Promises and Pitfalls of Consumer Use of Publicly Available Quality Information: Improving the Measures

Patrick S. Romano, MD MPH
Professor of Medicine and Pediatrics
University of California, Davis
June 25, 2012
What are the problems?

- Obvious to consumers and families
  - Too many measures; limited opportunities for personalization
  - Too few interesting measures that capture what’s important to patients

- Important but not obvious to consumers and families
  - Reliability of provider classification is often poor
  - Validity of measures varies and is often uncertain
National Quality Measures Clearinghouse (N=2075)

Measure Index

The Measure Index is a complete list of summaries published on the NQMC Web site. The list is organized alphabetically by submitting organization.

NQMC currently contains 2075 individual measure summaries.

Accreditation Association for Ambulatory Health Care Institute for Quality Improvement (1)
- Intra-procedure colonoscopy complication rate: percentage of patients who developed one or more intra-procedure complications.

Accreditation Association for Ambulatory Health Care Institute for Quality Improvement, Performance Measurement Initiative (1)
- Intra-procedure colonoscopy complication rate: percentage of patients who developed one or more intra-procedure complications.

Agency for Healthcare Research and Quality (AHRQ) (161)
- Abdominal aortic aneurysm (AAA) repair: mortality rate.
- Abdominal aortic aneurysm (AAA) repair: volume.
- Accidental puncture or laceration (area-level): rate per 100,000 population.
- Accidental puncture or laceration (provider-level): rate per 1,000 discharges.
- Accidental puncture or laceration: rate per 1,000 eligible discharges.
- Acute myocardial infarction (AMI): mortality rate, without transfer cases.
National Quality Forum Evaluation Criteria

- Importance to measure and report
  - What is the level of evidence for the measures?
  - Is there an opportunity for improvement?
  - Relation to a priority area or high impact area of care?

- Scientific acceptability of the measurement properties
  - What is the reliability and validity of the measure?

- Usability
  - Can the intended audiences understand and use the results for decision-making?

- Feasibility
  - Can the measure be implemented without undue burden, captured with electronic data/EHRs?
## Leading stewards of NQF-endorsed measures (web search 6/24/2012)

<table>
<thead>
<tr>
<th>Organization</th>
<th>Number (N=719)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center for Medicare &amp; Medicaid Services</td>
<td>129</td>
</tr>
<tr>
<td>National Committee for Quality Assurance</td>
<td>100</td>
</tr>
<tr>
<td>American Medical Association-PCPI</td>
<td>98</td>
</tr>
<tr>
<td>AHRQ</td>
<td>51</td>
</tr>
<tr>
<td>The Joint Commission</td>
<td>32</td>
</tr>
<tr>
<td>Society for Thoracic Surgeons</td>
<td>35</td>
</tr>
<tr>
<td>ActiveHealth Management, Inc.</td>
<td>24</td>
</tr>
<tr>
<td>Resolution Health, Inc.</td>
<td>24</td>
</tr>
<tr>
<td>Centers for Disease Control and Prevention</td>
<td>12</td>
</tr>
</tbody>
</table>
NQF portfolio of measures

- 719 cross-cutting and condition-specific measures
- 30% outcome measures
Too few interesting measures
Outcomes should reflect goals of treatment

- Mortality (inpatient, 30-day, 180-day)
- Morbidity (complications, adverse events)
- Functional status (return to school, work, usual/desired activities)
- Quality of life (freedom from pain or other distressing symptoms)
Healthcare Information Division

Coronary Artery Bypass Graft (CABG) Surgery in California

CABG surgery is the most common surgical procedure for treating coronary artery disease. In this surgery, a vein or artery from another part of the body is used to create a new path for blood to flow to the heart, bypassing the blocked artery. Coronary artery disease is the leading cause of all adult non-maternal admissions, representing nearly 9% of all admissions. It is a chronic condition in which cholesterol and fat solidify to form plaque along the linings of the coronary arteries. As the plaque continues to build up, blood vessels can be restricted or blocked leading to chest pain or a heart attack.


