also new rules that the run charts did not have. For example 1 of the rules (called a 3-sigma violation) occurs when a data point exceeds the UCL or LCL. Other rules help to detect what are referred to as abnormal data patterns, and relate to whether the data are bunching toward the outer regions of the chart, or hugging the centerline (ie, too many data clustered in close proximity to the mean). All of these tests are detailed in standard SPC texts.[4,7,17,19] In addition,

Table 2
Key terms in using the Shewhart chart decision tree

Subgroup	Observation	Area of Opportunity
How you organize your data (eg, by day, week or month)	The actual value (data) you collect	Applies to all attributes or counts charts
The label of your horizontal axis	The label of your vertical axis	Defines the area or frame in which a defective or defect can occur
Can be patients in equal or unequal sizes	May be single or multiple data points	Can be of equal or unequal sizes
Can be of equal or unequal sizes	Applies to all the charts	
Applies to all the charts		

SPC software packages automatically mark the presence of special cause variation on a Shewhart chart either by changing the color of the line when a special cause is detected or changing the symbol used to denote a data point on the chart.

In addition to the 7 basic Shewhart charts there are 2 other advanced charts that are appropriate for NICU data. These charts are known as the t-chart and the g-chart.[14] These charts are used when you are faced with 2 conditions. First, when you have small denominators (eg, fewer than 10 observations in the denominator) percentages can become volatile and show extreme swings in variation. For example, 2 out of 4 is 50% but it is not so strong as 50% that is based on 20 out of 40. Small denominators can be 1 reason to use a t- or g-chart. The second reason is that events happen so infrequently that they are considered to be rare. In both these circumstances (small denominators or rare events), the t-chart (plotting time between events) or the g-chart (successful cases between ones considered not successful) provide an alternative to the more traditional p-chart or u-chart. For example, if the unplanned extubation rate which was normally running about 15 per 1000 ventilator days was reduced to 1 or 2 per 100 ventilator days, you should consider moving the measure to a t- or g-chart. In this case we would plot the number of days that went by without an unplanned extubation, or the number of cases that had an ETT and never had an unplanned extubation (ie, a successful application of the ETT during the NICU stay). The goal with either type of chart is to have ever-increasing accumulation of successes without a failure. Every time you have a failure (ie, an unplanned extubation), you start counting the number of days or cases again. This method has been used successfully in manufacturing plants, construction, or the mining industry, where a sign is placed outside the work site stating ''147 days without a workplace injury.'' The next day the sign reads ''148 days'' and so on, until an injury on the job takes the counter back to zero and the count starts all over again.

SPC Examples Using Perinatology Data

Imagine that you are sitting in a meeting designed to review several quality measures for 2 NICUs within your system. The measures of interest for this meeting include:

- Average ventilator days for all babies with birth weight of 501 to 1500 g
- Catheter-associated bloodstream infections (BSIs) per 1000 line days for infants with a birth weight of 501 to 1500 g for NICU1 and NICU2.

Now imagine 2 scenarios for this meeting:

- Scenario 1: you are given tabular data
- Scenario 2: you are given SPC charts.

Think about how you would guide the group's discussion around these measures if you decided to use scenario 1 and the data shown for average ventilator days shown in **Table 3**. What do you conclude from the tabular data? Is the NICU getting better, staying the same, or getting worse? Do we have any special causes in the data? Are the data performing at or near expectations (target), or are the data demonstrating considerable variation and far from target? The tabular data make it difficult to answer these questions. If, on the other hand, we went into the meeting using scenario 2 and distributed the Shewhart chart shown in **Fig. 10**, we would set up a totally different context for the group's discussion. These data reveal the following:

- There is considerable variation in the average ventilator days. The overall average is 25.4 days; the minimum is 8.2 days and the maximum is 67.3 days. Although

