Q Manage Health Care
Vol. 22, No. 1, pp. 36–47

Sampling Considerations for Health Care Improvement

Rocco J. Perla, EdD; Lloyd P. Provost, MS; Sandra K. Murray, MA

Sampling in improvement work can pose challenges. How is it different from the sampling strategies many use with research, clinical trials, or regulatory programs? What should improvement teams consider when determining a useful approach to sampling and a useful sample size? The aim of this article is to introduce some of the concepts related to sampling for improvement. We give specific guidance related to determining a useful sample size to a wider health care audience so that it can be applied to improvement projects in hospitals and health systems.

BACKGROUND CONCEPTS

Using data for improvement

Data used for improvement are typically collected by those working within the health care system to monitor process performance, obtain ideas for improvement, test changes to see whether they are improvements, and to see whether improvements are maintained. Run or Shewhart charts are used to determine whether changes yield improvement and improvements are maintained. When collecting data for improvement, we are aware that we are doing so at the same time we are providing ongoing care to our patients and our community. Therefore, sampling is an important consideration to conserve resources and put those resources into testing and adapting changes rather than measuring as teams work toward improvement.

Using sampling for data collection in improvement projects can pose some challenges, though. Many of these challenges are rooted in the differences between expectations of data when they are used for improvement, accountability, or research.[1,2] The environment during an improvement project can be very different from that of a research project, with less emphasis on bias, less control of confounding variables, and not a fixed population, but rather an ongoing process sampled over time.[3] These differences impact the way we sample and the way we

Author Affiliations: *University of Massachusetts Medical School, Worcester (Dr Perla); Associates in Process Improvement, Austin, Texas (Mr Provost); and Corporate Transformation Concepts, Eugene, Oregon (Ms Murray).*

Correspondence: *Lloyd P. Provost, Associates in Process Improvement, 115 E 5th St, No. 300, Austin, TX 78701 (lprovost@apiweb.org).*

Key words: *data for improvement, judgment sampling, sample size, sampling*

The authors declare no conflicts of interest.

DOI: 10.1097/QMH.0b013e31827deadb

determine sample size. Using sampling strategies appropriate for research and accountability purposes when measuring for an improvement project is likely to lead to larger than necessary samples and wasted effort and resources.[4] Sampling is not simply about selecting the "right" computational sample size; all sampling schemes are associated with theoretical assumptions and concepts. Sampling for clinical trials or regulatory purposes most typically assumes a fixed population, a theoretical distribution of measures within this population, and uses these assumptions to devise sampling tables or strategies based on population size. Data for improvement do not come from a fixed population with a known distribution; rather, the data come from an ongoing process whose distribution may change as the future unwinds. Improvement projects in health care take place in an environment that is dynamic and contains many factors that are uncontrolled.[5] This reality impacts an appropriate sampling strategy.

Table 1 summarizes some of the theoretical assumptions underlying measurement for improvement, based on the work of Lewis[6] Shewhart,[7,8] and Deming.[9] On the basis of their guidance, we use the following 2 important concepts related to sampling in improvement projects:

1) Obtaining just enough data based on past experience to guide our learning into the future.
2) Making full use of subject matter expertise in selecting the most appropriate samples.

These 2 pivotal concepts are used as we address sampling for improvement throughout this article.

SAMPLING IN IMPROVEMENT

Sampling at 2 levels in improvement projects

Improvement projects involve both *global project measures* and measurement at the *Plan-Do-Study-Act* (*PDSA*) *cycle level*.[10] Sampling considerations are important at both these levels. Project or global measures focus at the project level and are maintained throughout the life of the improvement project. Data will be collected and summary statistics graphed for each of these measures at a regular time interval, using an annotated run or Shewhart chart. Typically, a project will have a "family" of 3 to 8 global outcome, process, and balancing measures. Table 2 defines these 3 categories and provides an example of each for a perioperative safety team working to reduce harm. In subsequent sections of this article, we address sample size for global measures.

Measures used during PDSA cycles for testing are more transient and usually do not endure for the entire lifespan of the improvement project. When the PDSA cycle is being used to test changes, the approach is sequential. Initial PDSA cycles testing an idea for improvement will typically start at a small scale (with a small sample size). As knowledge about the change increases, follow-up cycles are used to adapt and retest the idea, test it under more robust

Table 1

THEORETICAL ASSUMPTIONS ASSOCIATED WITH SAMPLING FOR THE PURPOSES OF IMPROVEMENT

Theorist	Theory/Philosophy	Implications on Sampling
C. I. Lewis	Conceptualistic pragmatism	Past experience influences what we expect to occur in the future and this prediction is based on small amounts of information being processed over time
Walter Shewhart	Cause systems	We need just enough observational data to tell us if our systems of production are stable or unstable
W. Edwards Deming	Profound knowledge[a]	Those closest to the process will have the greatest insight on where and how to obtain the most useful samples

[a]Combined with appropriate subject matter knowledge.

