Methods for Assessing Clinical Significance of Quality of Life Measures (worms, ducks, and elephants)

Jeff A. Sloan, Ph.D.
Quality of Life Research, Mayo Clinic, Rochester, MN

FDA Meeting, Silver Springs, MD, November 27, 2012
What am I talking about?

But if we didn’t measure things we wouldn’t know how good we were at measuring the things that we’re measuring!
Take home messages

• There are problems with assessing the clinical significance of QOL as indicators of clinical efficacy in clinical trials.

• There are scientifically sound solutions to these problems. The problems have been disseminated widely and consistently. The solutions have not.
Content

• 1) Clinical significance history
• 2) History of the $\frac{1}{2}$ sd for PROs
• 3) $\frac{1}{2}$ sd for survival, tumor response
• 4) Evolution: rise of the regulators
• 5) Responder analysis
• 6) the future?
So how big is a clinically significant difference?
Bottom Line

• Assessing the clinical significance of a PRO can be as simple as a 10-point change on a 100-point scale, if that is consistent with the goals of the scientific enquiry. The real issue underlying the controversy over QOL is the relative novelty and lack of experience that presently exists with QOL. With time and familiarity this too shall pass.

SYMPOSIUM ON THE CLINICAL SIGNIFICANCE OF
QUALITY-OF-LIFE MEASURES IN CANCER PATIENTS

Assessing Clinical Significance in Measuring Oncology Patient Quality of Life:
Introduction to the Symposium, Content Overview, and Definition of Terms—
J. A. Sloan, D. Cella, M. H. Frost, G. H. Guyatt, M. A. G. Sprangers, T. Symonds,
and the Clinical Significance Consensus Meeting Group

Methods to Explain the Clinical Significance of Health Status Measures—
G. H. Guyatt, D. Osoba, A. W. Wu, K. W. Wyevich, G. R. Norman, and
the Clinical Significance Consensus Meeting Group

Group vs Individual Approaches to Understanding the Clinical Significance of Differences
or Changes in Quality of Life—D. Cella, M. Bullinger, C. Scott, I. Barofsky, and
the Clinical Significance Consensus Meeting Group

Assessing the Clinical Significance of Single Items Relative to Summated Scores—
J. A. Sloan, N. Aaronson, J. C. Cappelleri, D. L. Fairclough, C. Varricchio, and
the Clinical Significance Consensus Meeting Group

Patient, Clinician, and Population Perspectives on Determining the Clinical Significance
of Quality-of-Life Scores—M. H. Frost, A. E. Bonomi, C. E. Ferrans, G. Y. Wong,
R. D. Hays, and the Clinical Significance Consensus Meeting Group

Assessing Meaningful Change in Quality of Life Over Time: A Users’ Guide
for Clinicians—M. A. G. Sprangers, C. M. Moinpour, T. J. Meynihan, D. L. Patrick,
D. A. Reivicki, and the Clinical Significance Consensus Meeting Group

The Clinical Significance of Quality-of-Life Results: Practical Considerations for
Specific Audiences—T. Symonds, R. Berzon, P. Marquis, T. A. Rummons, and
the Clinical Significance Consensus Meeting Group

Reprinted From Mayo Clinic Proceedings
Volume 77, April-June 2002
The Six Papers

1) Methods used to date
2) Group versus individual differences
3) Single item versus multi-item
4) Patient, clinician, population perspectives
5) Changes over time
6) Practical considerations for specific audiences

MCP, April, May, June 2002
Why is it difficult to define “clinical significance” for PROs?

• Blood pressure analogy
  • 100 years ago, clinical significance of BP scores was unknown
  • massage therapy was the gold standard
  • present guidelines for BP clinical significance today?

