When Do We Have Enough Evidence to Accept Migration of Patient-Reported Outcome Measures (PROMS) from Paper to Screen-based Formats without Additional Testing?

ISPOR European Congress 2018

Barcelona, Spain

November 14, 2018

Disclaimer

• The views and opinions expressed in the following slides are those of the individual presenters and should not be attributed to their respective organizations/companies or the Critical Path Institute.

• These slides are the intellectual property of the individual presenters and are protected under the copyright laws of the United States of America and other countries. Used by permission. All rights reserved. All trademarks are the property of their respective owners.
Presenters

– Bill Byrom, PhD
  – Vice President of Product Strategy and Innovation, CRF Bracket

– Sonya Eremenco, MA
  – Associate Director, Patient-Reported Outcome (PRO) Consortium, Critical Path Institute

– Willie Muehlhausen, DVM
  – Managing Director, Muehlhausen Ltd

Presentation Outline and Objectives

When Do We Have Enough Evidence to Accept Migration of Patient-Reported Outcome Measures (PROMS) From Paper to Screen-based Formats without Additional Testing?

• Introduction
• Presentation of key works
• Faithful migration best practices
• Discussion: case examples
Introduction

Sonya Eremenco, PRO Consortium, Critical Path Institute

Set the scene regarding electronic migration

• How did we get here?
• What are the current recommendations for evaluating equivalence?
• Where do we go next?
**Brief history of migration/equivalence recommendations**

- **2006**
  - FDA publishes Draft Guidance on PRO Measures in February
  - Changing mode is considered a modification of the instrument – validation may be necessary

- **2008**
  - ISPOR ePRO Task Force publishes recommendations for establishing measurement equivalence in November 2008 online

- **2009**
  - FDA publishes Final PRO Guidance in December
  - Electronic migration still considered a modification
  - Small non-randomized studies may be sufficient

- **2010**
  - ISPOR Task Force on Mixed Modes of PRO Data Collection convened

- **2014**
  - ISPOR Task Force Report on Mixed Modes of Data Collection published

---

**ISPOR ePRO Task Force Report Recommendations**

Table 1: PRO to ePRO measurement equivalence: instrument modification and supporting evidence

<table>
<thead>
<tr>
<th>Level of modification</th>
<th>Rationale</th>
<th>Examples</th>
<th>Level of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minor</td>
<td>The modification can be justified on the basis of logic or existing literature. No change in content or meaning.</td>
<td>1) Nonsubstantive changes in instructions (e.g., from circling the response to touching the response on a screen).</td>
<td>Cognitive debriefing, usability testing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Minor changes in format (e.g., one item per screen rather than multiple items on a page).</td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>Based on the current empirical literature, the modification cannot be justified as minor. May change content or meaning.</td>
<td>1) Changes in item wording or more significant changes in presentation that might alter interpretability.</td>
<td>Equivalence testing, usability testing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Change in mode of administration involving different cognitive processes (e.g., paper [visual] to IVR [aural]).</td>
<td></td>
</tr>
<tr>
<td>Substantial</td>
<td>There is no existing empirical support for the equivalence of the modification and the modification clearly changes content or meaning.</td>
<td>1) Substantial changes in item response options</td>
<td>Full psychometric testing, usability testing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Substantial changes in item wording</td>
<td></td>
</tr>
</tbody>
</table>

Adapted from Shields et al. [62].

ISPOR Mixed Modes Task Force Recommendations

1. Select appropriate mode(s) for trial
2. Perform a “faithful migration” ("migrate before you mix")
   - Only necessary changes to the format and instructions are made and that the content of the items and responses has not changed.
   - Subjects interpret and respond to the questions/items the same way regardless of mode
3. Evaluate equivalence between the modes migrated and/or to be mixed
   - Use appropriate study design
4. If above conditions are met, implement the mode or modes in the trial
   - Avoid mixing paper and electronic diaries; assess risks of other combinations
   - If deciding to mix other modes
     - Plan and implement carefully; mix at country level or higher
     - Assess statistical issues and poolability of data


Need to Establish Measurement Equivalence

Will PRO items be used for regulatory submission or labeling claim?

- No
- Yes

Is there published evidence of equivalence?

