Interpretation

Clinical significance: what does it mean?

Patrick Marquis, MD, MBA
Mapi Values - Boston

DIA workshop
Assessing Treatment Impact Using PROs:
Challenges in Study Design, Conduct and Analysis
Paris May 10-11, 2004
Context

- Context of the interpretation of clinical trial results assuming:
  - Absence of issue regarding trial design, quality and analysis
  - PRO questionnaire developed and validated according to recognized standards
  - Hit on primary end-point or clinical efficacy end-point
Question

- Statistical significance reached however...
- Null hypothesis may be rejected for very small differences by large sample sizes
- Differences may be of no practical importance

Clinical meaning and practical value?
ClinSig definition

- **Patient perspective**
  - A difference large enough to have an implication in the patient management
  - A value that patients recognize as minimal important difference (MID)

- **Clinical perspective**
  - Smallest effect size leading clinicians to recommend a treatment
ClinSig in PRO research
= like any other end-points

Specificities

- Concepts vary across questionnaires
- Scale label not always informative
- Number of items and scaling vary
- Differences in anchors, direction, and magnitude of change
Researcher perspective

\[ \Delta HRQL \]

Distribution → Paired t test → f, p

Distribution → Standard. Ratio → ES, SRM, RS

Distribution → Reliable change → SEM
Standardized Ratio

- Helpful to understand the size of the change based on a standardized unit
- Commonly used in clinical research
- Easy to interpret
- Interpretation guidelines (Cohen)
  - .2 to .49 small / .5 to .79 moderate / .8+ large
  - For sample size estimates
**Role of effect size**

- **Relative efficiency of measures to detect treatment effect for leflunomide versus placebo in RA (Tugwell, 2000)**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tender joint count</td>
<td>-4.69</td>
<td>-0.59</td>
</tr>
<tr>
<td>Swollen joint count</td>
<td>-2.78</td>
<td>-0.44</td>
</tr>
<tr>
<td>Pain intensity</td>
<td>-17.51</td>
<td>-0.65</td>
</tr>
<tr>
<td>ESR</td>
<td>-8.82</td>
<td>-0.41</td>
</tr>
<tr>
<td>CRP</td>
<td>-1.09</td>
<td>-0.47</td>
</tr>
<tr>
<td>HAQ disability index</td>
<td>-0.42</td>
<td>-0.80</td>
</tr>
<tr>
<td>SF-36 MCS</td>
<td>-0.65</td>
<td>-0.06</td>
</tr>
<tr>
<td>SF-36 PCS</td>
<td>-8.52</td>
<td>-0.67</td>
</tr>
<tr>
<td>SF-36 bodily pain</td>
<td>-15.76</td>
<td>-0.73</td>
</tr>
</tbody>
</table>
Standard Deviation as a key element

- Increased variance due to confounding factors such as gender and age
  - ES should integrate these confounders

- In the SEM, the higher the reliability:
  - The lower the true effect
  - The lower the corresponding ES

\[ \begin{align*}
  \text{SD} &= 20 \\
  r=.70 &\implies \text{SEM}=11 \implies \text{ES}=0.55 \\
  r=.80 &\implies \text{SEM}=8.9 \implies \text{ES}=0.45 \\
  r=.90 &\implies \text{SEM}=6.3 \implies \text{ES}=0.32
\end{align*} \]
△ HRQL

Researcher perspective

Distribution
- Paired t test
  - f, p
- Standard. Ratio
  - ES, SRM, RS
- Reliable change
  - SEM
- Transition rating
  - MID - MDD
- Clinical
  - MID - MCID - CID

Anchors
Transition rating

- Within-patient transition rating (magnitude of their global change)
- Between-patient difference rating (magnitude of perceived differences)
Within-patient transition rating

- Patient global rating of change after treatment (Juniper 1996)

  7  A very great deal change
  6  A great deal change
  5  A good deal change
  4  Moderately change
  3  Somewhat change
  2  A little change
  1  Almost the same
  0  No change

Mean HRQL change of patients considered as the MID = 0.5 for a 1 to 7 scale
“The subjective impact of asthma on patients’ perception of health was evaluated through use of a validated instrument called the Asthma Quality of Life Questionnaire (AQLQ). Patients receiving ADVAIR DISKUS 100/50 had clinically meaningful improvements in overall asthma-specific quality of life as defined by a difference between groups of >= 0.5 points in change from baseline AQLQ scores (difference in AQLQ score of 1.25 compared to placebo).”
Transition rating as the anchor?

