Patient-reported outcomes : Establishing equivalence between electronic (ePRO) and paper methods

Brian Tiplady

October 21st 2008

PRO Consulting®
Refining Patient Reported Outcomes

Why use ePRO?

- Only valid, in-range entries can be made
- Time-stamping of entries and time windows
  - Compliance can be documented
  - Retrospective entry can be prevented or managed
- Missing data can be reduced or eliminated
- Data available for prompt review.
- Reminders and feedback enhance compliance
- ePRO easy to use and generally preferred to paper
  - Including the elderly, and those without computer experience or skills

Migration to Electronic Format

- Where instrument is validated in paper format, electronic version should be equivalent to paper
- Changes in layout of questionnaire may be needed
- Could the changes made affect responses?

Changes in Migration

Level	Type of change	Action
Minor	No change in context or meaning	Cognitive debriefing
	Justified by existing literature	
Moderate	Changes in layout or wording that could affect interpretation	Equivalence testing
Substantial	Changes that clearly affect context or meaning	Psychometric Validation

Based on Shields et al. (2006) and Coons et al. (2008)

Examples of Minor Changes

- Tap on a screen
- Press a button

Rather than tick or circle a response on paper

- Change in length of visual analogue scale

Question Layout (SF-36)

3. The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

		Yes, limited a lot	Yes, limited a little	No, not limited at all
a	Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports	☐ 1	☐ 2	☐ 3
b	Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	☐ 1	☐ 2	☐ 3
c	Lifting or carrying groceries	☐ 1	☐ 2	☐ 3
d	Climbing several flights of stairs	☐ 1	☐ 2	☐ 3
e	Climbing one flight of stairs	☐ 1	☐ 2	☐ 3

Question Layout (SF-36)

3. The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports

- Yes, limited a lot
- Yes, limited a little
- No, not limited at all

Question Layout (SF-36)

3. The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf

- Yes, limited a lot
- Yes, limited a little
- No, not limited at all

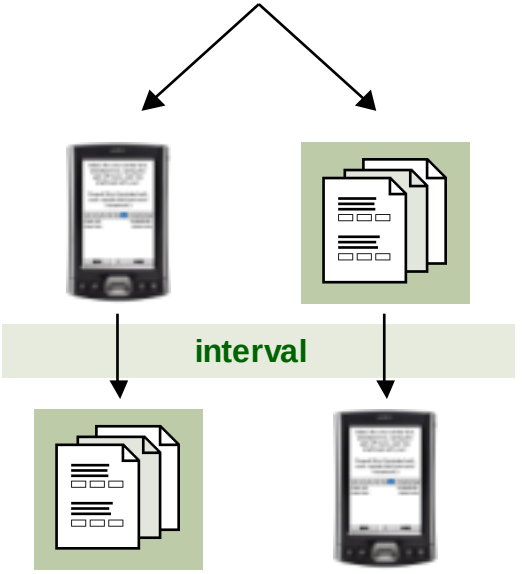
Evaluating Changes

- Existing Data (see e.g. Gwaltney et al. (2008)
  - Computer implementation *per se* not an issue
  - VAS unaffected by size over a wide range.
- Cognitive Interviewing (Ask the Patient)
  - Do patients construe questions in the same way?
  - Do they remember instructions as intended?
- Equivalence Testing may be required if changes are more than minor

Equivalence and Reliability

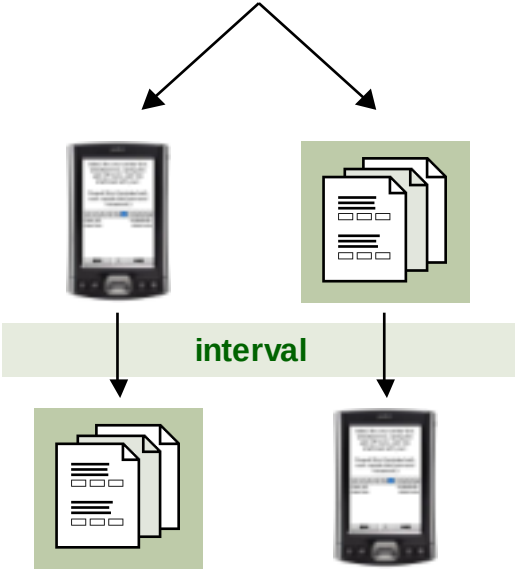
- Large literature on reliability of clinical assessments concerning either:
  - Test-retest reliability
  - Inter-rater reliability
- Principles and methods are closely similar for equivalence comparisons between two modes, such as paper and electronic.