- No
- Yes

What level of change is needed for migration?

- Minor
- Moderate

Document for later use in regulatory submission

Perform Cognitive Interviewing

Perform Equivalence Study

PRO, patient-reported outcome
Qualitative Study Design: Cognitive Interview

- **Purpose:** to evaluate if the migration has impacted how subjects *interpret and respond* to the items
  - Not intended to revisit content validity of the original instrument
- **Minor modifications to format or procedure**
- **Small sample size:** 5 to 10 subjects
- **Assess usability of instrument as a secondary goal**

Variations in study design include:
- Whether patients complete both modes during interview
- How responses can be compared
- Whether multiple interview rounds are necessary to allow for revising/retesting

Key questions answered:
- Why interpretation between modes may differ
- Why responses between modes may differ

“New” Literature on Equivalence

- **EuroQol 5-Dimension questionnaire (EQ-5D): IVR and Paper**

- **EORTC: IVR and Paper**

- **PROMIS Physical Function, Fatigue, Depression banks: personal computer (PC) vs. IVR, personal digital assistant (PDA), Paper, or PC**

- **Reviews of paper vs. electronic studies**

- **Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE): Web, IVR and Paper**

- **Bowel function instrument, linear analog scale assessment (LASA) quality-of-life (QOL) and Adapted Sydney Swallow Questionnaire (SSQ): Web, IVR and Paper**

- **Bring your own device (BYOD)**
Where Do We Go Next?

• What has changed since 2014?
  • Hundreds of unpublished qualitative migration studies conducted confirming equivalence
  • Usability issues are the more salient results

• Bring Your Own Device (BYOD) is becoming mainstream
  • Mixing is inherent in BYOD implementations
  • Not feasible to conduct equivalence studies among all possible devices

• Industry views “equivalence studies” as a requirement when implementing clinical outcome assessments electronically because of regulatory uncertainty

• A new ISPOR ePRO Task Force will update the previous recommendations
  • Outline the evidence required to ensure a faithful migration and suggest when, in light of the accumulated evidence, additional testing is not required
  • Identify aspects of instrument migration or study design that may jeopardize compatibility between modes or impact the operational integrity of the study
  • Explore the role of feasibility testing

Key Works

Bill Byrom, CRF Bracket
Growing body of evidence

- Equivalence study meta analysis (Muehlhausen) 2015
- Meta-synthesis of cognitive interview studies (Muehlhausen, Byrom) 2017
- BYOD equivalence study (ePRO/PRO Consortium) 2016-18

2008
- Equivalence study meta analysis (Gwaltney)

2016
- BYOD attitudes and opinions: industry survey (Byrom, Muehlhausen)

2018
- BYOD equivalence study (Byrom, Muehlhausen)

Published
In progress

Meta-analysis: Gwaltney et al. 2008

Gwaltney et al. (2008)
- 46 studies
- 278 PROM comparisons
- Paper vs. PC/handheld device
- Study n: 10 – 189
- Pooled correlation coefficient: 0.90 (95% CI: 0.87–0.92)
Meta-analysis: Muehlhausen et al. 2015

Muehlhausen et al. (2015)
- 72 studies
- 152 PROM comparisons
- Paper vs. PC, handheld device, IVRS
- 23 patient populations
- Ages: 6 – 68 years
- Pooled correlation coefficient: 0.875 (95% CI: 0.867 to 0.884)

“PROMs administered on paper are quantitatively comparable with measures administered on an electronic device” across multiple scales and patient groups.”

Meta synthesis of cognitive interview and usability studies

- Muehlhausen, Byrom, Skerritt et al. (2017)
  - All studies conducted by ICON from 2012 to 2015
  - 53 studies
  - Wide range of patient populations including:
    - Respiratory, Gastrointestinal, Oncology, Central Nervous System disorders, Rheumatology, Cardiovascular disorders, Dermatology, Gynaecology, Infectious disease, Metabolic, Urology, Vaccines.
  - 68 instruments
  - 101 PROM comparisons
  - Response scale types included: visual analogue scale (VAS), verbal rating scale (VRS), numeric rating scale (NRS), EQ-VAS (from EQ-5D)
Meta synthesis of cognitive interview and usability studies

"With the benefit of accumulating evidence, it is possible to relax the need to routinely conduct cognitive interview and usability studies when implementing minor changes during instrument migration. Application of design best practice and selecting vendor solutions with good user interface and user experience properties that have been assessed for usability in a representative group may enable many instrument migrations to be accepted without formal validation studies by instead conducting a structured expert screen review."