Table 3
Average ventilator days and number of patients by month

Month	Average Ventilator Days	Number of Patients
2003/01	18.9	22
2003/02	8.2	20
2003/03	18.1	29
2003/04	26.6	22
2003/05	28.8	24
2003/06	20	14
2003/07	23.6	13
2003/08	27.6	13
2003/09	13.8	18
2003/10	30.2	28
2003/11	13.5	22
2003/12	32.7	36
2004/01	12.9	18
2004/02	12.7	12
2004/03	22.6	28
2004/04	19.9	28
2004/05	26.2	16
2004/06	19.1	11
2004/07	18.7	17
2004/08	46.5	14
2004/09	36.9	16
2004/10	13.1	22
2004/11	31.4	16
2004/12	18.4	26
2005/01	19	17
2005/02	23.3	29
2005/03	16.1	20
2005/04	19.9	19
2005/05	32.7	23

these summary numbers could be calculated from the tabular data, the Shewhart chart provides a visual running record of the variation over time, which is lost in the tabular data.

- With the exception of the last data point (67.3 days), the variation is essentially common cause.
- The last data point is a special cause (above the UCL) and deserves investigation. Is this a data entry error? If it is accurate, then why is this average so high? Remember this is not 1 baby but the average for all 19 babies on a ventilator for the month of December 2008.
- If a target or other comparative reference data are available, the team could determine how far from the target the current process is performing.

Figs. 11 and **12** show the second measure (catheter-associated BSIs per 1000 line days for infants with a birth weight of 501 to 1500 g for NICU1 and NICU2) as a rate

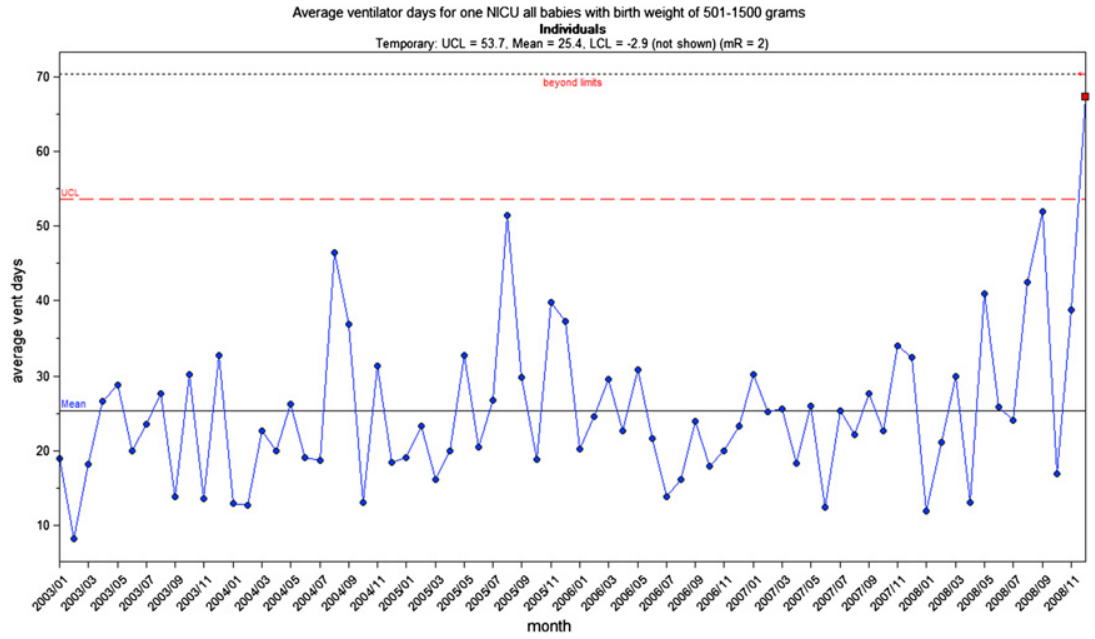

Fig. 10. Average ventilator days for all babies with birth weight of 501 to 1500 g.