Go to CABG Trends for: 2003-2008

Go to Other CABG Reports: impact of Public Reporting | The State of Cardiac Revascularization Outcomes Reporting
OSHPD’s CABG Report

Figure 3: O/E Ratios Over Time for 67 CCMRP Participating Hospitals that Have at Least Two Years of Continuous O/E Ratios Available Between 2000 and 2002 (Continued)
Redding Medical Center, Tenet, and “medicine gone awry”
Reliability and Validity

- **Reliability of provider classification (random error in classification or estimation)**
  - Stabilize estimates (shrinkage, smoothing)
  - Change tiering methods
  - Create composites

- **Validity of provider classification (systematic error in classification or estimation)**
  - Selection bias
  - Information bias
  - Confounding bias
Hierarchical Models

- Also referred to as smoothed rates or reliability-adjusted rates
- Endorsed by NQF for outcome measures
- Methods to separate the within and between provider level variation (random vs. systematic)
- Total variation = Within provider (noise) + Between provider variation (signal)
- Reliability \( w \) = Between / Total
  - Signal ratio = signal / (signal + noise)

Stabilization of estimates

- Smoothed rate is the (theoretical) best predictor of future quality because it borrows strength from other relevant information about provider performance.

- Smoothed rate (single provider, single indicator) =
  \[ \text{Hospital-type rate} \times (1 - w) + \]
  \[ \text{Hospital-specific rate} \times w \]

- Multivariate versions
  - Similar providers
  - Other years (auto-regression, forecasting)
  - Other measures (composites)
  - Contemporaneous innovations or shocks
Misclassification of Performance

- Misclassification is related to:*
  - The reliability of a measure
    - Which depends on sample size (which can vary provider to provider)
    - Variation between providers
  - Number of cutpoints in the classification scheme
  - How close the performance score is to the cutpoint

Traditional tiering methods

- Confidence intervals or p values
  - Each provider’s observed (or predicted) performance is compared with its expected performance
  - A low p-value rejects the null hypothesis
  - This approach tells us nothing about the relative performance of any two or more providers

- Ranking (tiers of size k=1)
Options in RAND White Paper

1. **Exclude providers** for whom the risk of misclassification due to chance is too high

2. **Exclude measures** for which the risk of misclassification due to chance is too high for too many providers

3. **Modify the classification system** used in the performance report
   - Report using fewer categories
   - Change the thresholds for deciding categories
   - Introduce a zone of uncertainty around cutpoints
   - Report shrunken estimates instead of categories
   - Newer option: report threshold exceedance probabilities (probability that true value exceeds $x$)

Friedberg and Damberg, Methodological Considerations in Generating Provider Performance Scores for Use in Public Reporting. AHRQ 2012.
Between-Provider Performance Variation

Lower between-provider variation
(harder to tell who is best)

Higher between-provider variation
(easier to tell who is best)

= average performance for each provider
Different Levels of Measurement Error (Uncertainty about the “true” average performance)

Higher measurement error (harder to tell who is best)

Lower measurement error (easier to tell who is best)

● = average performance for each provider

<> = range of uncertainty about “true” average performance
Various factors contribute to misclassification risk

- Higher misclassification risk
  - Lower reliability
  - Classification system: More categories
  - Lower within-provider measurement error
  - Lower between-provider variation in performance
  - Higher average error per observation
  - Lower number of observations
Why composite measures?  
(aka summary measures, roll-up measures)

- AHRQ: “condensing multiple quality measures into a single piece of information”:
  - Reduces cognitive burden for consumers, providing clearer “signal” and reducing the danger of “cognitive shortcuts”
  - Enhances reliability or ability to discriminate between higher-quality and lower-quality providers

- But remember two potential concerns:
  - Difficulty achieving consensus on composite construction and scoring, perhaps due to lack of professional consensus.
  - Loss of important information if the composite combines unrelated metrics in a manner that washes out meaningful differences on individual indicators.
Two conceptual approaches

- **Psychometric or reflective perspective** - an underlying, unmeasured factor ("quality") is the *cause* of what we observe; the observed data *reflect* this unmeasured factor
  - Requires a correlation *among* the measures included in the composite, because different measures can only reflect quality if they are correlated with each other.

