### Revisiting the Role of Measurement

Frank Federico & Robert Lloyd

---

#### The Three Faces of Performance Measurement

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Improvement</th>
<th>Accountability</th>
<th>Research</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aim</strong></td>
<td>Improvement of care (efficiency &amp; effectiveness)</td>
<td>Comparison, choice, reassurance, motivation for change</td>
<td>New knowledge (efficacy)</td>
</tr>
<tr>
<td><strong>Methods:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&amp; Test Observability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Test Observability</td>
<td>Test observable</td>
<td>No test, evaluate current performance</td>
<td>Test blinded or controlled</td>
</tr>
<tr>
<td>• Bias</td>
<td>Accept consistent bias</td>
<td>Measure and adjust to reduce bias</td>
<td>Design to eliminate bias</td>
</tr>
<tr>
<td>• Sample Size</td>
<td>“Just enough” data, small sequential samples</td>
<td>Obtain 100% of available, relevant data</td>
<td>“Just in case” data</td>
</tr>
<tr>
<td>• Flexibility of Hypothesis</td>
<td>Flexible hypotheses, changes as learning takes place</td>
<td>No hypothesis</td>
<td>Fixed hypothesis (null hypothesis)</td>
</tr>
<tr>
<td>• Testing Strategy</td>
<td>Sequential tests</td>
<td>No tests</td>
<td>One large test</td>
</tr>
<tr>
<td>• Determining if a change is an improvement</td>
<td>Analytic Statistics (statistical process control) Run &amp; Control charts</td>
<td>No change focus (maybe compute a percent change or rank order the results)</td>
<td>Enumerative Statistics (t-test, F-test, chi square, p-values)</td>
</tr>
<tr>
<td>• Confidentiality of the data</td>
<td>Data used only by those involved with improvement</td>
<td>Data available for public consumption and review</td>
<td>Research subjects’ identities protected</td>
</tr>
</tbody>
</table>

---

Where are you working?

- **Discovery**
  - Aim: Discover changes that might improve practice in your setting

- **Pilot**
  - Aim: demonstrate that the changes that worked elsewhere can work here

- **Diffusion**
  - Aim: spread prototype changes everywhere applicable

- **Implementation**
  - Changes are the way we work

What measures do you need?

- **During testing**
  - Initially measures based on questions/theory

- **As expand/diffuse**
  - Is the process being used as designed?
Measurement

- Small samples over time should be used to determine if the process is improving.
- Data should be collected by the team with strict attention to the agreed upon tempo.
- Data should be collected for segments (stratification).
- Process measurements should be the primary team measures.
- Outcome measures are needed but do not need to be collected by the team.

What will we look for?

- Initially process measures
  - Is the process being used?
- Over time: outcome measures
  - Have we made an impact on outcomes?
  - You will always need to report outcome measures.
Science and Outcomes

- Process reliability is linked to outcomes by science
- If the process is “reliable” and the outcome is not achieved either the science is wrong or the process really is not being done correctly
- Outcomes are linked to the processes by the confirmation the hypothesis

The Three Faces of Performance Measurement

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Improvement</th>
<th>Accountability</th>
<th>Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aim</td>
<td>Improvement of care (efficiency &amp; effectiveness)</td>
<td>Comparison, choice, reassurance, motivation for change</td>
<td>New knowledge (efficacy)</td>
</tr>
<tr>
<td>Methods:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Test Observability</td>
<td>Test observable</td>
<td>No test, evaluate current performance</td>
<td>Test blinded or controlled</td>
</tr>
<tr>
<td>• Bias</td>
<td>Accept consistent bias</td>
<td>Measure and adjust to reduce bias</td>
<td>Design to eliminate bias</td>
</tr>
<tr>
<td>• Sample Size</td>
<td>“Just enough” data, small sequential samples</td>
<td>Obtain 100% of available, relevant data</td>
<td>“Just in case” data</td>
</tr>
<tr>
<td>• Flexibility of Hypothesis</td>
<td>Flexible hypotheses, changes as learning takes place</td>
<td>No hypothesis</td>
<td>Fixed hypothesis (null hypothesis)</td>
</tr>
<tr>
<td>• Testing Strategy</td>
<td>Sequential tests</td>
<td>No tests</td>
<td>One large test</td>
</tr>
<tr>
<td>• Determining if a change is an improvement</td>
<td>Analytic Statistics (statistical process control) Run &amp; Control charts</td>
<td>No change focus (maybe compute a percent change or rank order the results)</td>
<td>Enumerative Statistics (t-test, F-test, chi square, p-values)</td>
</tr>
<tr>
<td>• Confidentiality of the data</td>
<td>Data used only by those involved with improvement</td>
<td>Data available for public consumption and review</td>
<td>Research subjects’ identities protected</td>
</tr>
</tbody>
</table>