(u-chart). Imagine what it would be like trying to make sense out of the tabular data for these 2 NICUs. But at a glance, you can see that the 2 sites have fundamentally different patterns. Questions we can ask include:

- Why does NICU1 have so many rates equal to zero, whereas NICU 2 has few zero points? NICU1 has so many of its data points at zero that this would be a perfect time to move this measure to a t- or g-chart and track the time between BSIs or the cases between BSIs. Note that when you have more than 50% of the data at zero or alternatively at 100%, this observation represents a sign that the t- or g-charts should be considered. The t- and g-charts would not be appropriate for NICU2, however.

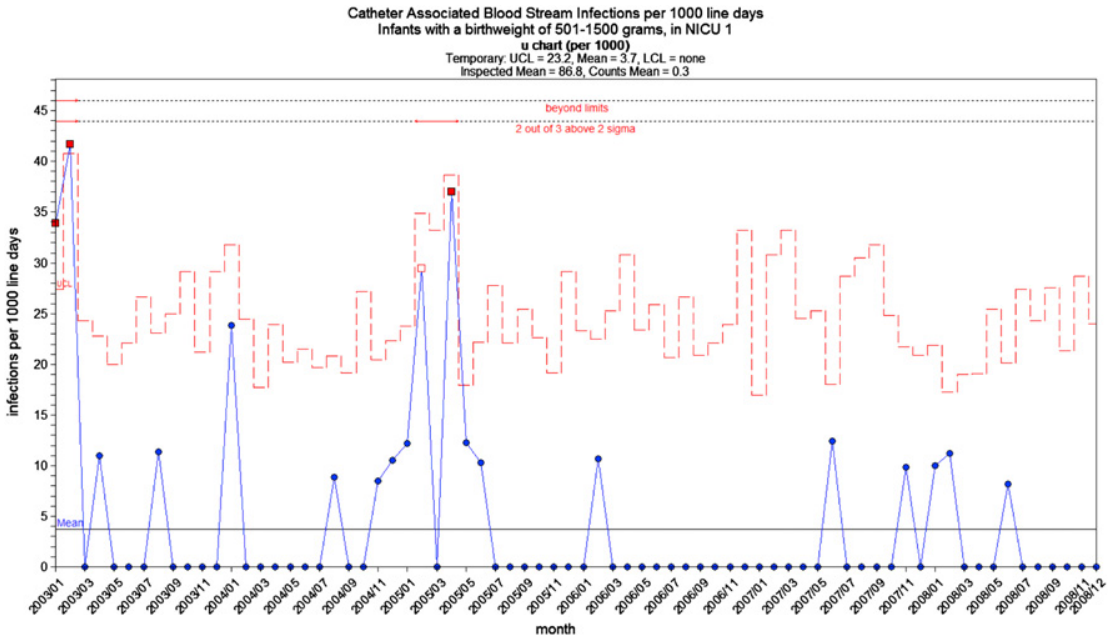

Fig. 11. Catheter-associated BSIs per 1000 line days for infants with a birth weight of 501 to 1500 g for NICU1.