- **Clinometric or formative perspective** – focus on guiding decision-making to optimize welfare instead of measuring an unobserved, latent factor
  - Use clinical judgment rather than empirical analysis to select component measures
  - *Formed from* or defined by specific indicators, so no correlation among component measures is required
## Approaches to scoring composite measures

<table>
<thead>
<tr>
<th>Scoring Method</th>
<th>Definition</th>
<th>Example</th>
<th>Adopter</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-or-none</td>
<td>Percentage of patients for whom all indicators triggered by that patient are met.</td>
<td>“Appropriate Care Measure” for 4 conditions (heart attack, heart failure, pneumonia, and SCIP).</td>
<td>PHCQA Progress and Performance Report of Hospital Quality</td>
</tr>
<tr>
<td>Overall Percentage (Opportunity weighting)</td>
<td>Percentage of care events that were properly delivered, where each opportunity to do the right thing counts equally.</td>
<td>149 hypertensive patients triggered 26 hypertension indicators 828 times. Required care was given 576 times yielding 69.9% (576/828).</td>
<td>CMS P4P Premier Hospital Quality Incentive Demonstration</td>
</tr>
<tr>
<td>Indicator Average (Equal event weighting)</td>
<td>Scores are averaged across all indicators to represent the mean adherence rate.</td>
<td>Hospital quality of care for acute myocardial infarction, congestive heart failure and pneumonia.</td>
<td>Hospital Quality Alliance (HQA)</td>
</tr>
<tr>
<td>Patient Average (Equal patient weighting)</td>
<td>The percentage of indicators successfully met is computed for each patient, and then averaged at the patient level.</td>
<td>None to our knowledge</td>
<td></td>
</tr>
<tr>
<td>Expert Opinion (Evidence-based)</td>
<td>Indicators are weighted based on evidence of impact on population health and/or effort required to achieve.</td>
<td>General Medical Services contract pays physicians more for achieving performance targets that require more time and other resources.</td>
<td>UK National Health Service</td>
</tr>
</tbody>
</table>
Different scoring methods can generate very different rankings.

Combining Multiple Indicators of Clinical Quality: An Evaluation of Different Analytic Approaches.
Reeves, David; Campbell, Stephen; Adams, John; Shekelle, Paul; MD, PhD; Kontopantelis, Evan; Roland, Martin
Diabetes composite reliability for PCPs in CA IHA

- Composite: 74%
- EYE: 1%
- HBA: 27%
- LDL: 27%
- MPM: 43%
All-or-none scoring may be driven by a single indicator.
All-or-none scoring ignores diminishing returns and public health impact

Figure 1. Stylized relative risks for development of various complications as a function of mean HbA1c during follow-up in the DCCT. For the purposes of illustration, the relative risk of various complications is set to 1 at HbA1c of 6%. The lines depict a stylized relationship for risk of: sustained progression of retinopathy (----- ); progression to clinical nephropathy (urinary albumin excretion ≥300 mg/24 h) (----- ); progression to severe nonproliferative or proliferative retinopathy (-- - - ); progression to clinical neuropathy (-- - - ); and progression to microalbuminuria (urinary albumin excretion ≥ 40 mg/24 h) (-- - - - - - ).
Optimal Weighting

- Weight by the measure’s *impact*
  - *Impact* reflects public health importance or opportunity for improvement

- Weight by the measure’s *reliability*
  - *Reliability* reflects sponsor’s level of confidence in the estimate for a given provider

- Weight by the measure’s *validity?*
  - A *distorted* performance measure is one that results in actions by the provider that are not perfectly aligned with the sponsor’s objective
AHRQ PSI Composite approach

- Impact is measured by the number of adverse events for each measure, potentially adjusted by excess charges or LOS.
- Reliability = signal / (signal + noise)
- Validity = false positive rate and/or false negative rate
Validity issues

- Selection bias
  - Which providers submit data for voluntary programs?
  - Which patients are omitted due to missing data?

