Measurement Properties of Total Ankle Replacement Outcome Measures: A Systematic Review

Ellie Pinsker, BA&Sc, PhD candidate
Taucha Inrig, BScN, RN, MDiv
Phillip Daniels, BA
Timothy R. Daniels, MD, FRCSC
Dorcas Beaton, BScOT, MSc, PhD
Disclosures

No conflict to disclose: EP, TI, PD, DB

Consultant/financial support: TD (Wright Medical, Carticept, Integra, Stryker)

Our disclosures are in the Final AOFAS Mobile App.

We have no potential conflicts with this presentation.
Introduction

- There are a wide range of outcome measures used to evaluate total ankle replacement (TAR). (Naal et al, 2010)
  - Valued for their ability to assess the effect of TAR and compare surgical techniques related to TAR or alternative interventions

- Researchers and clinicians selecting an instrument to evaluate TAR should base their decision on adequate evidence supporting its measurement properties (incl. validity, reliability, and responsiveness).

- However, no measure is simply valid or reliable; it is valid or reliable in a specific context or population. This evidence must be generated in all populations in which it is applied. (Beaton and Schemitsch, 2003; de Vet et al, 2011)

- **Purpose**: This review identifies region- or joint-specific outcome measures used for evaluating TAR outcomes and synthesizes evidence for their measurement properties.
Methods

- **Literature search:** A standard search strategy was conducted of electronic databases MEDLINE, EMBASE and CINAHL (to June 2015) to identify foot/ankle measures in use.
  - Ex. MEDLINE search terms: (1) “exp Arthroplasty, Replacement, Ankle/” OR “ankle arthritis”.mp; (2) “exp Psychometrics/” OR reliability OR responsiveness OR validity OR reproducibility of results OR interpretability OR MCID OR MDC”

- **Selection criteria:**
  1. Study reported on development and/or evaluation of at least one measurement property of a region-/joint-specific outcome measure tested in a TAR population;
  2. A full-text original article (i.e. not an abstract, conference proceeding, editorial or review);
  3. Published in English.

- Two reviewers (EP, PD) independently screened all titles and abstracts. Potentially relevant articles moved forward to full article review.

- Measurement properties covered in each article were tallied: validity, reliability, responsiveness, or interpretability.

- **Quality assessment:** Methodological quality of studies reporting on a measurement property rated (excellent, good, fair, poor) according to COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN). (Mokkink et al, 2006; 2010a; 2010b)
Methods

- Two reviewers (EP, TI) independently rated each study with all disagreements resolved through discussion.

**Best evidence synthesis: levels of evidence:**

- Approach considers the methodological quality of the study, quantity of studies evaluating the same measurement property and consistency of their results.
- Level of evidence (Table 1) is accompanied with overall rating for a measurement property: positive (+), indeterminate (?) or negative (-).
- Thus, a strong or moderate positive rating would indicate adequate evidence supporting the validity, reliability, or responsiveness of a measure.

**Table 1. Levels of evidence for the overall quality of the measurement property** (based on van Tulder et al, 2003)

<table>
<thead>
<tr>
<th>Level</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality</td>
</tr>
<tr>
<td>Moderate</td>
<td>Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality</td>
</tr>
<tr>
<td>Limited</td>
<td>One study of fair methodological quality</td>
</tr>
<tr>
<td>Conflicting</td>
<td>Conflicting findings</td>
</tr>
<tr>
<td>Unknown</td>
<td>Only studies of poor methodological quality</td>
</tr>
</tbody>
</table>
Results

Five measures were identified with psychometric evidence in a TAR population:

- American Orthopaedic Foot & Ankle Society Ankle Hindfoot Scale (AOFAS AHS) (Kitaoka et al, 1994)
- Foot Function Index (FFI) (Budiman-Mak et al, 1991)
- Ankle Osteoarthritis Scale (AOS) (Domsic and Saltzman, 1998)
- Foot and Ankle Outcome Score (FAOS) (Roos et al, 2001)
- Self-Reported Foot and Ankle Score (SEFAS) (Cöster et al, 2012)