Table 2

GLOBAL MEASURES: OUTCOME, PROCESS, AND BALANCING[a]

Type of Measure	Description	Perioperative Example
Outcome	The voice of the customer or patient How is the system performing? What is the result?	% of patients harmed % unplanned returns to the operating room % unplanned surgical readmission
Process	The voice of the workings of the process Logically linked to obtaining the outcomes Address how key parts/steps of the system are performing	% of patients with on-time antibiotic administration % of patients with appropriate DVT prophylaxis % of patients with appropriate β-blocker use
Balancing	Look at a system from different directions/dimensions What happened to the system as we improved the outcome and process measures? Could relate to unintended consequences or competing explanations for project success	Volume of surgical workload % of prophylactic antibiotics appropriately discontinued

Abbreviation: DVT, deep vein thrombosis.
[a]Adapted from Provost and Murray.[11]

conditions, or test on a larger scale. Sample size during these PDSA test cycles will vary depending upon a number of factors.

Sampling during PDSA cycles

In improvement, the goal of sampling is to select observational units in order to learn about systems and improve them in the most timely, efficient, and effective manner possible. Testing using the PDSA cycle is done in the context of human and social systems. We can plan an effective sampling strategy during PDSA cycles, using 3 major concepts related to this human context.[10] The size or scale (number of tests, time required, number of staff involved, amount of data) of a PDSA test of change should be decided by considering:

1. the team's degree of belief that the change will result in improvement;
2. the costs associated with a failed test; and
3. the readiness of those who will have to make the change.

As shown in Table 3, considering these 3 factors enables one to determine the size or scale, and thus the appropriate sample size, for the next PDSA cycle. For example:

- A team is testing a new admitting approach designed to improve billing accuracy and reduce time to bill. The team considers this change likely to produce great results. The staff involved in testing the idea was not eager. Negative consequences for their system could be considerable, however, if the new system does not work. Using the matrix, they determined that a very small-scale test would be appropriate (high degree of belief, high cost of failure, resistant staff). Their initial sample size was to try the change with 1 admission.
- The new online decision support guidelines had been reviewed by all of the clinicians and approved. The team had run some tests on the system, and it had worked well. If for some reason the system failed, a paper backup was ready for use. The team was tempted to move to implementation but decided on a large-scale test of change to build its confidence that it would work under all conditions (high degree of belief, cost of failure low, staff ready). Its testing

Table 3

DECIDING THE SCALE OF A TEST[a]

		Appropriate Scope for a PDSA Cycle		
		Staff/Clinician Readiness to Make Change		
Current Situation		**Resistant**	**Indifferent**	**Ready**
Low belief that change idea will lead to improvement	Cost of failure large	Very small-scale test	Very small-scale test	Very small-scale test
	Cost of failure small	Very small-scale test	Very small-scale test	Small-scale test
High belief that change idea will lead to improvement	Cost of failure large	Very small-scale test	Small-scale test	Large-scale test
	Cost of failure small	Small-scale test	Large-scale test	Implement

Abbreviation: PDSA, Plan-Do-Study-Act.
[a]Adapted from Langley et al.[10]

strategy for the next PDSA cycle consisted of piloting the guidelines in the entire organization for a month (large sample test).

When using the matrix in Table 3 to determine the scale of the next PDSA cycle, judgment is used to determine what constitutes a "very small," "small," or a "large" test of change. The decision about sample size for data collected as part of a PDSA cycle is to obtain just enough data to answer the questions posed in that cycle. When it turns out that additional information is needed, a follow-up cycle is available to obtain additional samples.