- Information (ascertainment) bias
  - False positive, false negative errors

- Confounding bias
  - Factors other than quality (e.g., unmeasured case mix) that actually explain variation in measure across providers
Participation varied from 8% (ID) to 100% (AL, CA, ME, NH, NM, VT, VA, WV, WY) across states.
Some CAHs submit data only to QIOs (circa 22%), not to Hospital Compare, and have poorer performance on average.
Treat Missing Values as "N/A"

Using AHA Data and Treating Missing Values as "N/A"
### Information bias related to PSIs

<table>
<thead>
<tr>
<th>Name</th>
<th>VA</th>
<th>AHRQ</th>
<th>UHC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPV (%) (95% CI)</td>
<td>Sample (n)*</td>
<td>PPV (%) (95% CI)</td>
<td>Sample (n)</td>
</tr>
<tr>
<td>Decubitus Ulcer</td>
<td>30 (22-40)</td>
<td>112</td>
<td>--</td>
</tr>
<tr>
<td>Foreign Body Left in During Procedure</td>
<td>46 (36-55)</td>
<td>93</td>
<td>--</td>
</tr>
<tr>
<td>Iatrogenic Pneumothorax</td>
<td>73 (64-81)</td>
<td>112</td>
<td>78 (73-82)</td>
</tr>
<tr>
<td>Central Venous Catheter-related Bloodstream Infections</td>
<td>38 (29-47)</td>
<td>112</td>
<td>61 (51-71)</td>
</tr>
<tr>
<td>Postoperative Hip Fracture</td>
<td>28 (15-43)</td>
<td>46</td>
<td>--</td>
</tr>
<tr>
<td>Postoperative Hemorrhage or Hematoma</td>
<td>75 (66-83)</td>
<td>112</td>
<td>--</td>
</tr>
<tr>
<td>Postoperative Physiologic and Metabolic Derangements</td>
<td>63 (54-72)</td>
<td>119*</td>
<td>--</td>
</tr>
<tr>
<td>Postoperative Respiratory Failure</td>
<td>67 (57-76)</td>
<td>112</td>
<td>--</td>
</tr>
<tr>
<td>Postoperative PE or DVT</td>
<td>43 (34-53)</td>
<td>112</td>
<td>47 (42-52)</td>
</tr>
<tr>
<td>Postoperative Sepsis</td>
<td>53 (42-64)</td>
<td>112</td>
<td>41 (28-54)</td>
</tr>
<tr>
<td>Postoperative Wound Dehiscence</td>
<td>87 (79-92)</td>
<td>112</td>
<td>--</td>
</tr>
<tr>
<td>Accidental Puncture or Laceration</td>
<td>85 (77-91)</td>
<td>112</td>
<td>91 (86-94)</td>
</tr>
</tbody>
</table>
453.4  Acute venous embolism and thrombosis of deep vessels of lower extremity
453.40  Acute venous embolism and thrombosis of unspecified deep vessels of lower extremity
  Deep vein thrombosis NOS
453.41  Acute venous embolism and thrombosis of deep vessels of proximal lower extremity
  Femoral, Iliac, Popliteal, Thigh, Upper leg NOS
453.42  Acute venous embolism and thrombosis of deep vessels of distal lower extremity
  Calf, Lower leg NOS, Peroneal, Tibial
453.5  Chronic venous embolism and thrombosis of deep vessels of lower extremity
  Excludes: personal history of venous thrombosis and embolism (V12.51)
453.50  Chronic venous embolism and thrombosis of unspecified deep vessels of lower extremity
453.51  Chronic venous embolism and thrombosis of deep vessels of proximal lower extremity
453.52  Chronic venous embolism and thrombosis of deep vessels of distal lower extremity
453.6  Venous embolism and thrombosis of superficial vessels of lower extremity
453.7  Chronic venous embolism and thrombosis of other specified vessels
  Excludes: personal history of venous thrombosis and embolism (V12.51)
453.71  Chronic venous embolism and thrombosis of superficial veins of upper extremity
453.72  Chronic venous embolism and thrombosis of deep veins of upper extremity
453.73  Chronic venous embolism and thrombosis of upper extremity, unspecified
453.74  Chronic venous embolism and thrombosis of axillary veins
453.75  Chronic venous embolism and thrombosis of subclavian veins
453.76  Chronic venous embolism and thrombosis of internal jugular veins
453.77  Chronic venous embolism and thrombosis of other thoracic veins
453.79  Chronic venous embolism and thrombosis of other specified veins
453.8  Acute venous embolism and thrombosis of other specified veins
  Excludes: cerebral, coronary, intracranial sinus, nonpyogenic, mesenteric, portal, precerebral, pulmonary
453.81  Acute venous embolism and thrombosis of superficial veins of upper extremity
453.82  Acute venous embolism and thrombosis of deep veins of upper extremity
453.83  Acute venous embolism and thrombosis of upper extremity, unspecified
453.84  Acute venous embolism and thrombosis of axillary veins
453.85  Acute venous embolism and thrombosis of subclavian veins
453.86  Acute venous embolism and thrombosis of internal jugular veins
453.87  Acute venous embolism and thrombosis of other thoracic veins
453.89  Acute venous embolism and thrombosis of other specified veins
453.9  Of unspecified site (embolism of vein, thrombosis (vein))
Methods

