Sampling plays a major role in quality improvement work. Random sampling (assumed by most traditional statistical methods) is the exception in improvement situations. In most cases, some type of “judgment sample” is used to collect data from a system. Unfortunately, judgment sampling is not well understood. Judgment sampling relies upon those with process and subject matter knowledge to select useful samples for learning about process performance and the impact of changes over time. It many cases, where the goal is to learn about or improve a specific process or system, judgment samples are not merely the most convenient and economical approach, they are technically and conceptually the most appropriate approach. This is because improvement work is done in the real world in complex situations involving specific areas of concern and focus; in these situations, the assumptions of classical measurement theory neither can be met nor should an attempt be made to meet them. The purpose of this article is to describe judgment sampling and its importance in quality improvement work and studies with a focus on health care settings.

WHAT IS A JUDGMENT SAMPLE?

A judgment sample is a type of nonprobability sample, which is selected on the basis of knowledge of

Key words: health care, judgment sampling, quality improvement

Author Affiliation: University of Massachusetts Medical School, Worcester, Massachusetts (Dr Perla); Associates in Process Improvement, Austin, Texas (Mr Provost).

Correspondence: Rocco J. Perla, EdD, University of Massachusetts Medical School, Department of Quantitative Health Sciences, 55 Lake Ave North, ACC Building 7th Floor, Worcester, MA 01655 (Rocco.Erdos4@gmail.com)

No conflicts of interest or funding sources to declare by all authors in association with this work.

DOI: 10.1097/QMH.0b013e31825e8806
a subject matter expert with knowledge of the process being studied. Deming first used the term as a contrast to probability-based sampling in the context of surveys in 1947 in *The Journal of Marketing* and again in his book *Some Theory of Sampling* in 1950. Deming described the difference as follows:

**Probability samples**—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.

**Judgment samples**—for which the biases and sampling errors cannot be calculated from the sample but instead must be settled by judgment.

Deming notes the following:

> It is more important to learn something about the biases of a judgment-sample than about its sampling errors. The usefulness of data from judgment-samples is judged by expert knowledge of the subject matter and comparisons with the results of previous surveys, not from knowledge of probability. A skilled statistical theorist would be helpless in the analysis of a judgment-sample if he were to depend on his knowledge of theory.

The point to stress here is not that some statistical calculations cannot be useful with judgment samples, but rather since there is no random sampling from a population, there is no theory to associate any probability statements or confidence level with such calculations. Both Deming and Shewhart were clear that any application of statistical concepts to understand the voice of the process (e.g., 3 sigma control limits) would have to be based on statistical logic and experience and therefore be viewed as an analytical heuristic versus a purely mathematical justification of method. Deming and others apply judgment sampling not only to surveys but to all methods to collect data for improvement, learning and experimentation where those closest to the process (subject matter experts) are required to help guide sample selection and stratification.

The distinction between probability and judgment (nonprobability) sampling is critical for quality improvement efforts in at least 3 ways. First, as mentioned earlier, because we rarely identify a population of interest from which to select a random sample, we cannot rely on classical statistical theory to associate any probability or confidence level to our observations. Second, even if we could identify such a population at a point in time, this population would quickly change as we predict performance into the future. Third, 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; in many cases, we look to the subject matter experts to guide which areas, segments, or fragments of the process are most important to study and understand.

Closely related to judgment sampling are 2 other types of sampling: convenience sampling and mechanical sampling. A convenience sample is also known as an opportunity sample, an accidental sample, or a “grab” sample. The specific sample is taken because it is readily available and easy to select: the first 5 charts in the file, the next 3 patients who register, the vials in the box that is already open, etc. Convenience sampling is probably the most common sampling method used because of its informality, simplicity, and cost-effectiveness. It can be considered a type of judgment sampling where a minimal amount of “judgment” is used in the decision to select a particular sample. The judgment might be as simple as that, for the purposes of learning at this time, and the sample will provide the information needed.