Selecting the sample

In sampling for improvement, it is useful to understand the concepts of probability- and nonprobability- (or judgment) based sampling. Deming[3,12] first introduced the term "judgment sampling" as a contrast to probability-based sampling in the context of surveys in 1947. He defined the difference as follows:

Probability samples: Samples selected by a random process for which the sampling errors can be calculated, and for which the biases of selection, nonresponse, and estimation are virtually eliminated or contained within known limits. Also called random sampling, the various methods of probability sampling (eg, simple, stratified, and systematic random samples) are discussed in basic statistics texts so are not discussed further here.[3]

Judgment samples: Samples selected by a nonrandom process for which the biases and sampling errors cannot be calculated from the sample but instead must be settled by subject matter knowledge. In research, the term purposive sampling has sometimes been used to describe a type of judgment sample.[13]

The distinction between probability and judgment sampling is important for quality improvement efforts and is the reason sampling strategies for improvement often differ from those used in clinical trials and accountability reports. Data used for improvement most often come from a stream of data from an ongoing process (eg, health care delivery process); we rarely have a fixed population of interest from which to select a random sample. Even if we could identify such a population at a point in time, this population would quickly change as we move into the future.[12] In addition, in improvement work, ensuring that all potential observational units in a population and sampling frame have equal probability of selection is often not the most desired or beneficial strategy. Rather, often we look to the subject matter experts to guide which areas, times of day, or segments of the population are most important to study and understand. For these reasons, judgment sampling is very useful in improvement work. Some examples of judgment samples are as follows:

- Select the times of day we should collect STAT laboratory turnaround times.
- Select the first 10 patients who arrive in the clinic after 2:00 PM.
- Select the charts from only patients with 3 or more comorbidities.
- Interview the next person with diabetes who comes in the office for care.

The desirability of judgment sampling is, in many ways, obvious whether the aim is to learn about a system or process that includes various factors and inputs that are likely to change over time. With judgment sampling, while we lose the ability to assess the precision of our results using traditional statistical methods, we gain the ability to generalize on the basis of samples selected under a wide range of conditions and over time as improvements are made. The work of Perla and Provost[14] provides more detail on the use of judgment samples in improvement.

Global project measures are those that are collected and displayed on an annotated run or Shewhart chart at least monthly for the duration of the improvement project. Sampling for global project measures involves deciding how often, how much, and which data we are going to collect in order to help us improve. Sometimes, the number of patients or other volume of work available is small enough that it makes sense to obtain all of the data in the set (ie, we only have 7 people with newly diagnosed diabetes each month, so we obtain data from all of them). And sometimes, electronic records (eg, use of time stamps for emergency department wait times) make large amounts of data readily available without having to sample. But when working with a large number of units requiring manual data collection, sampling can be an efficient way to collect enough data to accurately track performance yet save time and resources while doing so.[15]

SAMPLING CONSIDERATIONS USING RUN AND SHEWHART CHARTS

Improvement projects are, by their temporal nature, focused on performance as it moves forward in time in health care systems. Without a defined population, we cannot directly use probability-based sample size tables and formulas that are often used in research study designs and for regulatory applications. Despite this, we do have some practical guidance for obtaining appropriate sample sizes for improvement-related data where our selection is usually based to some degree on judgment.

Sampling for measures on run charts

A primary use of the annotated run chart in an improvement project is detection of improvement as changes are tested and implemented. The importance of viewing the impact of changes tested on the system over time sets a context for determining sample size for the test.[11] The number of data points (subgroups) plotted on the run chart needs to be adequate to detect patterns that may indicate improvement and impacts the overall sampling burden. Some guidelines for determining the number of data points plotted necessary to detect patterns on a run chart are shown in Table 4. Run charts are started when the first data point is available on a project. As each additional data point becomes available, it is added to the run chart for continuous learning. Analysis is principally visual, although rules for detecting improvement may be applied when enough data points

Table 4

MINIMUM NUMBER OF DATA POINTS FOR AN EFFECTIVE RUN CHART[a]

Situation	Data Points Required
Expensive tests, complex prototypes, or long periods between available data points, large effects anticipated	<10
Desire to discern patterns indicating improvements that are moderate or large	11-30
The effect of the change is expected to be small relative to the variation in the system	31-100

[a] Adapted from Provost and Murray.[11]

are plotted.[16] The run chart typically becomes quite powerful when 10 or more data points are available. The sampling burden for run charts then is impacted by both the frequency of subgroups and the size of the sample used to define subgroups. For example, a team wanted to develop and test changes to improve the turnaround time for a particular laboratory test. The team decided to obtain a judgment sample of 5 test times each day and plot the average turnaround time daily on a run chart. The team's sampling burden was 35 samples each week (5 per day × 7 days each week).