Equivalence Study Design

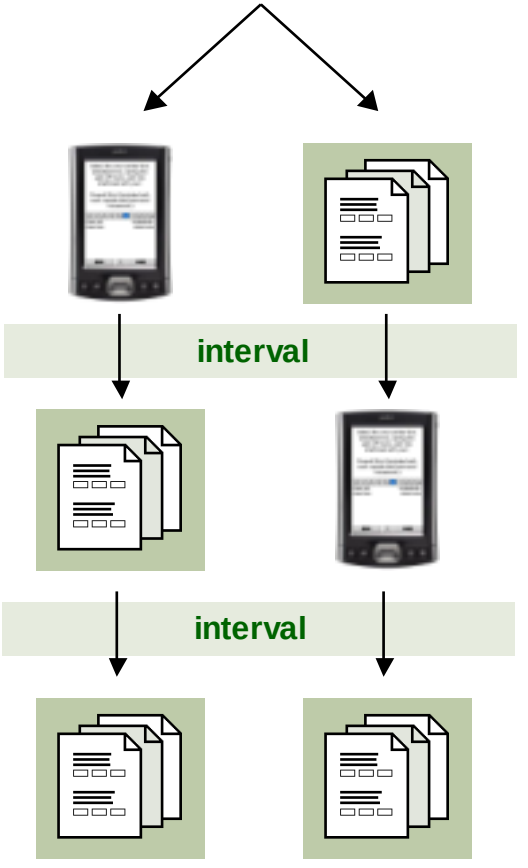


- Interval should be long enough that patients will not remember exact entries.

Equivalence Study Design



- How does reliability compare to test-retest reliability of existing instrument?



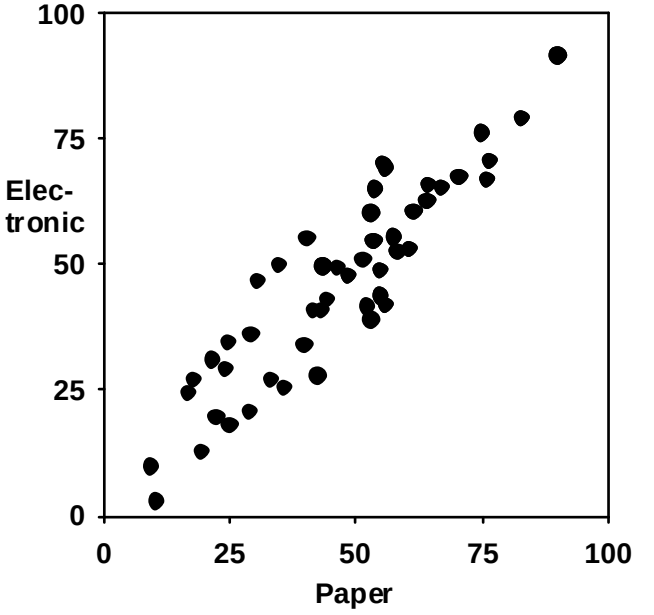
- Three-period approach allows P-E reliability to be compared to P-P

Quantifying Equivalence

- Correlational approach
  - Intraclass correlation coefficient (ICC) is preferred to Spearman's r.
  - Value from 0 (no agreement) to 1 (perfect agreement
- Numerical agreement approach
  - Based on precision of measurement
  - Agreement indicated by size and variance of Electronic-Paper differences in scale measure.