Mechanical sampling is most closely related to sampling of bulk materials. Various types of mechanical devices are used to select a sample from a larger amount of bulk material. For example, take a scoop from a bag of flour. In health care, mechanical methods are used to select tissue or fluid samples from patients. Deming emphasized study of the bias of mechanical sampling methods when they are used.

**APPLICATIONS OF JUDGMENT SAMPLING**

Table 1 highlights the difference in methods and perspectives when obtaining a probability-based sample versus a judgment sample. As can be seen in
Table 1, using judgment sampling, we trade the ability to quantify the precision of estimation and control the bias of selection of a defined population for learning about variation in the fragments of experience we are most interested in learning about—most often with an eye toward efficiency and getting “just enough” data to guide our learning and subsequent testing.

For example, in the first scenario in Table 1, the nurse manager on the unit may have chosen “the next 5 patients” because he wanted to know immediately whether the protocol was working, he knew staffing was stable over the next few days, and he was scheduled to work so could stay on top of the data collection. The case study below demonstrates the value of judgment sampling in health care and also serves to contrast the considerations of judgment sampling and probability sampling.

A CLINICAL EXAMPLE (FROM THE FIRST AUTHOR)

I was 36 years old when a small (<6 mm) mole on my leg was first noticed by my wife. It was so small and unalarming that I told my wife she was being paranoid and that there was no way this could be cause for concern. Nonetheless, I had an annual skin screening coming up with my nurse practitioner (NP) and I promised my wife I would point this mole out during the examination. When I met with my NP (one of the most thorough care providers I have ever had), she was not particularly concerned either. After all, it was less than 6 mm in diameter, not discolored, but did have some minor border irregularity. Just to be on the safe side, we decided to excise the mole and send it to the histology laboratory. I left the office and did not think about the situation further. A couple of weeks passed and I received a call from my NP stating that the mole was identified as melanoma in situ. I was stunned. My NP informed me that I would need to have a wide-margin excision to make sure the severely atypical cells were contained. Thankfully, results from the wide-margin excision were “clean” and I was given a clean bill of health, along with strict instructions to use sunscreen and to be extremely cautious about my degree of sun exposure. Most importantly, I was told to report to the office for a skin check every 6 months. Since then, I have had many samples of suspicious flesh sent to the laboratory, all coming back normal.

As I report to the dermatologist’s office every 6 months, I often wonder how many of the moles my care provider and I might incorrectly characterize as insignificant. I have complete trust in my care team’s judgment, but how can I be sure we are not missing any of the trouble moles?

Because this was a classic (and extremely important) sampling decision problem, I decided to list some of the sampling options my care team, or any care team, might have in this common clinical situation. I listed different sampling options, assuming a patient in my situation with no family history of melanoma, regardless of how practical they were. In this case, my body was the entire population from which a large number of biopsy samples could be taken over time. Three of the options are listed as follows:

Option 1: A clinician could decide to sample every mole identified on my skin. This “complete enumeration” would be costly, potentially painful, and disfiguring. It would also provide no guarantee that moles developing in the future would not be cancer-causing. Because this approach is costly and inefficient with no guarantees of future disease states, it is hardly ever exercised.

Option 2: Being trained as a scientist and recognizing the fact that clinical judgment is often inconsistent, a clinician could, in theory, identify and number each mole on my body. Following this procedure, a random number generator could be used to select, at random, a number of moles for excision (being sure to satisfy what ever degree of sampling error one might think acceptable). This of course would take the clinical expertise, training and experiential knowledge base of the clinician out of the process—an untenable position, for sure.