For data used as part of an improvement effort, getting samples across a wide range of conditions (locations, days of the week, shift, etc) is almost always more important than the number of samples collected under a specific condition. In general, more data (larger sample sizes) lead to more information and better precision of results. Unless the data are already collected and reported, larger sample sizes also involve more effort and cost.[4] As an illustration, in the simulated run charts in Figure 1 (adapted from the work of Provost and Murray[11]), we see average clinic waiting time plotted during the life of the project. The examples were simulated using results randomly generated with a mean of 50 minutes (SD = 10 minutes). An important improvement (a 30% reduction) in the average waiting time has been made after period 12. The 5 run charts show the simulated average waiting times of a sample of 1, 5, 10, 20, and 50 patients per week.

Is it obvious that a change in the waiting time has occurred looking at the data plotted with a sample size of 1 per week? What about a sample size of 5, 10, 20, and 50? The picture is not evident with a sample size of 1. The information gained with a sample size of 5 and 10 begins to provide convincing visual evidence that the change resulted in improvement. With a sample size of 10 observations per week, we observe the first signs of a nonrandom data pattern for a run chart in the form of a shift and too few runs.[11] A sample size of 20 or 50 per week is probably overkill and wasteful. Sample size issues in improvement efforts are a balance between resources (time, money, energy, slowing improvement efforts)

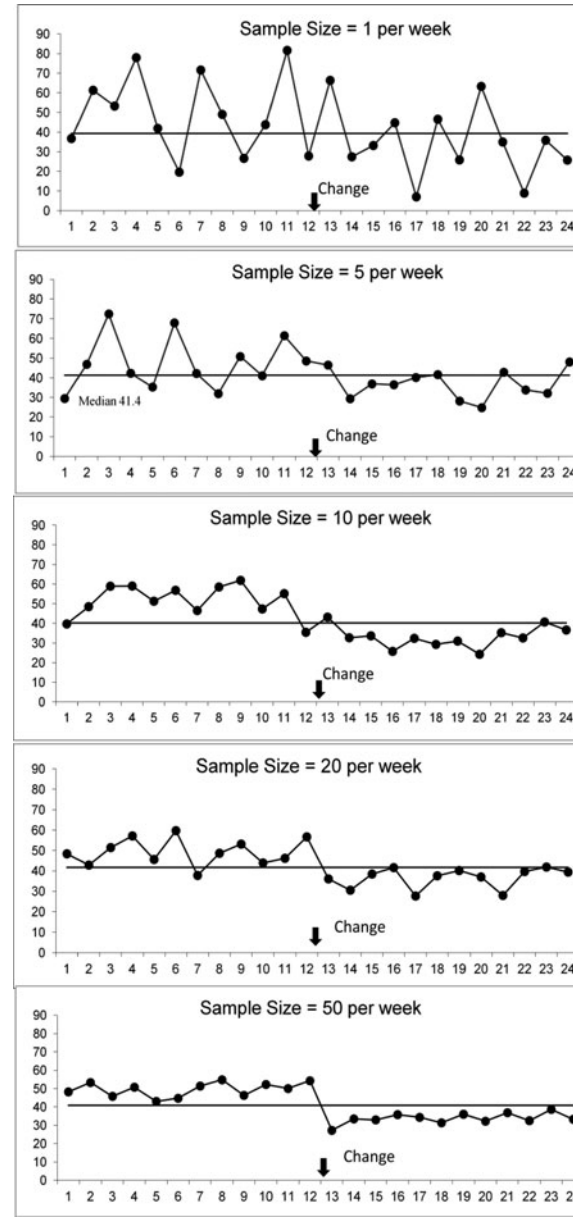

Figure 1. Sample size on a run chart and ability to detect change.

and the precision of the results desired.[4] We advise improvement teams to consider smaller samples at any one time but commit to collecting and graphing them over time (eg, daily, weekly, monthly) as a way to improve learning and conserve team resources. A visual review of simulated run charts for the project's

key measures can provide the "power calculation" usually done in research projects (which also assume baseline performance values).

Sampling for Shewhart charts

Shewhart charts are a pivotal tool for improvement work, providing an operational definition for Shewhart's theory of common and special causes.[7] These time-series charts are used to develop improvement strategies, detect the impact of changes, determine sustainability, and calculate process capability. While limits can be calculated with a small number of data points (called subgroups), they become most powerful when the limits are calculated using 20 or more subgroups. Shewhart charts are analyzed using statistically based rules. Shewhart charts come in a variety of types depending on the type of data being used and other factors.[11] Here, we address sampling for the most common types of Shewhart charts used in health care.

