The Yin and Yang of Patient Outcomes Measures: 
Cons of Using Outcomes Measures for 
Assessment of Patient Improvement

Alicia Emerson Kavchak 
Lexie Wright 
Chad Cook

Questions (True/False)

1. At present, the majority of PT’s within the United States use patient outcomes measures
2. An MCID is a single point measure (for example, for an NDI) that is consistent across multiple diagnoses

Questions (True/False)

3. Most clinicians (physicians) cherry pick which patients they want recorded for outcomes measures
4. Bias is eliminated (mostly) using patient report outcomes measures such as the Oswestry, NDI, etc.

Outline

• Why outcome utilization isn’t practical
• Limitations of Outcomes Measures
• Key ingredients to an outcome measure
• The problem with MCID’s
• Care seeking behavior
• Patient Satisfaction

Acceptability

• Balancing the ease to complete the outcome measure against those who find it difficult or impossible to do
• Six Minute Walk Test ?!!

Feasibility

• Can the staff involved be able to administer the outcome measure
• Administrative Burden
  – Resources required
  – Time to train
  – Productivity Loss

Minimized with electronic outcome databases? (Deutscher 2008)

http://www.redbond.com/ecuadorart/burden_closeup.jpg
Respondent Burden

- The average healthcare oriented outcomes measure takes 7.5 minutes to complete.
- +10 minutes paperwork, tardiness, overlapping patients.

A Majority Aren’t Even Using Outcomes Measures!

- Only 48% of clinicians currently use outcomes measures.
- Why not?
  - Lack of time, money, and human resources needed to collect, analyze, and then make use of the data.
  - Lack of IT support for storing and retrieving data.
  - Lack of clinician knowledge in this area.
  - No currently available instrument adequately meets the psychometric or practical criteria necessary for use with individual patients.
  - The meaningful interpretation of score changes from patient-based measures of health is also problematic.


Cherry Picking

- 75% of respondents indicated “cherry picking” patients for outcomes.
- Requires a denominator.
- Otherwise, inflated outcomes.
- P4P.


Key Ingredients to an Outcome Measure

- Appropriateness
- Acceptability
- Feasibility
- Precision
- Responsiveness
- Reliability
- Validity


Appropriateness

- Does the outcome measure address what is important to the patient? (face validity)
- Does the outcome measure address the appropriate psychometric properties of the tool?

Which One!?

- Appropriateness?
- There are over 100 spine-related, healthcare oriented outcomes measures reported in the literature.
Number of Appropriate Constructs one Needs to Measure

- Pain
- Function
- Generic Health Status
- Work Disability
- Patient Satisfaction (with current situation)

Current Scale that Captures all Five Constructs?


Item Stagnancy

<table>
<thead>
<tr>
<th>Constructs</th>
<th>Scales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biases</td>
<td>Common scale anchors, extreme response bias, interviewer bias, recall bias, acquiescence bias, item social desirability, item ambiguity, common scale anchors</td>
</tr>
</tbody>
</table>

Case Example

- 45 y/o male with c/o LBP x 2 weeks
- Completed 6 visits of PT

Modified Oswestry LBP questionnaire:

- Initial Visit: 22%
- Final Visit: 8%
- Was this a significant change?
- Which one was successful?

Case example

- 20 year old male
- Baseline LEFS 75/80
- 1 month LEFS 78/80
- 1 month GRCS +7

- 67 year old male
- Baseline LEFS 21/80
- 1 month LEFS 34/80
- 1 month GRCS +2

So which one is successful?
Responsiveness

- Ability of outcome measure to determine change that is important to the patient
- The focus is on performance (internal) characteristics of the instrument
- What is significant for the group may not be significant for the individual
- **Evaluative** measures are those measures designed to measure changes within patients over a period of time. Guyatt 1993 Annals of Internal Medicine

What is Important Change?

- How do we measure what is important?
- Statistical significance does not necessarily mean the change is clinically important

Minimal Detectable Change

- Represents the smallest change in score that likely reflects true change rather than measurement error alone. Stratford, Beadle Phys Ther 1996
- (AKA the reliable change or smallest real difference).
- MDC\textsuperscript{95} (minimal detectable change at a 95% confidence interval) is calculated as:
  - \( MDC^{95} = 1.96 \times \sqrt{2} \times SEM. \)
- To calculate the SEM (standard error of measure)
  - \( SEM = SD \times \sqrt{1-ICC}. \)

Interpretability

- Minimally important clinical differences or changes in functional status. The focus is on the respondents’ views about the domain being studied. Medical Outcomes Trust 2004

Global Ratings of Change (GROC)

The GROC is a 15-point global rating scale ranging from −7 ("a very great deal worse") to 0 ("about the same") to +7 ("a very great deal better").