• Pain analogy
  • 25 years ago physicians were the sole raters of patient pain
  • JCAHO 2000 guideline: every patient’s pain to be assessed upon intake on a 0-10 scale
What Clinical significance is NOT

- Statistical significance

- e.g. HSQ before / after scores on 1300 pts. (JCO, 2002)
  - all p-values <0.0001
  - stated conclusion: “clinically significant changes in all domains of QOL”
  - 80% power to detect a change of 1 unit on 0-100 point scale
### EORTC QLQ-LC13

<table>
<thead>
<tr>
<th>Item</th>
<th>n=537</th>
<th>n=346</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coughing</td>
<td>46.2</td>
<td>44.3</td>
<td>small</td>
</tr>
<tr>
<td>Dyspnea</td>
<td>17.2</td>
<td>16.2</td>
<td>small</td>
</tr>
<tr>
<td>Pain</td>
<td>26.9</td>
<td>25.5</td>
<td>small</td>
</tr>
</tbody>
</table>

- All p-values were statistically significant
The significance of statistical significance

- Medline search identified over 200 letters to journals since 1970 with this topic
- Numerous examples of published statistically significant results that clinicians responded to as being clinically irrelevant and would not impact practice
General Classification for Methods Assessing Clinical Significance

- Tool-specifications (norms, experience)
- Investigator-defined (ES, SEM, ERES, RSQR)
- a posteriori patient-defined (MCID, MID)
- a posteriori statistically-defined (emergent p-value)
- anchored to clinical outcome
  - (e.g., ability to walk, redundant)
    - Guyatt, 2002; Sloan, 2002, MCP)
Two general methods for clinical significance

- Anchor-based methods requirements
  - independent interpretable measure (the anchor) which has appreciable correlation between anchor and target

- Distribution-based methods
  - rely on expression of magnitude of effect in terms of measure of variability of results (effect size)
    - (Guyatt, 2002; Sloan, 2002)
Fig 1. Relationship between SSQ ratings of change and QLQ-C30 scores from T1 to T2 for patients receiving chemotherapy for either breast cancer (A) or SCLC (B). Columns represent mean scores + 2 SE. □, physical functioning; ■, emotional functioning; , social functioning; ■, global QL.
Foundation of the $\frac{1}{2}$ SD method for “clinical significance”

- “Is the difference in QOL the size of a worm, a duck or an elephant?”.

Definition: Juniper’s duck

“the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate in the absence of troublesome side-effects and excessive cost, a change in the patient’s management”

(Juniper et al, 1994)
The Empirical Rule

• Tchebyshev’s Theorem: at least $1 - \frac{1}{k^2}$ of any distribution will fall within $k$ standard deviations of the mean

• If the distribution is symmetric, 9% will fall within 3 standard deviations

• The pdf for the range is a function of the sd

• an estimate of the s.d can be obtained via
  • range = 6 s.d.
The ERES Approach

• QOL tool range = 6 standard Deviations
• SD Estimate = 100 percent / 6
  = 16.7% of theoretical range

• Two-sample t-test effect sizes (Cohen):
  small, moderate, large effect (0.2, 0.5, 0.8 SD shift)

• S,M,L effects = 3%, 8%, 12% of range
Power curves for small, medium, and large effect sizes (worms, ducks and elephants via a two-sample t-test with a 5% type I error rate and two-sided alternative hypothesis.
Methods equated

- Cohen - 1/2 SD is moderate effect

- MCID - 1/2 point on 7-point Likert
  - 7-1 = 6 point range ==> SD of 1 unit
  - so 1/2 point ==> 1/2 SD

- Cella - 10 point on FACT-G
  - 10/1.12 = 8.9% / 16.7% = 1/2 SD

- Feinstein - correlation approach
  - Cohen was arbitrary, should be 0.6 SD
Clinically Significant Effect Sizes from the Literature

- Norman et al (Med Care 2003 May; 41(5):582-92.)
  - Remarkable universality
  - consensus
  - Triangulation

... and many more all recognizing that irrespective of the technique, the answers were all similar (comparable, same ballpark)
Effect Sizes from the Literature Expressed as a Percentage of a QOL measure’s 0-100 Point Theoretical Range.

