Accounting For Rater Severity/Leniency In Endpoint Measures

An Example in Neurobehavioral Functioning of Adults with Severe TBI

Trudy Mallinson, PhD, OTR/L
Visiting Associate Professor

Office for Clinical Practice Innovation &
Department of Clinical Research and Leadership
Collaborators

- Theresa Pape, DrPH, MA, CCC-SLP/L
  - Center for Innovation in Complex Chronic Healthcare, Edward Hines VA Hospital, Hines, IL

- Ann Guernon, MS, CCC-SLP/L, CCRC
  - Marianjoy Rehabilitation Hospital, Wheaton, IL

Study funded by Department of Veterans Affairs, Office of Research and Development, Health Services Research and Development Service Merit Grant # CCN-07-133 to T. Pape.
Endpoints

● Per the FDA, an endpoint is a direct measure of:
  – Improved survival
  – Detectable benefit to the patient (e.g. symptoms or functional capacity)
  – Decreased chance of development a condition
  – May be a surrogate (e.g. biomarkers)
Endpoints and the FDA

- Assess the benefits of treatment (i.e., medication, devices)
  - Endpoints include survival, biomarkers, and clinical outcomes

- Clinical Outcome Assessments (COAs)
  - Patient-reported outcomes (PROs)
  - Clinician-reported outcomes (ClinROs)
  - Observer-reported outcomes (ObsROs)
  - Performance outcomes (PerfOs)

- The primary goal of FDA measurement development is to support labeling claims

- FDA has clear guidance for developing PROs as endpoints but is developing this guidance for ClinROs and ObsROs
Relevance of Endpoints for Rehabilitation

- ClinROs are currently a major part of drug efficacy endpoints

- ClinROs are often relevant in the rehabilitation-setting because:
  - Patients may be unable to self-report (e.g. altered states of consciousness)
  - Clinician observation adds important dimension (e.g. understanding altered functional capacity after a traumatic event)

- FDA guidance on ClinROs has some direct relevance to rehabilitation (for medications such as Amantadine or devices such as FES)

- But also relevant for improving the interpretability and quality of rehabilitation-focused clinical trials
Rater severity/leniency affects ALL rater-mediated clinical outcome assessments (COAs)
- Including clinician-reported and observer-reported outcomes (ClinROs and ObsROs).
Defining Rater Severity/Leniency

- Rater severity/leniency should *not* be confused with “interrater reliability.”

“The consistent tendency on the part of the rater to give a score that is higher or lower than appropriate, which is usually interpreted to mean higher or lower than the average of the other raters.”

~ (Wilson & Case, 2000)
Why rater severity/leniency matters

● Applies to any situation where clinician observe and then make a judgement about patient “performance”

● Problem is more acute when different clinicians rate the same patient at different points in time
  – For example, monitoring recovery in patients with severe TBI involves repeated measures, often by multiple different clinicians

● These scores will be inaccurate to the extent that different clinicians rate the same patient more severely or leniently

● It reflects a clinician’s worldview, training is generally ineffective in removing severity/leniency (Eckes, 2008)
When is change real clinical change?

A typical run of rehabilitation outcome measures

Unadjusted measure
● Any factor that contributes to making a person appear to have more or less function than they actually do.

● Examples of common facets include:
  – Rater severity/leniency
  – Test stimuli/task difficulty
  – Time period
Multi-Faceted Rasch Measurement

- **MFRM**, an extension of the normal 1-parameter Rasch model, uses the principal of conjoint additivity to essentially remove the effect of rater severity/leniency from person measures (Linacre, et al., 1994).

\[
\log \left( \frac{P}{1-P} \right) = B_n - D_i
\]
A facet is an additive contribution to the log odds of observing one category (vs. the next lowest)

\[ \log \left( \frac{P}{1-P} \right) = B_n - D_i - C_j \]
And so on ...

\[ \log \left( \frac{P}{1-P} \right) = B_n - (D_i + C_j + T_k \ldots) \]
To examine the impact of rater severity/leniency on measures of neurobehavioral functioning (NBF) derived from the Disorders of Consciousness Scale (DOCS) using the Multi-Faceted Rasch Model (MFRM). Change in NBF is a clinically-meaningful endpoint for patients with severe TBI.