Equivalence with variable screen size (BYOD)

- 156 subjects
  - 19 to 69 years old (48.6 ± 13.1)
  - Female: 83 (54%)
  - Male: 72 (46%)
  - Conditions resulting in chronic pain
  - Broad range of educational backgrounds
- SF-20
  - VRS
  - Y>3, Y<3, N
  - Likert
- VAS and NRS-11 (pain)
### Equivalence with variable screen size (BYOD)

<table>
<thead>
<tr>
<th>Paper</th>
<th>iOS</th>
<th>Android</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
</tr>
</tbody>
</table>

- **Very high correlation between the three modes of administration:**
  - **ICCs:** 0.816 to 0.974
  - **Lower bound of the 95% confidence interval:** > 0.70

- **Very high correlation between paper and BYOD**
  - **ICCs:** 0.806 to 0.974
  - **Lower bound of the 95% confidence interval:** > 0.70

- **Very high correlation between site device and BYOD**
  - **ICCs:** 0.791 to 0.966
  - **Lower bound of the 95% confidence interval:** > 0.70

- **Very high correlation between the three modes of administration for each response scale type:**
  - **VRS:** ICC: 0.97 (0.96 – 0.98)
  - **NRS:** ICC: 0.98 (0.97 – 0.98)
  - **VAS:** ICC: 0.94 (0.91 – 0.95)
Good summary of all the evidence

Therapeutic Innovation and Regulatory Science.
https://doi.org/10.1177/2168479018793369

Thoughts:
Faithful Migration Best Practices

Willie Muehlhausen, Muehlhausen Ltd.
Why are we doing this?

- Clinical Trials are including more patient input (Patient-centric)
- Virtual Clinical trials
- Real World Evidence
- Patient Care and Remote Monitoring
- Bring Your Own Device (BYOD)
Instrument “Controls” / “Widgets”

• Most instruments are composed of a small number of controls/widgets
  
  • Numeric Rating Scale (NRS)

  • Visual Analogue Scale (VAS)

  • Verbal Rating Scale (VRS)

Definition:

• A graphical control element or widget is an element of interaction in a graphical user interface (GUI), such as a button or a scroll bar.

• A “Control” or “Widget” is an interface element (e.g., numeric rating scale)
Instrument “Controls” / “Widgets”

- Can we validate a control/widget independent of the context it applies to?
  - Example: numeric rating scale components

![Numeric Rating Scale](image)

- Can we appeal to previous validation of this widget when considering the need to perform future equivalence testing?

Meta synthesis of cognitive interview and usability studies

- Muehlhausen, Byrom, Skerritt et al. (2017)
  - All studies conducted by ICON from 2012 to 2015
  - 53 studies
  - Wide range of patient populations including:
    - Respiratory, GI, Oncology, CNS, Rheumatology, CV, Dermatology, Gynaecology, Infectious disease, Metabolic, Urology, Vaccines.
  - 68 instruments
  - 101 PROM comparisons
  - Response scale types included: VAS, VRS, NRS, EQ-VAS
Equivalence with variable screen size (BYOD)

- Very high correlation between the three modes of administration:
  - ICCs: 0.816 to 0.974
  - Lower bound of the 95% confidence interval > 0.70
- Very high correlation between paper and BYOD
  - ICCs: 0.806 to 0.974
  - Lower bound of the 95% confidence interval > 0.70
- Very high correlation between site device and BYOD
  - ICCs: 0.791 to 0.966
  - Lower bound of the 95% confidence interval > 0.70

- Very high correlation between the three modes of administration for each response scale type:
  - VRS: ICC: 0.97 (0.96 – 0.98)
  - NRS: ICC: 0.98 (0.97 – 0.98)
  - VAS: ICC: 0.94 (0.91 – 0.95)