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Small Effect</th>
<th>Moderate Effect</th>
<th>Large Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaeschke et al.\textsuperscript{22}</td>
<td>8%</td>
<td>15%</td>
<td>20%+</td>
</tr>
<tr>
<td>Juniper et. al.\textsuperscript{26}</td>
<td>8%</td>
<td>15%</td>
<td>20%+</td>
</tr>
<tr>
<td>Redelmeir et al.\textsuperscript{27}</td>
<td>8%</td>
<td>17%</td>
<td>25%+</td>
</tr>
<tr>
<td>Osoba et al.\textsuperscript{41}</td>
<td>5-10%</td>
<td>10-20%</td>
<td>20%+</td>
</tr>
</tbody>
</table>
Proposal: A Unified Theory Approach

• All methods can be equated

• define a movement equivalent to 1/2 S.D. as a NON-IGNORABLE required shift for clinical significance on any domain or individual item

• protocol-specific exceptions can be defined a priori with supportive evidence
  • - (e.g. 1/4 S.D. is important here because…)
Presenting global solutions is always interesting
A duck or a robin?

• Cella et al, JCE 55: 285-295, 2002 proposed 1/3 SD as a “robin” for LCS and TOI.

• In ERES terms, Cella’s robin = 5.5% of TR, compared to Sloan’s duck = 8% of TR.

• The difference represents answering one category different on one question out of 7 or 20 questions.

• Duck is a guideline, robin is a potential alternative buoyant waterfowl: justification?
Percent of Patients Achieving a 1/2 SD Improvement

<table>
<thead>
<tr>
<th></th>
<th>Music</th>
<th>No Music</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diastolic Blood Pressure</td>
<td>29%</td>
<td>16%</td>
<td>0.0529</td>
</tr>
<tr>
<td>Systolic Blood Pressure</td>
<td>37%</td>
<td>22%</td>
<td>0.0348</td>
</tr>
<tr>
<td>Pulse</td>
<td>20%</td>
<td>16%</td>
<td>0.7012</td>
</tr>
<tr>
<td>Respiration</td>
<td>2%</td>
<td>5%</td>
<td>0.4440</td>
</tr>
<tr>
<td>Physical Well-Being</td>
<td>51%</td>
<td>23%</td>
<td>0.0001</td>
</tr>
<tr>
<td>Emotional Well-Being</td>
<td>47%</td>
<td>14%</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Spiritual Well-Being</td>
<td>34%</td>
<td>4%</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Intellectual Well-Being</td>
<td>30%</td>
<td>10%</td>
<td>0.0008</td>
</tr>
<tr>
<td>Overall Well-Being</td>
<td>58%</td>
<td>20%</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Fatigue</td>
<td>40%</td>
<td>29%</td>
<td>0.1634</td>
</tr>
<tr>
<td>Nausea</td>
<td>21%</td>
<td>13%</td>
<td>0.2370</td>
</tr>
<tr>
<td>Anxiety</td>
<td>39%</td>
<td>20%</td>
<td>0.0056</td>
</tr>
<tr>
<td>Sadness</td>
<td>37%</td>
<td>20%</td>
<td>0.0136</td>
</tr>
<tr>
<td>Pain</td>
<td>29%</td>
<td>28%</td>
<td>1.0000</td>
</tr>
<tr>
<td>Relaxed</td>
<td>47%</td>
<td>13%</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Overall Quality of Life</td>
<td>21%</td>
<td>12%</td>
<td>0.1617</td>
</tr>
</tbody>
</table>
Clinically Significant Difference in Average Treatment Effects

* = 1/2 SD (clinically significant)
So I Had Another Idea...

If you believe in yourself...

You are a fool!
How would the $\frac{1}{2}$ SD work for other “hard” clinical endpoints?