Classification of data: P charts

One of the most commonly used Shewhart charts is the P chart, which is used to analyze percentages based on classification (yes/no, compliance/noncompliance, conforming/nonconforming) data. The limits for the P chart are based on the standard error of a binomial distribution. To develop an effective chart, it is desirable to avoid too many zeros, to have a symmetric distribution of the statistic plotted, and sometimes important to have both an upper and lower control limit to detect both improvement and deterioration. The sample size used to create the subgroups plays the primary role in achieving these characteristics. A sample size large enough to be able to detect the nonconforming units is needed in order to tell whether the percentage measure is improving. For example, if we had a process in which 1% of the patient assessments did not meet standards and we collected a sample of 10 assessments each week, we would be unlikely to find any in that sample of 10 that did not meet standards, much less be able to tell when we improved the process to less than 1% noncompliance.

The minimum subgroup size for an effective P chart depends on the average percentage (P_{bar}). Table 5 (adapted from the work of Provost and Murray[11]) summarizes the minimum subgroup size required to create P charts with specific features:

1. To expect 25% or less of the subgroups with a subgroup value of 0 (or 100%) that is necessary for a useful chart.
2. To expect a symmetric distribution of P based on the guideline of $n > 300/P$.
3. To expect a lower control limit for the P chart (or upper limit if near 100%).

These sample size guidelines are useful for both random and judgment samples. The decision to use 1 of the 3 strategies (columns in Table 5) to select a subgroup size for P charts will be impacted by cost, feasibility, and data access. To make this decision, use the following decision scheme:

Step 1: *If feasible to use a subgroup size that provides a lower limit, then select this option*. The power of the Shewhart chart over a run chart is the limits used to detect special causes.

Step 2: *If not possible to select a subgroup size that provides a lower limit, then consider a subgroup size that satisfies* $n > 300/P$. When the subgroup size meets this criterion, the distribution of the data points is expected to be symmetric enough that all the rules for special causes used with the chart will be useful.

Step 3: *If not possible to select a subgroup size that satisfies steps 1 or 2, then select a minimum subgroup size needed to have 25% or less zero values for P.* At a minimum, construction of a P chart should use a subgroup size needed to have 25% or less zero values for P.

Step 4: If it is not possible to obtain subgroup sizes to meet this third criterion, the P chart will not be useful. An alternative Shewhart chart tracking count or time between rare events data (G or T chart) should be considered.[11]

As an example of the use of Table 5, a team was interested in determining a sample size for percentage of eligible patients discharged from a particular unit

Table 5

MINIMUM SUBGROUP SIZE FOR P CHARTS BASED ON 3 DIFFERENT CRITERIA[a]

Estimated Average % Nonconforming Units (P_{bar})	Minimum Subgroup Size (n) Needed to Have ≤25% Zeros for *P*s	Minimum Subgroup Size Guideline (n >300/*P*)	Minimum Subgroup Size Needed to Have LCL >0
0.1	1400	3000	9000
0.5	280	600	1800
1	140	300	900
1.5	93	200	600
2	70	150	450
3	47	100	300
4	35	75	220
5	28	60	175
6	24	50	142
8	17	38	104
10	14	30	81
12	12	25	66
15	9	20	51
20	7	15	36
25	5	12	28
30	4	10	22
40	3	8	14
50	2	6	10

Abbreviation: LCL, lower control limit.
[a]For $P > 50$, use $100 - P$ to enter the table (eg, for $P = 70\%$, use table P of 30%; for $P = 99\%$, use table P of 1%; etc).
Adapted from Provost and Murray.[11]

who received prophylaxis for deep vein thrombosis during their hospital stay. Their historical data revealed that 30% of the patients did not receive deep vein thrombosis prophylaxis. Using the guidance in Table 5, it determined that using a sample size of 22 would yield both upper and lower limits on the chart. This unit had as few as 40 and as many as 120 patients discharged per week. This team wanted to obtain a data point each week, and had sufficient data to do so, but did not have an automated system from which to obtain the data. The team realized that a manual record review of 22 records per week was too time consuming, given its resources. The team decided to obtain a sample of 6 medical records per week for record review and plot a data point monthly. By aggregating weekly data to monthly data, it was able to obtain a subgroup size that produced a chart with both lower and upper control limits.