Best Practices
Recommendations

- Keep it simple!!
- Text Art only when proven beneficial (i.e., bold, italic, underline, capitals)
- Colour only when proven beneficial
- One item per screen (even if there is space for more on the tablet)
- Use of neutral verbiage during development (i.e., “Select” vs. “Circle”)
- Use basic widgets and avoid creative combinations

Then:
- Conduct an Expert Screen Review and possibly Usability Testing only

<table>
<thead>
<tr>
<th>ePRO Design Best Practice *</th>
<th>Usability</th>
<th>Expert Screen Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Provide robust instructions on use of the application.</td>
<td>- Usability should cover the app and all common widgets</td>
<td>1. Overall instructional information</td>
</tr>
<tr>
<td>- Ensure font size is suitable, clear, and readable.</td>
<td>- Usability evidence from representative groups is sufficient</td>
<td>- Instrument and application instructions</td>
</tr>
<tr>
<td>- Present a single question and response scale option per screen.</td>
<td>- Patients or healthy volunteers</td>
<td>- Recall period representation</td>
</tr>
<tr>
<td>- Take care not to modify the original instrument text beyond minor changes.</td>
<td>- Similar age range to target population, range of educational and socioeconomic backgrounds</td>
<td>- Author-specific requirements</td>
</tr>
<tr>
<td>- Precede question with instructional text screen if cannot be presented together.</td>
<td>- Additional representative groups may include (as needed)</td>
<td>2. Usability, including font size and navigation</td>
</tr>
<tr>
<td>- Ensure equal screen area, font and line spacing used for each response option.</td>
<td>- children/adolescents,</td>
<td>- Clarity, colour and font size per screen</td>
</tr>
<tr>
<td>- Use indicator arrows to identify the location of anchor text if needed.</td>
<td>- dexterity-challenged subjects</td>
<td>- Consistency, visibility and size of controls</td>
</tr>
<tr>
<td>- Present the recall period with each item as opposed to only in initial instrument instructions.</td>
<td>- technology-naive subjects (e.g., very elderly subjects)</td>
<td>- Device-orientation changes</td>
</tr>
<tr>
<td>-</td>
<td>- cognitively challenged subjects</td>
<td>- Back and forward navigation</td>
</tr>
<tr>
<td>-</td>
<td>- partially sighted subjects.</td>
<td>- End of questionnaire review (where included)</td>
</tr>
</tbody>
</table>

* Consolidated from Critical Path Institute’s ePRO Consortium Recommendations and ICON research.
Next steps

• Request for help:
  • Screenshots of instruments in published/unpublished
  • Equivalence studies
  • Cognitive interview/usability studies
  • Expert Screen Review

• Share Experience with Regulators
  • Bring Your Own Device

• Distribute the C-Path Best Practice documents and use them!
  • c-path.org/programs/epr
Example 1: SF-36

- Current evidence
  - 7 studies within Gwaltney et al. 2008
  - 9 studies within Muehlhausen et al. 2015
  - 3 studies in meta-synthesis, 2018
  - Equivalence of VRS (BYOD study, 2018)
  - Meta-analysis of 25 studies: SF-36 only (White et al., 2018)
  - Unpublished CI/UT studies on various vendor platforms

1. Is there enough evidence to not require additional testing?

2. If so, what conditions would be required for this to be the case?

3. How would we package the evidence to support migration comparability?

Example 2: Instrument in different population

- Current evidence
  - Demonstrated equivalence in population 1

- Required evidence
  - Evidence to support measurement equivalence in population 2

1. Examples of populations that would not require additional evidence

2. Examples of populations that would require additional evidence
Example 3: Visual analogue scale
Baseline

- Keep it simple and consistent
- Spot the difference?!
How do you think I feel about these?

Example 4: New instrument w/without standard widgets

Examples:

1. What evidence would be needed to demonstrate migration equivalence?

2. How could this evidence be generated and reported to enable its re-use to support other studies?
Example 5: Apple Research Kit or non-standard response scale types

1. What evidence would be needed to demonstrate migration equivalence?

2. How could this evidence be generated and reported to enable its re-use to support other studies?

Example 6: New vendor platform

- What you would need to do if it was completely a new platform
- Usability: screen review vs. data entry
- Confirm best practices are being followed
- Button vs. selecting answer
Thank you