May 13, 2020

Measure Testing Webinar: Part 3
Measure Testing in eCQMs – An Overview
• Welcome and Housekeeping
• Introductions
• Presentation
• Discussion
Housekeeping

• This forum is being recorded
• All lines will be muted
Heather Tinsley, MSPH  
Project Manager, Measure Testing & Analytics

Deborah Harper, MA  
Program Manager, Measure Testing & Analytics
Overview

• Goals of Measure Testing
• Evolution of Measure Testing
• Testing with Electronic Data Sources
• Testing Methodologies
  – Review of commonly used methodologies
Introduction

• Goals of measure testing
  – Demonstrate feasibility and scientific acceptability
  – Increase provider confidence in measures as an indicator of quality
  – Inclusion in national quality reporting programs
  – Demonstrate improved health care outcomes through correlations
Evolution of Measure Testing

Paper Records & Administrative Claims

Electronic Data Sources

“Test early, test often.”
WHAT LIES BENEATH?
A COMPARISON OF CLAIMS DATA AND EHR DATA AVAILABLE FOR 500 PATIENTS

CLAIMS DATA
The foundation for most healthcare analytics, Claims Data is easy to come by, but delayed, and short on details.

EHR DATA
Real-time, rich in clinical content, and right under your fingertips. EHR data enhances risk algorithms and informs outcomes-based measures.

© 2015 Arcadia Solutions

© 2020 PCPI Foundation. All rights reserved.

Changes in the Testing Landscape

• In September 2019 NQF release updated testing guidance in their 2019 NQF Endorsement Criteria
  – New changes:
    • Increased emphasis on disparities data for all measures to show importance and performance gap
    • All eCQMs, new and existing, must be tested for reliability and validity at the data element level.
    • Bonnie testing alone no longer sufficient for validity/reliability

• In November 2019 CMS released the 2020 Quality Payment Program Final Rule
  – New changes:
    • All QCDR measures submitted for the *2021 performance period MUST be fully tested as defined in the MMS Blueprint/NQF Evaluation Criteria
  – Previously, QCDR measures were not required to submit testing results

*This guidance has since been extended
Changes in the Testing Landscape

• In April 2020 CMS released additional guidance in response to COVID-19
  – For candidate MIPS measures (e.g., eCQMs):
    • Measure submission deadline extended to June 30, 2020
    • Extension of submission for new or updated testing-related supporting data to September 4, 2020 *IF* you submit your measure by June 30, 2020 *AND* send an extension request to preruemaking@battelle.org
  – For QCDR measures
    • Delay of implementation of the QCDR measure testing and data collection policies by 1 year, from the 2021 performance period to the 2022 performance period
<table>
<thead>
<tr>
<th></th>
<th>Feasibility</th>
<th>Reliability</th>
<th>Validity</th>
<th>Bonnie Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>eCQM</td>
<td>New Measure</td>
<td>NQF Scorecard</td>
<td>Data Element*</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Existing Measure</td>
<td>NQF Scorecard</td>
<td>Data Element*</td>
<td>Yes</td>
</tr>
<tr>
<td>Registry</td>
<td>New Measure</td>
<td>No formal tool required. Attest to measure’s feasibility by answering 3 Q’s</td>
<td>Measure Score</td>
<td>Face</td>
</tr>
<tr>
<td></td>
<td>Existing Measure</td>
<td>No formal tool required. Attest to measure’s feasibility by answering 3 Q’s</td>
<td>Measure Score</td>
<td>Empirical, or Face with justification</td>
</tr>
</tbody>
</table>

* No reliability testing required if data element validity testing conducted and results are adequate; however, additional score level testing is encouraged.
Feasibility Testing

• Gather qualitative information that may contribute to further refinement of measure concepts and technical specifications
• Identify potential challenges to implementation
• Formulate strategy for data collection and analysis
• Develop plans for integration
• Anticipate unintended consequences
Feasibility Methodologies

- NQF Feasibility Scorecard
- Simulated Patient Deck
- Provider Interviews
- Patient Interviews
- Focus Groups
An excel worksheet intended to rate data elements within a measure on a 0 to 1 scale related to:

- Data availability: the extent to which the data are readily available in a structured format across EHR systems
- Data accuracy: the extent to which the information contained in the data is correct
- Data standards: the extent to which the data element is coded using a nationally accepted terminology standard (vocabulary) and mapped to the Quality Data Model (QDM)
- Workflow: the extent to which capturing the data element impacts the typical workflow for that user
• A minimum of two test sites using different EHR systems is required

• PCPI adds a comment sheet to the excel file to gather additional qualitative information as it pertains to each data element

• For each data element that scores a 0 among the four categories you must detail a feasibility plan
Test Sites

• Test sites play a crucial role in measure development and testing

• They provide data, time, knowledge, expertise, experience, and opportunities for future collaboration