Count data: C and U charts

For count data (the number of errors, instances, occurrences, or nonconformities in the subgroup), the standard error from the Poisson distribution is used to develop the limits on the Shewhart C or U chart. The minimum sample size is related to the opportunity for occurrences (time period, physical area, or number of units) that determines the average number of nonconformities (C_{bar}) or average rate (U_{bar}). Table 6 provides guidance for determining opportunity size (subgroup size or area of opportunity) for each data point plotted for a C chart.[11]

In selecting the sample size (area of opportunity) for a subgroup on the C chart, select the time period, physical or geographic area, or number of units it takes to expect at least an average of 1.4 or more nonconformities to get a useful C chart, or, ideally, where the average is greater than 9.0 to get a lower

Table 6

GUIDELINE FOR C CHART CONSTRUCTION

To Have a Lower Control Limit	To Have ≤25% Zeros
C_{bar} must be >9.0	C_{bar} must be ≥1.4

limit. For example, a team wanted to reduce billing errors for a particular category of bill. Its historic data averaged 0.6 billing errors per invoice. It decided that a sample of 20 bills would safely yield more than 9 errors and would result in an effective C chart with a lower limit to quickly detect improvement. This organization generated about 55 bills a week of the type the team was interested in tracking. The team decided that it could obtain a sample 20 bills every 2 weeks and plot the resulting data on a C chart.

This guideline for a C chart (C_{bar} >9 for a lower limit and C_{bar} >1.4 for a useful chart with ≤25% zeros plotted) can also be used to design effective U charts. In U charts, the "area of opportunity" (subgroup size) is allowed to vary. So, to determine a useful sample size, a "standard" area of opportunity (1 day, 100 admissions, 1000 catheter days, 10 000 deliveries, 100 000 hours worked, etc) is set and the "number of standard areas" needed to meet the criteria in Table 6 is determined using the following steps:

- Pick an appropriate standard area of opportunity (based on subject matter knowledge).
- Estimate the current average rate for the measure of interest.
- Divide 1.4 by the average rate to get the minimum number of "standard areas of opportunities" required to have a useful U chart.
- Then divide 9 by the average rate to get the minimum number of "standard areas of opportunities" required for the U chart to have a lower limit.
- Use the resulting information to plan the subgroup size.

For example, a team was chartered to reduce the hospital-wide pressure ulcer rate. The team wanted to develop a U chart for its baseline data from the past 2 years. The pressure ulcer rate was currently reported monthly as 5 per 1000 occupied bed days. To determine the minimum subgroup strategy, the team divided 1.4 by 5.0 to get 0.28 standard areas of opportunities. So to develop a useful U chart, it needed a sample of 280 bed days (0.28 × 1000 occupied bed days). This meant the team could have separate monthly charts (or subgroups) for wards with more than 10 occupied beds (10 beds × 30 days per month = 300 occupied bed days per month). To get a U chart with a lower limit, the team would need 1.8 standard areas of opportunity (9 divided by 5.0) or 1800 bed days. This would work well for the hospital-wide monthly chart since the hospital had 110 beds and typically had more than 3000 occupied bed days each month.

Shewhart charts for continuous data (I charts and X_{bar} S and I charts)

When planning the measurement strategy for improvement projects, measures based on a continuous scale are usually preferred to count or classification data when the option is available. We can learn more quickly from continuous than attribute data (measuring rare events is one exception to this). So, for example, in learning about length of stay (LOS), creating subgroups that contain LOS data from 3 to 10 patients will provide more learning about the causes of variation than subgroups of 30 to 300 attribute measures of LOS (eg, percentage of times >3 days).

When multiple samples are available to create subgroups, the X_{bar} (average of subgroup data) and S charts (standard deviation of subgroup data) can be used effectively with subgroup sizes of as few as 2 or 3. These charts can also handle very large sample sizes such as those from administrative or automated databases. The sensitivity of an X_{bar} chart depends on the variation of the measure of interest. This variation is documented and studied in the S chart. The central limit theorem shows that the precision of averages of multiple data points is greater than that of the original data and therefore the width of the control limits on an X_{bar} chart is inversely related to the square root of the subgroup size. This relationship can be used to design an X_{bar} chart with limits

that are equal to changes of specific magnitude using: $n = (3 \times \sigma/c)$,[2] where n is the required subgroup size, sigma is the common cause variation of the measure documented on the S chart (average within subgroup standard deviation), and c is the size of change of interest. For example, if an hourly length of stay measure had a σ of 20 minutes and it was desired to detect a change of 10 minutes, $n = (3 \times 20/10)^2 = 36$ for a subgroup size that will give a limit 10 minutes above or below the center line (average) on the X_{bar} chart. So if a sample size of n = 36 patients was used to define each subgroup, then each subgroup after a shift in the average of 10 minutes would have a 50% chance of exceeding the control limits and thus signal a special cause. If it was only important to detect a change of 30 minutes, a subgroup size of n = 4 would be adequate to give limits that are plus or minus 30 minutes from the center line.