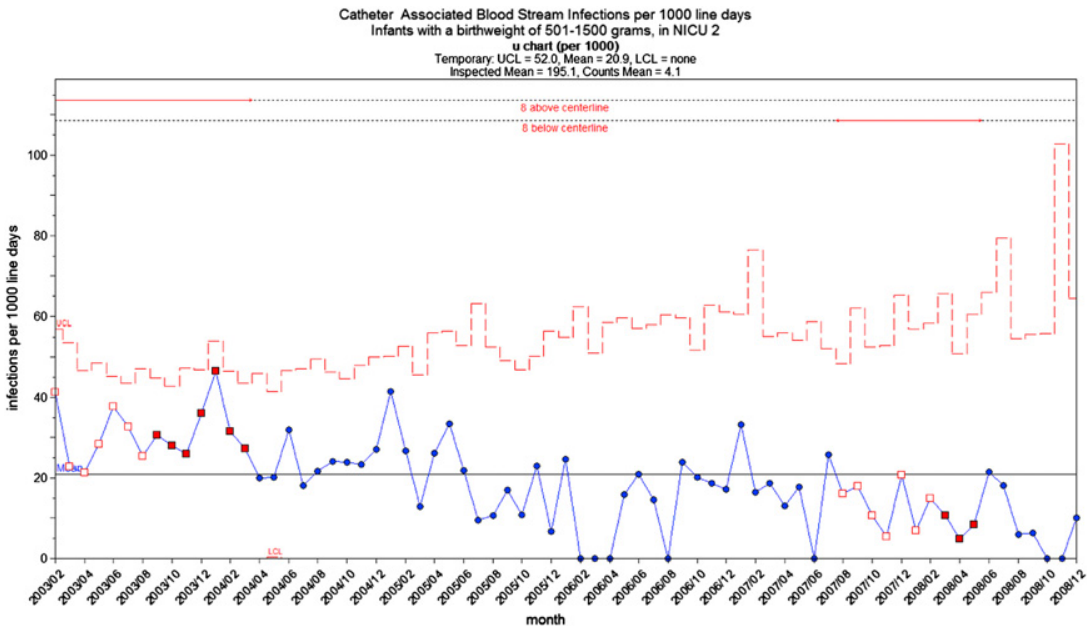

Fig. 12. Catheter-associated BSIs per 1000 line days for infants with a birth weight of 501 to 1500 g for NICU2.

- Note that the average for NICU1 is low, whereas the mean for NICU is considerably higher. Are these units fundamentally different in size, complexity of patients, or types of populations being served?
- The measure has shifted downward at NICU2. This finding signals that improvements may have been put in place. We would want to understand what has caused this downward shift (note the changes in the color of the dots and the connecting lines, which signal special causes in the data). There is also an opportunity to define 2 sets of control limits on the chart. One set would be for the left side of the chart, which is performing at a higher level, and the second set would be used for the data after they shifted downward.

In summary, the Shewhart charts provide a fundamentally different view of the data. The charts should enable dialog and learning. Typically, the tabular data lead the team to engage in shallow levels of learning, boredom, or worse yet, jumping to conclusions. Quality and safety cannot be improved by looking at tabular data and summary statistics. The context for learning comes when you plot data over time and understand the variation in the entire data set.

Linking Measurement to Improvement

Joshua Slocum was well known for keeping detailed diaries and data on his sailing adventures. But he did not collect data and measure his progress just to fill the many lonely hours while circumnavigating the globe. He collected data to help him make better decisions. Slocum was by all accounts a most intriguing yet enigmatic individual. What is clear from reading his diaries, however, is that he understood the linkage between measurement and improvement.

All the preceding milestones and steps in the QMJ are designed to lead to improvement. Data without a context or plan for action give the team a false sense of accomplishment. It is not until you identify change concepts that you believe will move performance in the desired direction and conduct tests of change that the journey becomes complete. All too often health care managers and leaders see data as the beginning and end of the journey. These individuals need to spend a little time with

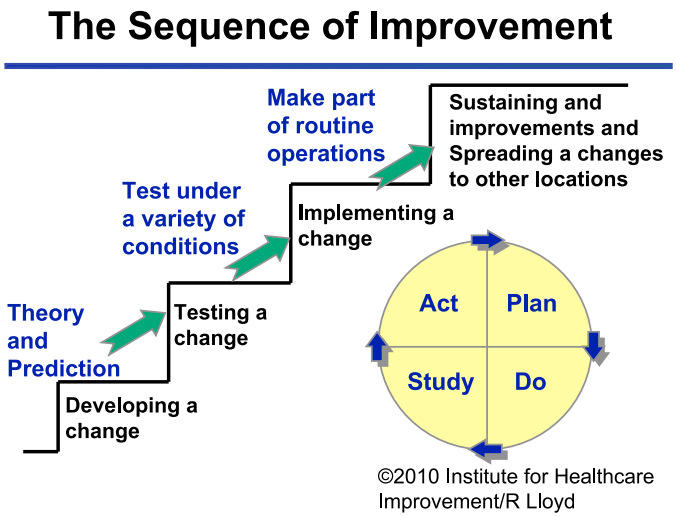

Fig. 13. The sequence for improvement. (*Courtesy of* Institute for Health Improvement.)