• Choosing the right test site depends:
  – Type of measure
  – Clinical focus area

• Start recruitment early
  – Can take up to 4 months to recruit sites
BONNIE Testing Tool: Background

- Web-based application for testing eCQMs
- Bonnie acts like an EHR
- Ensuring the measure logic, as specified, will behave properly in an EHR
- Build patients (test cases), with the goal of achieving 100% passing & 100% coverage
Test Cases

Goal:

– Take the characteristics and clinical experiences found within the specifications and combine them to create a test case that behaves according to each of the following categories for the measure logic:

• Initial Patient Population (Fail), Denominator (Fail), Denominator Exception (Fail), Numerator (Fail)
• Initial Patient Population (Pass), Denominator (Fail), Denominator Exception (Fail), Numerator (Fail)
• Initial Patient Population (Pass), Denominator (Pass), Denominator Exception (Fail), Numerator (Fail)
• Initial Patient Population (Pass), Denominator (Pass), Denominator Exception (Pass), Numerator (Fail)
• Initial Patient Population (Pass), Denominator (Pass), Denominator Exception (Fail), Numerator (Pass)

© 2020 PCPI Foundation. All rights reserved.
Scientific Acceptability

The opportunity to analyze the scientific acceptability of measures, using empirical data
Reliability – Are the results stable and repeatable?
Validity – Are you measuring what you intended?

by Experiment-Resources.com
• Conduct testing on a sample of accountable entities (e.g., hospital, physician) at the appropriate unit of analysis (e.g., physician, hospital, home health agency)

• Represent the variety of entities whose performance will be measured

• Units of measurement and patients within units should be randomly selected when possible

• Size should include adequate numbers of units of measurement and adequate numbers of patients
Reliability

• The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability.

• Commonly used methodologies
  – Signal-to-Noise Ratio Analysis
  – Data Element Testing / Parallel Forms
SNR Reliability Methodology

- The proportion of variation in the performance scores due to systematic differences across the measured entities (or signal) in relation to random error (or noise)
- Tests at the measure score level while comparing providers (or groups)
- Calculated as the ratio of the provider-to-provider variance divided by the sum of the provider-to-providers variance plus the error variance specific to a provider
  - Reliability 0 implies that all variability is due to measurement error.
  - Reliability of 1 implies that all variability is attributable to differences in provider performance.
- Reliability is assessed by using a beta-binomial model. The beta distribution is defined by two parameters, alpha and beta.
- Sampling considerations
SNR Analysis

- Reliability is estimated for each unit of analysis and we report
  - Overall reliability
  - Reliability distribution
- Higher reliability increases the chances you will classify a provider correctly in benchmarking
  - What level of reliability is sufficient?
    - High (90%)
    - Medium to High (70-80%)
    - Low (<70%)
Validity

• Validity Testing: empirical analysis of the measure as specified that demonstrates that data are correct and/or conclusions about quality of care based on the computed measure score are correct
• Commonly used validity testing methodologies
  – Validity against the gold standard (Parallel forms/Data element testing)
  – Face validity
  – Empirical validity
Parallel Forms Methodology

• Comparison of an automated report vs. manual abstraction (gold standard) from the medical record
• Requires testing EHR systems from more than one vendor
• Data is expected to be in structured fields
• Appropriate for EHR and Registry data
• More expensive and requires more time and resources
  • Site recruitment / contracts / costs
  • Training and consultation with sites
Parallel Forms

• Sites draw a sample from patients that have been determined to meet the Initial Patient Population for the measure

• Testing at the level of data elements requires that all critical data elements be tested

• Inter rater reliability statistical analysis is completed, using Cohen’s Kappa coefficient which takes into account any agreement occurring by chance

• Percent agreement and correlation is reported

• Resources:
Face Validity

• The extent to which an empirical measurement appears to reflect that which it is intended to, at “face value.”

• Assessed by a panel with relevant expertise
  – Did not participate in measure development
  – Have expertise on the measure topic and/or in measure development

• Collect data utilizing a Likert scale to indicate level of agreement

• Report the percent agreement
  – Number of Agree/Strongly Agree responses over the total eligible responses.
Empirical Validity

• The extent to which the measure quantifies what it claims to be measuring

• Correlation analysis of two scores at the provider level is conducted and a reliability score is calculated
  – Process measure: most common to compare against related measures that have been validated, often based on similarities in numerator or denominator
  – Outcome measure: most common to compare the prediction of outcomes across multiple points in time
Threats to Validity

- Exclusions & exceptions
- Measure scores from different data sources
- Missing or incorrect data
- Differences in patient mix
  - Stratification
  - Risk Adjustment
Discussion
The slides and recording will be posted at http://thepcpi.org

Stay tuned to your email for updates.

For general questions please contact us at info@thepcpi.org

For measurement questions, contact us at PCPImeasures@thepcpi.org