When Life Gives You Lemons, Make Lemonade

Unfortunately, life would also have to give you sugar for that. Enjoy your lemon juice.
A Bit of Background Math

• assume survival time of patient x follows an exponential distribution of \( f(x) = \frac{1}{t} \exp(x/t) \) for \( x \geq 0 \)

• Where \( t = \) mean lifetime,

• Then it follows directly that
  • \( E(x) = t \)
  • \( V(x) = t^2 \)
  • \( Sd(x) = t \)

And finally \( \Rightarrow \frac{1}{2} \text{SD}(x) = \frac{t}{2} \)
A Bit of Background Math

- Similarly for median survival, we find that

- $S_d(x) = \frac{t}{2\ln2}$

- And finally $\frac{1}{2} SD(x) = \frac{t}{2\ln2}$
### ½ SD for Mean Survival Analysis

<table>
<thead>
<tr>
<th></th>
<th>Mean survival = 6 months</th>
<th>Mean survival = 1 year</th>
</tr>
</thead>
<tbody>
<tr>
<td>½ SD</td>
<td>3 months</td>
<td>6 months</td>
</tr>
<tr>
<td>¼ SD</td>
<td>1.5 months</td>
<td>3 months</td>
</tr>
<tr>
<td>1/5 SD</td>
<td>1.2 months</td>
<td>2.4 months</td>
</tr>
</tbody>
</table>
### ½ SD for Median Survival Analysis

<table>
<thead>
<tr>
<th></th>
<th>Median survival = 6 months</th>
<th>Median survival = 1 year</th>
</tr>
</thead>
<tbody>
<tr>
<td>½ SD</td>
<td>4.3 months</td>
<td>8.6 months</td>
</tr>
<tr>
<td>¼ SD</td>
<td>2.1 months</td>
<td>4.0 months</td>
</tr>
<tr>
<td>1/5 SD</td>
<td>1.2 months</td>
<td>2.4 months</td>
</tr>
</tbody>
</table>
½ SD for survival: An example

- Patients were randomized to receive single agent lapatinib (1500 mg/daily) or a combination of lapatinib (1000 mg p.o. daily) plus trastuzumab (2 mg/kg).

- Women treated with monotherapy lapatinib experienced a median overall survival of 9.5 months compared with 14 months when treated with the combination (median HR: 0.74, p=0.026).

...could be calibrated by saying that the effect is one standard deviation which we know to be a HUGE effect size. Note the p-value might be interpreted as some as “barely significant”.
½ SD for Survival: A review of the literature

- Tannock, JCO 2008, 321 RCT trials
- Survival effect size in design ranged between 0.01SD and 1.5 SD. Huge heterogeneity without cause
- Average effect size was 0.18 – 0.3 SD depending upon tumor type
- Generally smaller effect sizes than ½ SD declared “significant”
A Bit of Background Math

- Similarly for tumor response, we find that
  - \( S_d(x) = \sqrt{p(1-p)} \)
  - And \( \frac{1}{2} \) \( SD(x) = \frac{1}{2} \times \sqrt{(p(1-p))} \)
### ½ SD for Tumor Response Analysis

<table>
<thead>
<tr>
<th>Tumor Response = 50%</th>
<th>Tumor Response = 25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>½ SD</td>
<td>25%</td>
</tr>
<tr>
<td>¼ SD</td>
<td>12%</td>
</tr>
<tr>
<td>1/5 SD</td>
<td>10%</td>
</tr>
</tbody>
</table>
CONCLUSION: The $\frac{1}{2}$ sd method allows for more ready interpretation of the clinical significance of survival and response studies. It allows for direct cross study comparison even if the endpoints are different. It also facilitates study design as it builds clinical significance into the study directly.
Meanwhile, the regulatory agencies got involved,
With mixed results
Enter the PRO guidance process
Enter the Committees and the Politics and our Tendencies

A committee is a group of people who individually can do nothing but together can decide that nothing can be done. - Fred Allen
The state of the science?