**HYPOTHESIS:** That raters will be sufficiently uniform in severity/leniency such that a patient’s Disorders of Consciousness Scale (DOCS) measure will not require adjustment for rater variation.
**Study Design**

- **Prospective, observational cohort study** of 172 patients at 7 post-acute rehabilitation facilities.

- **Participants:** Patients with severe TBI, classified as vegetative or minimally conscious at enrollment, <180 days post-injury.

- **Raters:** 48 rehabilitation clinicians trained to score the DOCS (occupational therapists n=13, physical therapists n=8, speech-language pathologists n=20, other disciplines n=7).

- **Outcome Measure:** Behaviors elicited by 25 DOCS sensory stimuli.
Disorders of Consciousness Scale

• The DOCS involves administering 25 items that include auditory-verbal, visual, somatosensory, and gustatory/olfactory sensory stimuli (Pape, Mallinson, & Guernon, 2014).

• Clinicians observe patients and record the best responses to the stimuli.

• Responses are scored on a 3-point rating scale:
  • 0 = No Response (no apparent motor response to stimulus)
  • 1 = Generalized Response (motor response not related to stimulus e.g. moves foot in response to juice)
  • 2 = Localized Response (motor response appropriate to context of stimulus e.g. licks lips in response to juice)
Need to have some paired ratings in order for the analysis to appropriately determine amount of rater severity/leniency.

Does not require complete overlap. Just cannot have single patient/rater dyads.
**Data Analysis:** Data were analyzed with and without adjustment for rater severity and the results examined for differences greater than the established minimally detectable change (MDC), 5 units. On 137 (20%) occasions, a pair of clinicians jointly observed but separately scored the patient responses, creating linkage across all 690 observations.

**MFRM:** Facets® software, which adjusts for rater severity.

**2-facet:** Winsteps® software; no adjustment for rater severity
Principal Findings

Fig. 1. Most raters are severe/lentient within an acceptable range (± 5 units = MDC₉₅). However, 241/690 (35%) of patient measures exceeded the MDC after adjustment; 183 under-estimated, 58 overestimated.
DOCS measures were significantly different after adjusting for rater severity/leniency.

- Mean unadjusted measure 50.8 ± 0.9 units; mean adjusted measure 51.7 units ± 1.0 units; t=-2.25, P=0.03

35% of DOCS measures differed more than the established minimally detectable change (MDC95) of 5 units, after adjusting for rater severity/leniency.

- 28 patients were rated too severely:
  - median -6.7 units; IQR -9.9 – -5.6
- 13 were patients rated too leniently:
  - median 9.15 units; IQR 7.0 – 14.0
Implications of Findings

Fig. 2. Blue line shows adjusted patient measures ± minimally detectable change (MDC). At weeks 3 & 5, adjusted measures exceed the MDC. At both times, patient would be classified as minimally conscious by the unadjusted measure when they are more likely in a vegetative state.
Implications of Findings

Fig. 3. Blue line shows adjusted patient measures ± minimally detectable change (MDC). By unadjusted measures, the patient appears to improve week 1 to week 2 but is actually stable. Without adjustment, clinicians could be misled about the possible effect of medications.
Conclusions

- Rater severity/leniency affects ALL rater-mediated COAs including ClinROs and ObsROs (Eckes, 2009) and has significant potential to misrepresent a drug’s effect.

- MFRM quantified the impact of rater severity/leniency on DOCS measures. About one third of DOCS measures changed by more than 1 MDC after adjustment.

- Common interrater reliability coefficients e.g., ICCs, are really only useful as a population parameter since they cannot be used to adjust individual patient scores for a particular rater’s judgments (Stemler, 2004).

- MFRM removes the effect of rater severity/leniency on patient measures if data are appropriately collected (sufficient linkage).
To qualify as a COA and to be useful as an endpoint for adequate and well-controlled clinical trials providing substantial evidence of drug effectiveness, rater-mediated measures must account for rater severity/leniency.
Questions