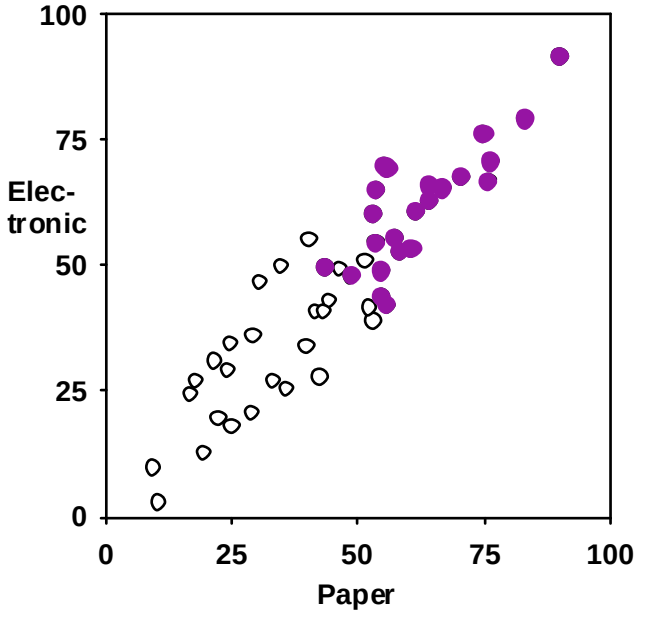
Intraclass Correlation

ICC = 0.90



ICC and range

ICC = 0.75

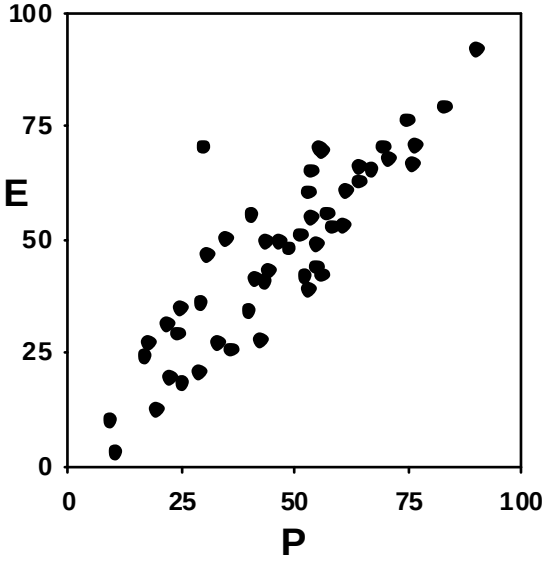


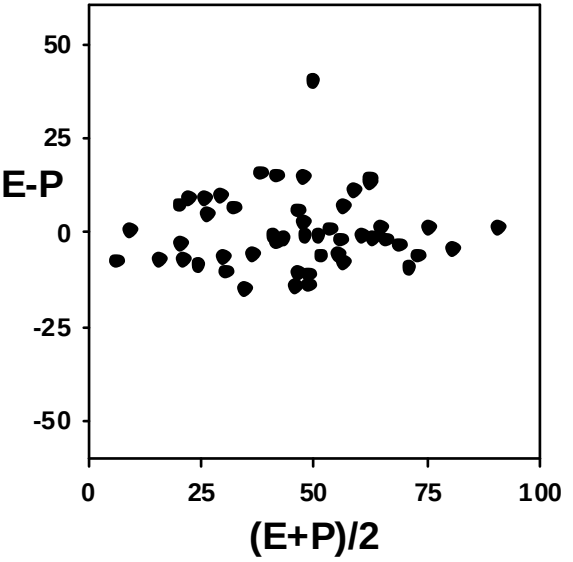
ICC and Range

- Correlation-based measures such as ICC are very dependent on range of values in sample
- ICC will increase if variability is high
  - Wide range of severity in sample
- ICC will be lower if variability is restricted
  - Normal population vs clinical group
  - Inclusion criteria specifying minimum severity

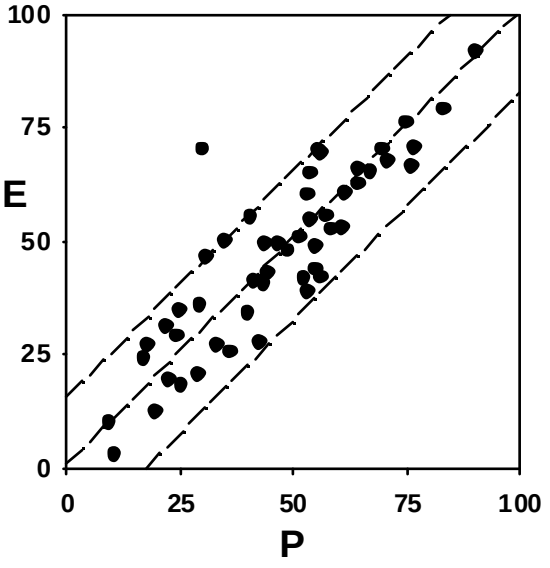
Other measures should be used as well as ICC