Distribution-based
Anchor-based

NO, RALPH, THERE ARE TWELVE INCHES IN A FOOT...NOT THIRTEEN!

When Dinosaurs ruled the Earth
MID

• Why is MID not included in Final PRO Guidance?
  • Term is interpreted inconsistently (intra-patient change vs. inter-group difference of mean change from baseline)
  • Responder definitions offer a direct approach to intra-patient change and treatment differences across a range of clinical anchors that can be presented in a cumulative distribution function
Kathy Wyrwich sums up the result
The MID is DEAD

(Evidencemattersonline www.united biosource.com 2010)

Thou shalt define a responder and use a cumulative distribution function
PRO’s and the FDA: where are we and where are we going?
PRO Guidance – Interpretation of Data

i. Hypothesize Conceptual Framework
- Outline hypothesized concepts and potential claims
- Determine intended population
- Determine intended application/characteristics (type of scores, mode and frequency of administration)
- Perform literature/expert review
- Develop hypothesized conceptual framework
- Place PROs within preliminary endpoint model
- Document preliminary instrument development

v. Modify Instrument
- Change wording of items, populations, response options, recall period, or mode/method of administration/data collection
- Translate and culturally adapt to other languages
- Evaluate modifications as appropriate
- Document all changes

ii. Adjust Conceptual Framework and Draft Instrument
- Obtain patient input
- Generate new items
- Select recall period, response options and format
- Select mode/method of administration/data collection
- Conduct patient cognitive interviewing
- Pilot test draft instrument
- Document content validity

iv. Collect, Analyze, and Interpret Data
- Prepare protocol and statistical analysis plan (final endpoint model and responder definition)
- Collect and analyze data
- Evaluate treatment response using cumulative distribution and responder definition
- Document interpretation of treatment benefit in relation to claim

iii. Confirm Conceptual Framework and Assess Other Measurement Properties
- Confirm conceptual framework with scoring rule
- Assess score reliability, construct validity, and ability to detect change
- Finalize instrument content, formats, scoring, procedures and training materials
- Document measurement development
Motivation

• Define a threshold for the change from baseline in the continuous variable

• Then define a patient to be a “responder” if the patient’s change value is on a particular side of this threshold

• FDA reviewers will evaluate a PRO instrument’s responder definition in the context of each specific clinical trial

• Two reasons for doing so
  • To simplify analysis and interpretation
  • To ensure that a reported statistically significant result represents a clinically meaningful benefit
Responder Analyses as Primary or Co-Primary Analyses in Late Registration Studies: Examples

• Acute stroke, with response being functional independence based on a cut point on Barthel Index or the Modified Ranking Scale

• Depression, with response being a 50% improvement on accepted rating scales

• Parkinson’s disease, with response being the degree of symptom reduction from baseline

• Rheumatoid arthritis, with response being 70% or greater improvement in a combination of signs and symptoms
Cumulative Distribution Function

• An alternative or supplement to responder analysis

• Encompasses all available data

• Mentioned prominently in the FDA Guidance on PRO label and promotional claims
Illustrative Cumulative Distribution Function: Experimental Treatment (solid line) better than Control Treatment (dash line) -- Negative changes indicate improvement
CDF results that do not demonstrate the comparative efficacy of Drug A or Drug B.
Example Result for Demonstrating the Efficacy of Drug A over Drug B
Figure 5. Cumulative Percentage of Patients with Specified Changes from Baseline ADAS-cog Scores. The Percentages of Randomized Patients Within Each Treatment Group Who Completed the Study Were: Placebo 93%, 5 mg/day 90% and 10 mg/day 82%.
Figure 1: Percentage of Patients Achieving Various Levels of Pain Relief as Measured by 24-Hour Average Pain Severity - Study 1
So where are we now?

- FDA reconvening meetings to define clinical significance in various committees (osteoporosis this month)
- Mayo clinical significance group reforming to update state of the science
- Stay tuned....
• Hopefully we can work together to come up with a practical solution for all concerned.