The Global Rating of Change Score (GRoC) Poorly Correlates with Functional Measures and is Not Temporally Stable: A Clinimetrics Corner

Not Longitudinally Stable
Methods

Anchor based
- Sensitivity and specificity based approach
- Within-patients score change
- Between-patients score change
- Social comparison approach

Distribution based
- Standard error of measurement
- Minimal detectable change
- Standard deviation
- Effect size

Sensitivity and Specificity based approach
- Use of ROC curves to discriminate between the "improved" and "unchanged" patients

Within patients score change
- MCID is defined as the average change of the patients defined as "improved"
- Thus, the mean change score of only those patients scoring > +5 on the GROC
- Arbitrary selection of cut-point on GROC

Between patients score change
- MCID is defined as the score difference between two adjacent levels on a GROC
- Thus, the mean change score of patients scoring > +5 on the GROC minus the mean changes score of patients scoring ≤ +5 on the GROC
- Again, selection of cut point on the GROC is arbitrary

Social comparison approach
- Patients are paired with other patients to discuss their health situation
- Following discussion, patients rate themselves as the same or to varying degrees of worse or better than the patient with whom they spoke
- MCID is defined as the difference in scores of patients who rate themselves as "a little better" or "a little worse" instead of "about the same" as compared to the other patient
- NOT WIDELY USED!!

SEM
- Defined as the variation in the scores due to the unreliability of the scale or measure used.
- A change smaller than the identified SEM is likely the result of error.
- Statistic alone, no patient importance noted.

Copay, AG. Understanding the minimum clinically important difference: a review of concepts and methods. The Spine Journal 2007;7:541-546
MDC
• Smallest detectable change that can be considered above the measurement error with a given level of confidence
• MDC95

Standard deviation
• Variation among a group of scores

Effect size
• Standardized measure of change obtained by dividing the difference in scores from baseline to post-treatment by the SD of the baseline scores
• The value of the effect size represents the number of SDs by which the scores have changed from baseline to post-treatment
• i.e. effect size should be small in patients reporting no change and large in patients reporting great improvement
• Change in scores corresponding to the small effect size is considered the MCID
• MCID = SD(baseline) x 0.2(small effect size)

Which one do I choose?
• There are a range of acceptable values reported in the literature based on the methodology chosen!

It depends on the patient!!
• Would you use the same MCID??

her?

Vs.

him?
Subjects: N = 6,651; Orthopedic knee impairments
Treatment: Outpatient physical therapy
Outcome measures:
- LEFS (0-100 scale)
- GROC (15 point scale); ≥ 3

Results: Based on baseline intake value

<table>
<thead>
<tr>
<th>Variable</th>
<th>Improved (Change ≥ 15)</th>
<th>No Change</th>
<th>LEFS</th>
<th>Final</th>
<th>LEFS/100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.04</td>
<td>0.96</td>
<td>0.91</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>Female</td>
<td>0.96</td>
<td>0.04</td>
<td>0.91</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>Sex</td>
<td>0.51</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.51</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Result: Based on sex, symptom acuity, and age

Participants who were male, younger, or had more acute symptoms required more change to report meaningful change
MCII is context specific and not a fixed attribute!

Case example

- 45 y/o male with c/o LBP x 2 weeks
- Completed 6 visits of PT

Modified Oswestry LBP questionnaire:
Initial Visit: 22%
Final Visit: 8%
Was this a significant change?
Which one was successful?

Case example

- 20 year old male
- Baseline LEFS 75/80
- 1 month LEFS 78/80
- 1 month GRCS +7

- 67 year old male
- Baseline LEFS 21/80
- 1 month LEFS 34/80
- 1 month GRCS +2

So which one is successful?
Misclassification

- Most MCID estimates are extremely specific and lack sensitivity
- At the individual patient level, this will misclassify people below the mean as not having experienced an important change when in fact they have

Beaton, 2002

Problems with the mean

- Most studies produce mean change scores in those patients who experienced a small but important change.
- Often, investigators and clinicians assume that whether the difference in the means is less than the MIC, the treatment effect is unimportant.
- However, individual patients may have shown clinically important improvement.

Let's use an example

- Is manual therapy an effective treatment for improving function in patients with hip OA?
- Outcome measure used is the WOMAC function subscale
- MCID = 6.5 points
- 70 patients underwent MT treatment
- Mean change score = 6 points
- Conclusion: Manual therapy is ineffective in improving function in patients with hip OA
- WHY??!

Ostelo, 2008

We are asking the wrong question!

- “How much?” vs. “How important?”
- What defines “important” to the patient?
  - + 3 “somewhat better”?
  - + 5 “quite a bit better”?

Terwee, 2010

MCID ≠ MDC

- The minimal detectable change is the smallest change in score than can be detected beyond random error and is dependent upon sample distribution.
- MCID can occur on either side of any statistical threshold and is determined by the patients in quality of life measures

Example

- **QuickDASH**
  - MCID 8.1 % points
  - MDC 11.2 % points


Conclusions

- **PROCEED WITH CAUTION!!**
- Consider alternative outcome measures

Pay for performance

- Reward for those meeting pre-established targets for performance
  - i.e. Those meeting the MCID receive reimbursement vs. those who do not are subject to reductions in payment
- Are you confident in using the MCID or other quality indicators that are not evidence based?