- Two parallel studies were conducted to update previous PPV estimates for PSI 12 and to identify actionable opportunities to improve care:
  - 7 volunteer hospitals recruited through AHRQ QI listserve, including flagged cases only
  - 15 academic health systems recruited through UHC, including both flagged and unflagged cases with TKA surgery

- AHRQ PSI 12 Version 4.1 software was applied to eligible cases from participating hospitals, using “present on admission” (POA) flags.
  - Hospital’s own data (AHRQ) or Clinical Database (UHC)

- Flagged cases were reviewed by trained QI nurses at each hospital, using detailed chart abstraction tool and guidelines, with detailed review of discrepant cases.
Summary of findings from volunteer community hospitals

- Records from volunteer hospitals in AHRQ study were sampled in sequential reverse order from 6/30/2010 back to 10/1/2009, up to N=30
- PPV much better than in previous studies of PSI 12 (81% versus 43-47%)
- Of 30 false positive cases:
  - 15 cases were POA
  - 8 cases were upper extremity VT
  - 1 case was SVC (central VT)
  - 3 cases were superficial VT
  - 3 cases were chronic
Postoperative DVT/PE after TKA
Follow-up study of PPV in 15 academic centers

126 VTE flagged by PSI 12 (+4 Readmission)

- Positive Predictive Value
  \[
  \frac{TP}{TP + FP} = \frac{125}{125 + 1} = 0.992
  \]

125 cases True Positive postop lower ext DVT or PE
1 case clinical False Positive (superficial) saphenous Vein
Postoperative DVT/PE after TKA
Follow-up study of NPV in 15 academic centers

- 463 Not flagged as VTE by PSI 12

- 5 cases had VTE per UHC abstract
- 458 cases had no VTE (TN)

- 3 cases False Negative
- 2 cases superficial or upper extremity thromboses

Negative Predictive Value
\[ \text{NPV} = \frac{\text{TN}}{\text{FN} + \text{TN}} \]
\[ \text{NPV} = \frac{458}{458+3} = 0.993 \]

- Previous sensitivity estimate from 33 teaching hospitals:
  96% (95% CI: 86-100%)
  100% if limited to acute DVT or PE
Conclusions

Science of measure development is improving, with particular attention to:

- New domains of outcome measurement
- Improving reliability through stabilization, attention to tiering, and composites
- Attention to validity by improving code sets and registry element definitions