Data on measurement properties drawn from six studies (Pena et al, 2007; Cöster et al, 2012; Madeley et al, 2012; Coe et al, 2015; Pinsker et al, 2015; Mani et al, 2015)
<table>
<thead>
<tr>
<th>Tested Measures</th>
<th>Studies</th>
<th>Validity Quality</th>
<th>Reliability Quality</th>
<th>Responsiveness Quality</th>
<th>Interpretability Reported</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Content</td>
<td>Construct</td>
<td>Internal Consistency**</td>
<td>Test-Retest</td>
</tr>
<tr>
<td>AOFAS</td>
<td>Madeley et al., 2012*</td>
<td>Fair (+)</td>
<td></td>
<td></td>
<td>Fair (+)</td>
</tr>
<tr>
<td></td>
<td>Pena et al., 2007</td>
<td>Poor</td>
<td></td>
<td></td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td>Pinsker et al., 2015*</td>
<td>Good (+)</td>
<td>Good (+)</td>
<td>Good (+)</td>
<td></td>
</tr>
<tr>
<td>AOS</td>
<td>Coe et al., 2015</td>
<td>Fair (+)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Madeley et al., 2012</td>
<td>Fair (+)***</td>
<td></td>
<td></td>
<td>Fair (+)</td>
</tr>
<tr>
<td></td>
<td>Pinsker et al., 2015</td>
<td>Good (+)</td>
<td>Good (+)</td>
<td>Good (+)</td>
<td></td>
</tr>
<tr>
<td>FAOS</td>
<td>Côster et al., 2012</td>
<td>Good (+)</td>
<td></td>
<td></td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td>Mani et al., 2015</td>
<td>Fair (+)</td>
<td>Fair (+/-)</td>
<td>Fair (+)</td>
<td>Poor</td>
</tr>
<tr>
<td>FFI</td>
<td>Madeley et al., 2012</td>
<td>Fair (+)***</td>
<td></td>
<td></td>
<td>Fair (+)****</td>
</tr>
<tr>
<td></td>
<td>Pinsker et al., 2015</td>
<td>Good (+)</td>
<td>Good (+)</td>
<td>Good (+)</td>
<td></td>
</tr>
<tr>
<td>SEFAS</td>
<td>Côster et al., 2012</td>
<td>Good (+)</td>
<td>Good (+)</td>
<td>Good (+)</td>
<td>Poor</td>
</tr>
</tbody>
</table>

*5 subjective items only; **Factor analysis item not considered; ***Total scores only; ****No assessment of FFI Activity Limitations subscale.
Table 3. Best Evidence Synthesis

<table>
<thead>
<tr>
<th>Outcome Measures</th>
<th>Validity</th>
<th>Reliability</th>
<th>Responsiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Content</td>
<td>Construct (Hypothesis Testing)</td>
<td>Internal Consistency</td>
</tr>
<tr>
<td>AOFAS</td>
<td>Untested</td>
<td>Unknown</td>
<td>Untested</td>
</tr>
<tr>
<td>AOFAS (5 subjective items)</td>
<td>Untested</td>
<td>Moderate (+)</td>
<td>Moderate (+)</td>
</tr>
<tr>
<td>AOS</td>
<td>Untested</td>
<td>Moderate (+)</td>
<td>Moderate (+)</td>
</tr>
<tr>
<td>FAOS</td>
<td>Limited (+)</td>
<td>Limited (?)</td>
<td>Moderate (+)</td>
</tr>
<tr>
<td>FFI</td>
<td>Untested</td>
<td>Moderate (+)</td>
<td>Moderate (+)</td>
</tr>
<tr>
<td>SEFAS</td>
<td>Untested</td>
<td>Moderate (+)</td>
<td>Moderate (+)</td>
</tr>
</tbody>
</table>

Note: +, positive rating; ?, indeterminate rating; -, negative rating. Interpretability is not considered a measurement property, but an important characteristic of a outcome measure.
Results

- Studies rarely evaluate the content validity of a measure.
  - Content validity is degree to which an instrument measures the most relevant and comprehensive aspects of a concept.

- Responsiveness studies offer limited to no evidence (i.e. unknown due to risk of bias, thus not considered in synthesis).
  - Three studies use effect size (ES) and standardized response mean (SRM), which are acceptable measures of responsiveness only in presence of \textit{a priori} defined hypotheses (incl. magnitude and direction of correlations). (de Vet et al, 2011)

- AOS, FFI, and SEFAS have moderate supporting evidence for construct validity, internal consistency, and test-retest reliability.

- Common weaknesses relating to study design, which can introduce bias:
  - Absent/unclear \textit{a priori} defined hypotheses (incl. magnitude and direction) for responsiveness and construct validity.
  - Inadequate sample size
  - Lack of information on measurement properties of comparator instruments and the constructs they measure
  - Lack of information on missing items (i.e. skipped questions) and how they were handled
Discussion

- This review describes the measurement properties of five foot/ankle-specific outcome measures which have been tested in a TAR population.
  - None of the identified measures have evidence on all measurement properties (missing content validity and/or responsiveness).
  - None have “strong” evidence of measurement properties in TAR.
  - Three had moderate strength of evidence supporting their performance (AOS, FFI, SEFAS).

- In particular, content validity of AOFAS, FFI, AOS, and SEFAS should be evaluated in a TAR population.
  - Content validity is considered critically important. (FDA, 2009; Reeve et al, 2012) “Testing other measurement properties will not replace or rectify problems with content validity.” (FDA, 2009)

- Interpretability is not considered a measurement property, but rather an important characteristic of an outcome measure.
  - AOS and SEFAS were the only measures without reported ceiling/floor effects (>15-20% of patients achieve highest/lowest score).
Conclusion

- This review offers a basis for choosing the most appropriate outcome measure for evaluating TAR outcomes.

- Based on current evidence, AOS or SEFAS are among the best options.

- COSMIN is a strict judge of methodological quality but is based on reporting of the study in publication.
  - With time, reporting of measurement studies may improve and come closer to COSMIN standards leading to better quality evidence.

- Additional high quality studies in TAR populations would benefit all identified measures to form a stronger evidence base supporting their use, particularly:
  - Qualitative study to answer the impact of ankle arthritis and reconstructive outcomes that matter to patients.
  - Longitudinal study to further evaluate responsiveness based on COSMIN or instrument development guidelines.
References