Captain Slocum to learn the true value of data collection. Data allow us merely to set the direction of our improvement journey, not define the end of the journey.

The sequence for improvement is shown in **Fig. 13**. Note that although data are used throughout this sequence, the primary objective is to start with small tests of new ideas, build on the success and failures of these tests, and move to testing under different conditions to determine how robust and reliable the new ideas are. When sufficient testing has been accomplished, it is time to implement the new ideas and make them a permanent part of the daily work in the pilot or demonstration area. Once implementation has been successful, it is time to turn your attention to sustaining the gains that have been realized and then start to make plans to spread the improved practices to other locations. Other articles in this issue address the steps in the improvement journey and should be consulted for additional guidance.

ACKNOWLEDGMENTS

The author wishes to acknowledge Dr John Chuo and William Peters for their contributions to this article. Dr Chuo, a neonatologist at Children's Hospital of Philadelphia, provided extensive background information on neonatal unexpected extubations. His willingness to share his knowledge and the planned experiment he has developed to address this issue are greatly appreciated. Peters, an Improvement Advisor and statistician, gave generously of his time to prepare the control charts used in this article.

REFERENCES

1. Teller WM. The voyages of Joshua Slocum. Dobbs Ferry (NY): Sheridan House Inc; 2002.
2. Solberg L, Mosser G, McDonald S. The three faces of performance measurement. Journal on Quality Improvement 1997;23(3):135–47.
3. Brook R, Kamberg C, McGlynn E. Health system reform and quality. JAMA 1996; 276(6):476–80.
4. Lloyd R. Quality health care: a guide to developing and using measures. Sudbury (MA): Jones and Bartlett; 2004.
5. Lloyd R. The search for a few good indicators. In: Ransom S, Joshi M, Nash D, editors. The healthcare quality book: vision, strategy and tools. Chicago (IL): Health Administration Press; 2005. p. 89–116.

6. Deming WE. Out of the crisis. Cambridge (MA): MIT Press; 1992.
7. Provost L, Murray S. The data guide. Austin (TX): Associates in Process Improvement and Corporate Transformation Concepts; 2007. p. 3–15.
8. Babbie ER. The practice of social research. Belmont (CA): Wadsworth; 1979.
9. Deming WE. On probability as basis for action. Am Stat 1975;29(4):146–52.
10. Schultz L. Profiles in quality. New York: Quality Resources; 1994.
11. Swed F, Eisenhart C. Tables for testing randomness of grouping in a sequence of alternatives. Ann Math Stat 1943;XIV:66–87, Tables II and III.
12. Wheeler D. Advanced topics in statistical process control. Knoxville (TN): SPC Press; 1995.
13. Duncan AJ. Quality control and industrial statistics. Homewood (IL): Irwin Press; 1986.
14. Benneyan J, Lloyd R, Plsek P. Statistical process control as a tool for research and health care improvement. Qual Saf Health Care 2003;12:458–64.
15. Carey R. Improving healthcare with control charts. Milwaukee (WI): ASQ Quality Press; 2003.
16. Carey R, Lloyd R. Measuring quality improvement in healthcare: a guide to statistical process control applications. Milwaukee (WI): ASQ Quality Press; 2001.
17. Western Electric Co. Statistical quality control handbook. Indianapolis (IN): AT&T Technologies; 1985.
18. Mohamed MA, Worthington P, Woodall WH. Plotting basic control charts: tutorial notes for healthcare practitioners. Qual Saf Health Care 2008;17:137–45.
19. Wheeler D, Chambers D. Understanding statistical process control. Knoxville (TN): SPC Press; 1992.