Option 3: A clinician could use established dermatologic criteria (asymmetry, border irregularity, color, and diameter—known as the ABCD rule) and
<table>
<thead>
<tr>
<th>Example</th>
<th>Situation (Context, Aim)</th>
<th>Method to Obtain Probability-Based Sample</th>
<th>Method to Obtain Judgment Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A care team on unit 3 was testing a new “quiet time” program during the third shift. They wanted to know whether this program was having a positive impact on minimizing sleep disturbance among patients.</td>
<td>After randomly selecting one of the next 3 patients, select every third patient admitted to Unit 3 to complete a brief sleep survey on discharge for 1 month (a systematic random sample)</td>
<td>Have the nurse manager give the sleep survey to the next 5 patients who have been in the unit for a third shift as they are discharged from unit 3 that experienced the new program.</td>
</tr>
<tr>
<td>2</td>
<td>A hospital was monitoring pneumonia vaccine core measure compliance on a unit with poor performance (unit B) after a process change.</td>
<td>Sample is determined by a third party data vendor using a simple random sample for all eligible patients producing an estimate with a 95% CL and 5% margin of error. (This sample may not equally represent all units of various size on any given period)</td>
<td>Sample 5 vaccine eligible patients each day on unit B for the next 3 weeks to assess whether the process change is working.</td>
</tr>
<tr>
<td>3</td>
<td>A thoracic surgery team is trying to increase the number of patients who go from “cancer diagnosis to first treatment” in 30 days or less using a patient navigator. Recent feedback from some patients is that they did not know the navigator program was even available.</td>
<td>Use a stratified random sample of 20 patients from the thoracic clinical database by gender and age groupings and review each case to see whether a patient navigator was assigned.</td>
<td>Monitor the next 2 patients who are supposed to be assigned a navigator.</td>
</tr>
<tr>
<td>4</td>
<td>Develop baseline information for an improvement initiative on safety in the ED</td>
<td>Select a sample of nurses to complete a safety survey. Obtain a list of all nurses in the ED. Assign a number to each nurse and then select the numbers at random (without replacement). Select the size of the sample with a response rate in mind to compute the margin of error per item in the survey.</td>
<td>Do 5 patient tracers in the ED with nursing staff on each shift to understand where safety concerns exist.</td>
</tr>
<tr>
<td>5</td>
<td>Nurse leadership wants to test the impact of a new pressure ulcer bundle on inpatient pressure ulcer prevalence, and data resources are very limited at the hospital.</td>
<td>Obtain a list of all the units in the hospital and randomly select 50% of the units. On a specific day prior to the intervention, nurse wound specialists will assess each patient for worsening pressure ulcers on these selected units. This will also be done 2 months after the intervention.</td>
<td>Sample 5 patients each week who are highest risk for pressure ulcers using a validated scale on the rehabilitation unit only (which has the highest historical rate of worsening pressure ulcers). Track results over time. (continues)</td>
</tr>
</tbody>
</table>
### TABLE 1.
DIFFERENCES BETWEEN JUDGMENT AND PROBABILITY SAMPLESa (Continued)

<table>
<thead>
<tr>
<th>Example</th>
<th>Situation (Context, Aim)</th>
<th>Method to Obtain Probability-Based Sample</th>
<th>Method to Obtain Judgment Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>An oncology manager wants to know whether patients get proper education from nurses when starting a new drug following the introduction of a new education flow sheet in the medical record.</td>
<td>Simple random sample used to select charts from all patients who started a new chemotherapeutic drug in the last 3 months. These charts will be reviewed manually by the nurse manager. (Some patients may not be in the practice anymore)</td>
<td>Select the most recent patients who have just had their drugs changed or modified for chart review.</td>
</tr>
</tbody>
</table>

This table contrasts probability and judgment (nonprobability) samples. However, there are a number of methods and tools to help improvers determine the minimum required sample size for different projects that go beyond this table. These include guidelines on how large a test should be performed, selecting a minimum subgroup size to create lower control limits on Shewhart charts, use of run chart simulations as a visual assessment of power, and more advanced techniques to gauge power to detect change under different scenarios such as operating characteristics curves. Although beyond the focus of this article, readers are referred to reference no. 10 for more information on these techniques.

DISCUSSION

If options 1 to 3 described earlier seem absurd, this is because they suggest discounting the expertise of the medical professionals. Nevertheless, the reality of medical practice often requires a balance between thoroughness and efficiency. The table above illustrates the differences between probability and judgment samples, highlighting the importance of considering both methods in quality improvement projects.