- Easy to implement and intuitive
- Consistency with other methods
- Used for most questionnaires
Transition rating as one anchor

- **Issues**
  - Single measure
  - Pattern of correlation pre-post test not usually as expected
  - Arbitrary decisions and sensitivity to wording
  - Absence of clinical anchor / interpretation
  - Mainly developed for within-patient changes but used for between-patient decision
Clinical anchor

- Meaningful clinically and for patients
- “Some” level of correlation with the clinical anchor
- Threshold or calibration needed for the clinical anchor
Clinical anchor

Within-group MID = -0.73

Between-group MID = 1.75

Slope of the line

Change in HB (g/dl)

Application of MID - MDD - CID

- Benchmark for interpreting change
  - Not a hard threshold
- Patient Benefit from treatment
  - Responder analysis
- NNT Number Needed to treat

- Interpretation of change over different time points
- Sample size calculation
Interpretation issue in QoL

Control of SAR symptoms
Improvement of HRQL
- F vs PCB  p = 0.005
- F vs L   p = 0.03

(Clin Exp Allergy 2000; 30: 891-899)
Percentage reaching the MID

- F: 74%
- L: 64%
- PCB: 61%
## NNT results

### F

<table>
<thead>
<tr>
<th>L</th>
<th>Improved (0.74)</th>
<th>Unchanged (0.23)</th>
<th>Deteriorated (0.03)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved (0.64)</td>
<td>0.4691</td>
<td>0.1494</td>
<td>0.0174</td>
</tr>
<tr>
<td>Unchanged (0.26)</td>
<td>0.1919</td>
<td>0.0611</td>
<td>0.0071</td>
</tr>
<tr>
<td>Deteriorated (0.10)</td>
<td>0.0768</td>
<td>0.0244</td>
<td>0.0028</td>
</tr>
</tbody>
</table>

**NNT = 1/(Net Benefit)**

**Net Benefit = 0.2931 – 0.1739 = 0.1192**

- **NNT = 8.4 for F vs L**
- **NNT = 7.2 for F vs PCB**
- **NNT = 57.2 for L vs PCB**
Interpretation of the NNT

- About 7 patients should be treated with F to get an additional meaningful HRQL benefit over placebo
- About 8 patients should be treated with F to get an additional meaningful HRQL benefit over L
- About 57 patients should be treated with L to get an additional meaningful HRQL benefit over placebo
Integration of multiple time points
Time to definitive 5-point HRQL deterioration

Arm A (n=154)
3 deaths (9.1%)
HRQL deterioration in 30 (90.9%)

Arm B (n=138)
3 deaths (7.5%)
HRQL deterioration in 37 (92.5%)

P = 0.03
Universality of half-SD?

- Consistency of MID determined by anchor-based methods with ES?
- 38 studies reviewed
- For 32 studies MID closed to half-SD (.495 ± .155)
  - No influence of the number of response options and type of questionnaires
- Larger ES for negative changes & acute conditions
- Smaller ES for population-based estimation

Norman G, Med Care 2003 41,5, 582-592
△ HRQL

Distribution
- Paired t test
  - f, p
- Standard. Ratio
  - ES, SRM, RS
- Reliable change
  - SEM
- Transition rating
  - MID - MDD
- Clinical
  - MID - MCID - CID

Anchors

Practical Value
- Mortality Resource used...
  - Correlation Predictive value

Researcher perspective
Practical Value / Added value

- Prediction of mortality, resource utilization
  - See first session
- Strong message
- Final confirmation of questionnaire validity
- Limitation to define the “minimum” different concept
- Used for establishing the robustness of the questionnaire
- Not commonly used for RCT interpretation
Researcher perspective

△ HRQL

Distribution

Anchors

Practical value

Distribution

Norms / other known-groups

Full distribution

Scores Overtime & at EP
Treatment effect in PAR compared to normative data
Decision-making perspective

- Drug approval
- Label development
  - Not the indication but useful additional information for prescribers that should be mentioned in the SPC
- Promotional material approval
  - Evidence supporting the promotional claim
Reviewer perspective

△ HRQL

- Distribution
  - ES, SRM, SEM…
  - Magnitude of differences
- Anchors
  - MID, MDD, MCID, CID …
  - Map to benchmarks
- Distribution
  - Distribution norms/ known groups
  - Map to “normative” levels
  - Simple figures…

Scores overtime & at EP
Recommendations

- Establish links between clinical variable, PRO data and clinical significance
- Consistency with other results key factor
- Integrate clinical context as will always be different
  - Cancer
  - Asthma
  - Anemia
  - Irritable Bowel Disease
Conclusion

- No universally accepted approaches to determine the ClinSig
- No single approach is perfect
- Half-SD seems “safe” but
  - More research needed
  - Smaller ES can be meaningful +++
    - see Cohen, Miller, and SEM
  - Comparison of active treatments
  - Increased variance due to confounders
Facilitate the task of reviewers!

- Provide supportive evidence:
  - Documentation on the development
    ⇒ “face value”
  - Documentation of the validation with executive summary
    ⇒ “properties”
  - Interpretation guidelines (different approaches)
    ⇒ “meaning”
- Standardization & quality of PRO submission...
Points of Discussion

- Is the clinical significance only a matter of method and technique?
- Should it be universal (mathematical model) and or specific to the clinical context?
- Is the search for thresholds mandatory?
- How to choose the right anchor for determining the MID?
- Are post-doc analyses acceptable to further analyze the clinical significant?