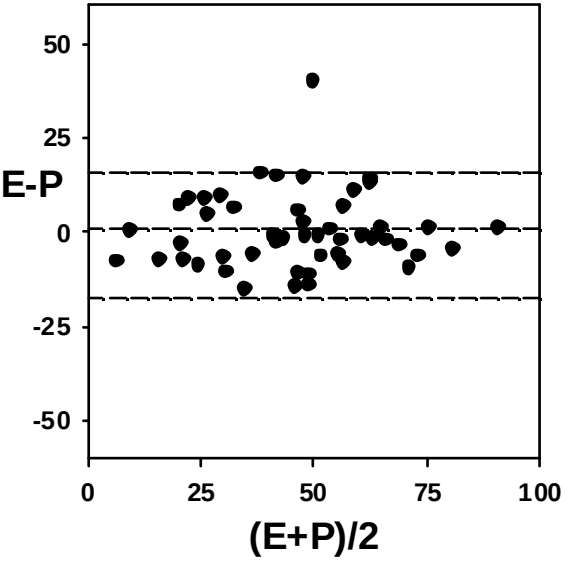
Bland-Altman Plot





Bland-Altman Plot





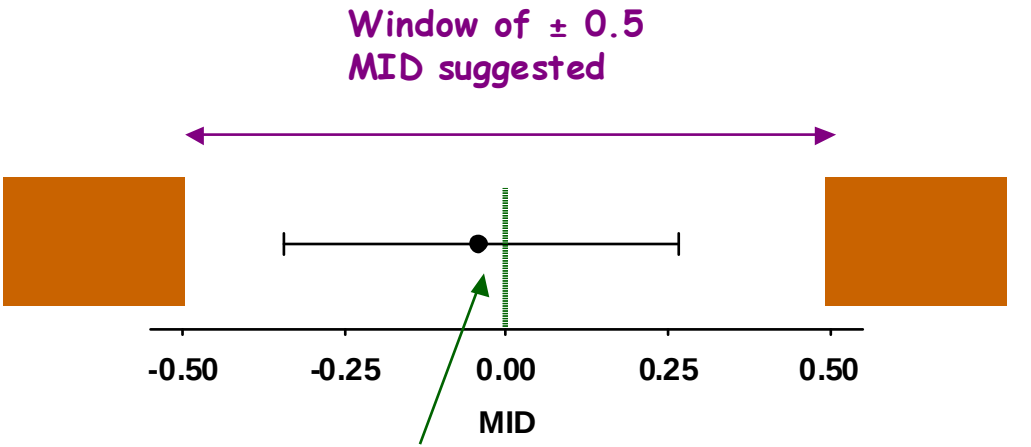
Differences between Modes

- Simple E - P difference on its own is not very helpful. It needs to be compared to something
  - Scale Length (usually as %)
  - Variability (“Effect Size”, ES)
  - Minimum Important Difference (MID)
- When expressed in one of these ways, differences can be compared between scales

The Bioequivalence Model

- The observed difference between modes and its confidence interval must be completely contained within a specified range around zero
- The range can be specified in terms of one of the comparative measures: MID, ES, or % scale.

Bioequivalence Model



CI of mean should be completely within the window

Mean difference should be close to zero

Difference between modes is expressed in terms of MID

The Bioequivalence Model

- Allows degree of equivalence between modes to be expressed in terms of the sizes of differences we wish to demonstrate in a clinical trial
- Is not affected by the variability of the sample used in assessment
- Provides useful basis for power calculations

Power and Sample Size

There is no null hypothesis in the usual sense. Several criteria can be used:

- ICC is significantly greater than a given value, say 0.75
- Precision of ICC is within stated limits, say ± 0.1
- CI of differences between modes is contained within defined window

Power and Sample Size

The bottom line:

- 40 – 50 is usually a sensible sample size

Equivalence Data on SF-36 (Aggregate Physical Factor)

Two-period crossover in Rheumatoid Arthritis

- ICC = 0.92
- Very good agreement
- Changes to layout had no significant effect on data obtained
- Mean E-P difference is very small: Effect size about 0.04 (ES of <0.2 considered unimportant)
- Confidence intervals are well within bioequivalence target window

Tiplady et al., in preparation

Conclusions

- Research data supports validity of electronic diaries and questionnaires
- Where specific data are required to demonstrate validity of a specific scale, established reliability methods are appropriate
- Very close agreement between electronic and paper scales can be achieved by paying attention to the fundamental construct being measured