For the I chart, each individual measure is used to define the subgroup. So the "sample size" issues have to do with the number of subgroups and the guidance for run charts is useful.

Power calculations for Shewhart charts

The issue of power comes up when people ask about the "sensitivity" of the Shewhart chart to detect changes. In the improvement literature, power calculations lead to operating characteristic (OC) curves, but an assumption of random sampling is needed to perform these calculations.[17] This step is very acceptable to researchers and others who use random selection in their work.[13,18] However, as mentioned earlier, improvement efforts often use judgment samples. If the subject matter expert thinks that the method of selecting samples will simulate random sampling, viewing OC curves for different sample sizes under various assumptions can be helpful in developing an effective measurement strategy. OC curves can still be useful as indicators of the sensitivity of Shewhart charts when using such samples.

A health system wanted to use a Shewhart P chart to monitor the proportion of patients in its different long-term care facilities compliant with medication reconciliation on admission to gauge the impact of different improvement attempts. Each facility was at a different stage of testing and implementing changes. Various scenarios were discussed and alternative sampling plans were created with the ability to detect an improvement as a signal outside the control limits on the P chart (a point above the upper limit). OC curves were graphed for alternative monthly subgroup sizes (sample sizes) that show the chance of detecting a change each month using the Shewhart chart for different amounts of improvement (Figure 2).

One of the long-term care facilities had a baseline compliance of 30% and wanted to know what sample size would be needed to detect within 1 month an improvement in compliance by 30% using the P chart. Looking at Figure 2A, they settled on a monthly sample size of 33 randomly selected patient records that gave them a 79% chance of detecting a 30% improvement in 1 month.

Figure 2A-C shows some of the possible OC curves that could be used to determine the power of a Shewhart chart in this situation. With just a small amount of baseline information and knowing what degree of improvement is desired, managers and leaders can use these curves to more effectively select subgroup sizes for various projects and balance power to detect change with the resources needed to collect additional samples. The key assumption made is the sampling method deployed simulates a random sample. The subject matter experts have to make this decision on the basis of their knowledge of the sampling method deployed. We have found many situations in which these calculations were useful in developing sampling strategies where various forms of judgment samples were used.

SUMMARY

This article provided background on sampling concepts and methods for quality improvement, making an important distinction between probability and nonprobability sampling. It emphasized the importance of judgment sampling in improvement work. Sample size considerations for improvement were