In practice, options 1 to 3 are rarely used, even for patients with a history of melanoma or melanoma in situ. Instead, clinicians rely on their years of clinical training and experience to select the moles they believe are suspicious. They realize they will not identify every problematic mole, but they have enough confidence in their professional judgment and information gathering techniques (e.g., good family histories) that they will not miss most problematic moles. Instead, they rely on their years of clinical training and experience to identify suspicious moles. They then use a combination of clinical judgment, randomization, and criteria-based weights to select the moles for further evaluation.

In conclusion, the use of probability samples in healthcare quality improvement is important, but judgment samples are also crucial in recognizing the potential for errors in clinical practice. The table above provides a useful framework for understanding the differences between these two approaches and how they can be applied in practice.

---

a This table contrasts probability and judgment (nonprobability) samples. However, there are a number of methods and tools to help improvers determine the minimum required sample size for different projects that go beyond this table. These include guidelines on how large a test should be performed, selecting a minimum subgroup size to create lower control limits on Shewhart charts, use of run chart simulations as a visual assessment of power, and more advanced techniques to gauge power to detect change under different scenarios such as operating characteristics curves. Although beyond the focus of this article, readers are referred to reference no. 10 for more information on these techniques.
Judgment Sampling

and perspective of the subject matter experts who are not merely experts in a chosen field (dermatology) but also practice this discipline in a busy and complex work environment, where the practical considerations of care and treatment are a daily reality. Although judgment sampling is not appropriate for every type of study, it is often the most appropriate form of sampling in an improvement context where our focus is on prediction into the future where the population of interest is unknown, unknowable, or beyond our resources to characterize.5

For example, a large dermatology clinic was interested to know whether the rate of positive biopsies for their general population differed between its specialist providers, which included 4 physicians, 3 NPs, and 1 physician assistant. Because the practice was using a hybrid medical record (part electronic and part paper) making historical laboratory data difficult to obtain and because one of the providers had only recently started with the group, a decision was made to study and plot the rate of positive biopsies by provider for the next 2 months (June and July 2011) (Figure 1).

The group thought this was an acceptable approach to take initially because billing data suggested they all performed a relatively large number of biopsies during a given month and the patient mix across the practice was similar. The data revealed consistency among the 4 physicians, a higher percentage of positive results for the NPs (especially NP 1) and a low percentage for the physician assistant.

In a noncompetitive and patient-centered manner, the group wanted to learn why the results might be different across provider types and had information to guide their learning and further studies. They decided to extend the data collection for another 2-month period so they could have a higher degree of belief in the results and also to gauge the potential impact of any changes they might make on the basis of what they learned from each other. By monitoring performance over time versus aggregating large amounts of data in an annual or biannual summary statistic, the practice gained the advantage of knowing when detection rates began to shift by provider and provider type over time.

This example highlights how limitations in historical data need not get in the way of a group that wants to learn about practice variation quickly with data that are available and are a logical starting point for improvement. Subsequent studies could use additional judgment samples to focus on a particular subset of patients (eg, those with a history of melanoma, family history of melanoma, adults older than 65 years). As the signs and symptoms of disease states change, real time tracking in clinical practice becomes more important.

**SUMMARY**

Judgment sampling relies upon those with process and subject matter knowledge to select useful samples for learning about process performance and the impact of changes over time. In many cases, where the goal is to learn about or improve a specific process or system, judgment samples are not merely the most convenient and economical approach, they are technically and conceptually the most appropriate approach. This is because improvement work is often done in the real world in complex situations involving specific areas of concern and focus; in these situations the assumptions of classical sampling theory cannot be met nor would it even be desirable to try to meet those assumptions. Although not all types of studies will benefit from judgment sampling, far more opportunities exist to use judgment samples.
than people realize. In many cases, like the clinical example addressed here, they are already routinely being used.

REFERENCES