Also......P4P

- Inability to identify the part of the healthcare process (PT, MD, natural history) that is responsible for the change in a patient
- Assumes care is related to outcome

Not as Precise as one Might Think

- Quality measures that focus on clinical outcomes rather than specific processes of care create other problems of attrition.


Westerman et al. Listen to their answers! Response behaviour in the measurement of physical and role functioning. Qual Life Res. 2008 May;17(4):549-58.
Do Condition Specific Measures Capture problems from the Condition?

<table>
<thead>
<tr>
<th>Condition</th>
<th>Specific Measures</th>
<th>Conditions</th>
<th>Outcome Results</th>
<th>Policies</th>
<th>Ethical Policy Decisions</th>
<th>Research on This</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression, BMI, education, and smoking status</td>
<td>changed disability scores</td>
<td>Goode and Cook. Pain Practice. 2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Logistic Regression Model Beta Coefficients and Odds Ratios for Severe Back Pain and Severe Leg Pain

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>SE</th>
<th>Odds Ratio</th>
<th>CI (95%)</th>
<th>SE</th>
<th>Odds Ratio</th>
<th>CI (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe back and leg pain</td>
<td>0.58 (0.04)</td>
<td>0.02</td>
<td>1.40</td>
<td>0.99 (1.19)</td>
<td>0.03</td>
<td>1.40</td>
<td>0.99 (1.19)</td>
</tr>
<tr>
<td>Severe complaint leg pain</td>
<td>0.68 (0.03)</td>
<td>0.02</td>
<td>1.98</td>
<td>1.66 (2.38)</td>
<td>0.03</td>
<td>1.98</td>
<td>1.66 (2.38)</td>
</tr>
</tbody>
</table>

Percent: (NADA)

0 Percent

• How many times CMS adjusts payment rate because of case mix or severity of illness
• No assessment of co-morbidities

This vs This


Policy Decisions

• Outcome results are used to make policy decisions
• Politicians make decisions based on the betterment of people’s well being, right?

Ethical Policy Decisions

• Strength in numbers (e.g., number of studies with confirming information)
• Strength in results (e.g., results are simply too strong to deny)
• Admiration from others (overwhelming evidence for support)

Abood. S. Influencing Health Care in the Legislative Arena. OJIN 2007;12:1

Research on This

• Clinicians and researchers often employ different epistemologies.
• Clinicians often fear that the research is designed to evaluate “them” not the intervention
• Researchers answer questions no one is asking

**What Drives Health Use?**

1. Demographic and Social Characteristics
2. Available Resources
3. Perceived need of medical attention

Anderson et al. J Health Soc Behav 1995

**Others**

- 1) Gravity of problem, 2) knowledge of mechanisms of self treatment, 3) faith in treatment, 4) Faith in efficacy of treatment and 5) Accessibility of treatment


**Predisposing Factor and Health Seeking Behavior**

Adverse / stressful event is associated with
- Development of chronic LBP
- Depression
- Disability
- Health seeking behavior


**Why do Those with Chronic Hip or Knee Pain Seek Health Care?**

- Mobility issues (OR: 2.62)
- They live in an urban area (OR: 2.40)
- Pain (OR: 1.72)


**What about those with LBP?**

- Disabling pain (OR: 6.6)
- Fear (OR: 2.2)

A True Litmus Test?

- Care seeking behavior after care has been received

Best Predictor of Care Seeking?

<table>
<thead>
<tr>
<th>Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>General practitioner only</td>
<td>5</td>
</tr>
<tr>
<td>Physiotherapy</td>
<td>13</td>
</tr>
<tr>
<td>Hospital consultant</td>
<td>11</td>
</tr>
<tr>
<td>Osteopathy</td>
<td>43</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>76</td>
</tr>
</tbody>
</table>

- Depression


Good Disability Scores Initially

Do Outcomes Link with Care Seeking Behavior?

- Consider LBP
- Maybe, not necessarily, especially in cases that involve recurrence
- Pain and Disability seem to trigger care seeking (but we are uncertain if it predicts it)


Patient Satisfaction?

- CONNECT data
- Doesn’t seem to matter
- Patient Loyalty likely matters more

PT / Patient Relationship

Positive Alliances influences:
1) The patient’s overall perceived effect of treatment
2) The patient’s satisfaction with treatment
3) Change in pain
4) Physical function
5) Depression
6) General health status

### What Really Matters to the Patient?

- Retrospective analysis of change was more sensitive than outcome measures completed repeatedly


### What Should Really Matter to You and Your Patients?

- Utilization of outcome measures equally among all patients – make it a habit!
- Attempt for disease specific measures
- Include Global Rating of Change
- Include Patient Satisfaction
- Weigh policy decisions carefully