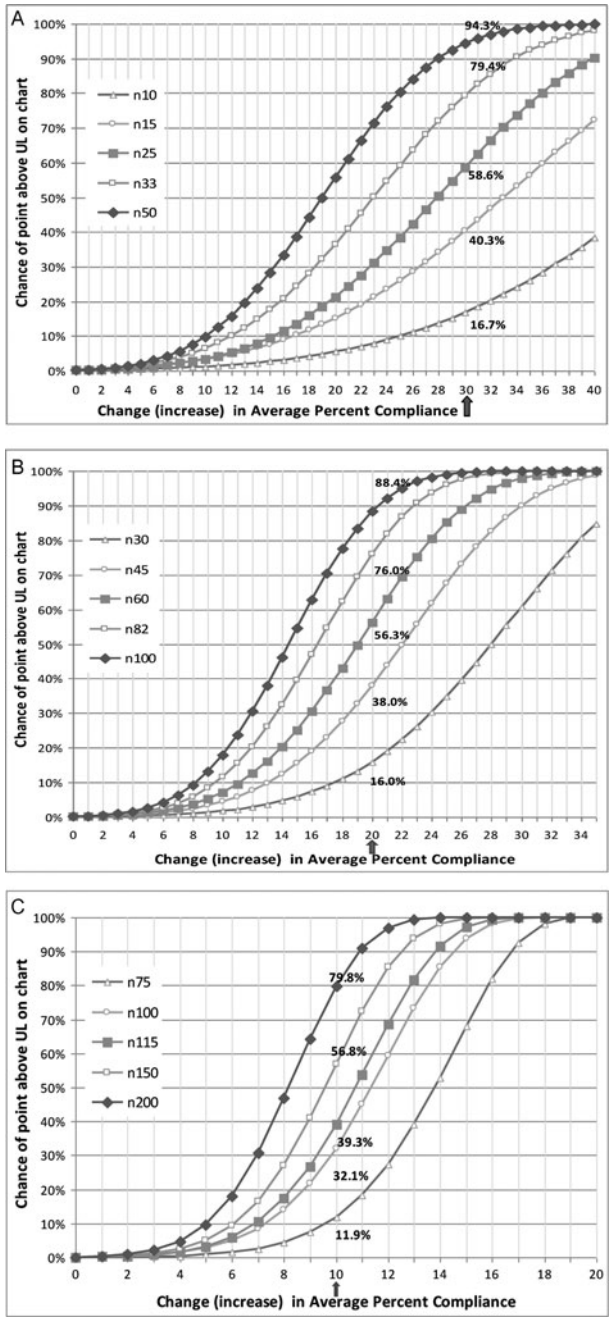

Figure 2. Chance of detecting improvement in medication reconciliation using operating characteristic curves (arrow indicates percent change desired to be detected). **(A)** Chance of detecting improvement on P chart with 30% baseline performance. **(B)** Chance of detecting improvement on P chart with 50% baseline performance. **(C)** Chance of detecting improvement on P chart with 80% baseline performance.

discussed both for the PDSA level of measurement and for global project measures. The importance of viewing the impact of the change on the system over time sets a context for determining sample size when using run and Shewhart charts. With Shewhart charts for attribute data, such as the P, C, and U charts, it is possible to address the subgroup sizes needed to detect improvement using criteria related to the average percent nonconforming (P chart) or average rate of nonconformities (C and U charts). When using Shewhart charts for continuous data, learning occurs more quickly and requires smaller subgroup sizes. OC curves can be used to effectively determine subgroup size when the subject matter expert believes that the judgment sampling method simulates random samples.

REFERENCES

1. Solberg LI, Mosser G, McDonald S. The three faces of performance measurement: improvement, accountability and research. *Jt Comm J Qual Improv*. 1997;23:135-147.
2. Perla RJ, Provost LP, Lloyd RL. Accountability measure to promote quality improvement. *N Engl J Med*. 2010;363:1975-1976.
3. Deming WE. *Some Theory of Sampling*. New York, NY: Wiley; 1950:10-11. (Reprinted by Doer Publishing 1960.)
4. Perla RJ, Allen BD. Balancing cost and precision in hospital accountability sampling. *J Healthc Qual*. 2009;33:5-9.
5. Provost LP. Analytical studies: a framework for quality improvement design and analysis. *BMJ Qual Saf*. 2011;20 (suppl 1):i92-i96.
6. Lewis CI. *Mind and the World Order: Outline of a Theory of Knowledge*. New York, NY: Dover Publications; 1929.
7. Shewhart WA. *The Economic Control of Quality of Manufactured Product*. New York, NY: D Van Nostrand; 1931.
8. Shewhart WA. *Statistical Method From the Viewpoint of Quality Control*. Washington, DC: Department of Agriculture, The Graduate School; 1939.
9. Deming WE. *Out of the Crisis*. Cambridge, MA: MIT; 1982.
10. Langley GJ, Moen RD, Nolan KM, et al. *The Improvement Guide: A Practical Approach to Enhancing Organizational Performance*. San Francisco, CA: Jossey-Bass; 2009.
11. Provost LP, Murray SK. *The Healthcare Data Guide: Learning From Data for Improvement*. San Francisco, CA: Jossey-Bass; 2011.
12. Deming WE. On probability as a basis for action. *Am Stat*. 1975;29:146-152.
13. Onwuegbuzie AJ, Leech NI. A call for qualitative power analyses. *Qual Quantity*. 2007;41:105-121.
14. Perla RJ, Provost LP. Judgment sampling: a health care improvement perspective. *Qual Manag Health Care*. 2012;21(3):169-175.

15. Cochran WG. *Sampling Techniques*. New York, NY: John Wiley & Sons Inc; 1977.
16. Perla RJ, Provost LP, Murray S. The run chart: a simple analytical tool for learning from variation in healthcare processes. *BMJ Qual Saf*. 2011;20:46-51.
17. Wadsworth HM, Stephens KS, Godfrey AB. *Modern Methods for Quality Control and Improvement*. New York, NY: John Wiley & Sons; 1986.
18. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. *Emerg Med J*. 2003;20:453-458.