Solberg, L I; Mosser, G; McDonald, S “The three faces of performance measurement: improvement, accountability, and research.” The Joint Commission journal on quality improvement 23, No. 3 1997, pp. 135-47.
The Quality Measurement Journey

AIM (How good? By when?)
Concept
Measure
Operational Definitions
Data Collection Plan
Data Collection
Analysis

Discussion Topics

- Why should I stratify our data?
- To sample or not to sample... that is the question!
- How long must this measurement madness go on?

Stratification (or Segmentation)

- Separation & classification of data according to predetermined categories
- Designed to discover patterns in the data
- For example, are there differences by shift, time of day, day of week, severity of patients, age, gender or type of procedure?
- Consider stratification **BEFORE** you collect the data
Approaches to Sampling

Research Approach: a Pond

- Pull one sample from this spot and walk away!

QI Approach: a River

- But, how do you pull a sample from a moving process?

Fixed population-universe, frame
Random sampling
Probability based
Purpose:
- determine how much variation in a sample
- apply learning to the sample (should not extrapolate)
- reject or do not reject sampled population
Hypothesis, statistical tests (t-test, F-test, chi square, p-values)

No fixed population
Population-ongoing "stream" of data
Also uses judgment sampling
Not totally based on probability
Purpose:
- take action on underlying process to improve future outcome of process
- Run charts or Shewhart control charts

Sampling Methods

Probability Sampling Methods

- Simple random sampling
- Stratified random sampling
- Stratified proportional random sampling
- Systematic sampling
- Cluster sampling

Non-probability Sampling Methods

- Convenience sampling
- Quota sampling
- Judgment sampling

Sampling Options

Research Approaches

Simple Random Sampling

Stratified proportional Random Sampling

Judgment Sampling

Improvement Approach

Judgment Sampling

Especially useful for PDSA testing. Someone with process knowledge selects items to be sampled.

Characteristics of a Judgment Sample:

- Include a wide range of conditions
- Selection criteria may change as understanding increases
- Successive small samples instead of one large sample
Judgment Sampling takes advantage of the knowledge of those who own the process.

Are things pretty much the same in this ward day in and day out?

Things are pretty quiet after 3 PM.

We are absolutely crazy around here between 9 and 11 AM!

But I don’t work afternoon shift so I really don’t know what goes on then.

Not really. Let’s see...

Sample Size and the Ability to Detect Change

The 5 run charts in this figure show how sample size can affect the decisions you make.

A 30% reduction in patient wait time at the clinic was observed after week 12.

The average wait time in each chart is about 40 minutes.

Which of the 5 sample sizes provides the minimum amount of evidence that the change has occurred?

Sample size issues for QI are a balance between resources (time, money, energy) and the clarity of the results desired.

How long must this measurement madness go on?

It is not uncommon for a team to want to stop collecting data, especially after they have been at it for a year or two or improvement is noted!

The reliability of the process, your need to know, the criticality of the measures and the amount of data required to make conclusions should all drive your decisions about the frequency and duration of data collection.
Let’s clarify two key terms…

**Frequency** – the period of time in which you collect data (i.e., how often will you dip into the process to analyze the variation that exists?)
- Moment by moment (continuous monitoring)?
- Every hour? Every day?
- Once a week? Once a month?
- Once a quarter? Once a year?

**Duration** – how long you need to continue collecting data
- Do you collect data on an on-going basis and not end until the measure is always at the specified target or goal?
- Do you conduct periodic audits?
- Do you just collect data at a single point in time to “check the pulse of the process”?

Three aspects of a measure

Frequent data tracking (i.e., daily, weekly or monthly) should occur until the measure is stable, capable and sustainable.

- **Stable** means that the measure does not swing widely (i.e., +/- 10% variation in the numbers) and reflects non-random or common cause variation.
- **Capable** means that the measure is meeting the defined target or goal (e.g., 75% reduction in pressure ulcers as an outcome measure or 95% of all patients get their skin assessed at a specified time interval as a process measure).
- **Sustainable** means that the measure is reliable over time (i.e., not just hitting the target once but hitting it 5 or more times in a row with only +/- 10% random variation around the new level of performance).
A Simple Rule for Outcome Measures

**Outcome Measures – Always!**

As long as you are concerned about the quality and safety of the care that you deliver, you should always be tracking the outcomes!

**How long should these outcomes be measured?**

- When do you stop measuring your financial results?
- When should a diabetic patient stop tracking his or her blood glucose?
- How long should we monitor the vital signs of an ICU patient?
- When should airport security stop assessing passengers for weapons?
- How long does a local water authority need to measure the quality of the water going through its pipes?
- When should schools stop measuring the progress of students?

A Simple Rule for Process Measures

**Process Measures – it depends!**

- Process measures usually demonstrate improvement before outcome measures.
- Process measures may be revised during an improvement project which means that new data will then need to be collected.
- A process measure should demonstrate improvement (against the RCRs) and then STAY at or around the new level of performance for at least 5 reporting periods to be considered “sustained.”
Frequency of Process Measures

• **Regularly (daily, weekly or monthly)**
  Done initially to improve a specific measure and understand the variation inherent in this measure. Continued until the measure is stable and the variation understood.

• **Periodically (once every 2 - 3 months)**
  Done when statistical improvement has been noted using run chart rules, sustained AND the process is highly reliable, then a periodic audit approach to data collection and measurement can be used.

• **Once or twice a year (why bother?)**

• **Stop measuring?**
  Done only when the measure is no longer relevant.

---

Summary Guidelines for Measurement

• Initially when starting out, daily, weekly or monthly data collection is required if you really want to understand the variation in the measure and establish a baseline.

• Run chart rules (RCRs) should be used to then determine if the interventions (PDSA tests) have been able to move the measure in the desired direction. This will most often be detected by the RCR for a shift (6 or more consecutive data points above or below the median) or a trend (5 or more consecutive data points increasing or decreasing) in the data against the baseline.

• When a measure has been determined to be stable and capable as well as being sustained over time at the new level of desired performance then it is time to consider backing off the frequency of data collection.
Exercise: Data Collection Strategies
(frequency, duration and sampling)

• This exercise has been designed to test your knowledge of and skill with developing a data collection plan.
• In the table on the next page is a list of eight measures.
• For each measure identify:
  – The frequency and duration of data collection.
  – Whether you would pull a sample or collect all the data on each measure.
  – If you would pull a sample of data, indicate what specific type of sample you would pull.
• Spend a few minutes working on your own then compare your ideas with others at your table.

Exercise: Data Collection Strategies
(frequency, duration and sampling)

The need to know, the criticality of the measure and the amount of data required to make a conclusion should drive the frequency, duration and whether you need to sample decisions.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Frequency and Duration</th>
<th>Would you sample? If yes, what type?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vital signs for a patient connected to full telemetry in the cardiac ICU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blood pressure (systolic and diastolic) to determine if the newly prescribed medication and dosage are having the desired impact</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent compliance with a hand hygiene protocol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cholesterol levels (LDL, HDL, triglycerides) in a patient recently placed on new statin medication</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent of patients receiving daily pressure ulcer risk assessments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central line bloodstream infection rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent of inpatients receiving multi-disciplinary rounds</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent of surgical patients given prophylactic antibiotics within 1 hour prior to surgical incision</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>