Developing a Core Outcome Set for (stage IV) melanoma trials

C.A.C. (Sanna) Prinsen, MSc, PhD
Clinical epidemiologist
VU University Medical Center
Amsterdam, The Netherlands

Annual Cochrane Skin Group Meeting
Dresden, Germany
17-Mar-2015
A ‘core outcome set’ (COS) is a recommendation of **what** should be measured and reported in all clinical trials.

Once COS are defined, it is then important to achieve consensus on **how** these outcomes should be measured.

What = outcome

How = outcome measurement instrument
What is the problem?

There is lack of **standardization in outcome reporting** in melanoma clinical trials.

This hampers the **usefulness of clinical trial evidence** to inform clinicians.

At the cost of the best possible **care** for melanoma patients.
What is the solution?

**Consensus:**
International consensus among relevant stakeholders on the ‘core outcomes’, including recommendations on outcome measurement instruments that can be used to measure these core outcomes

**Recommendation:**
The core outcomes should be measured and reported as a *minimum* in all future melanoma clinical trials

Schmitt *et al.* (2015)
Objective

**Overall aim:**
To develop a multi-disciplinary, consensus-based set of core outcomes, including their relevant outcome measurement instruments, for melanoma clinical trials
<table>
<thead>
<tr>
<th>Name</th>
<th>Background</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanna Prinsen</td>
<td>Clinical epidemiologist, postdoc COSMIN and COMET, member HOME research groups</td>
<td>The Netherlands</td>
</tr>
<tr>
<td>Caroline Terwee</td>
<td>Epidemiologist, co-founder of COSMIN</td>
<td>The Netherlands</td>
</tr>
<tr>
<td>Phyllis Spuls</td>
<td>Dermatologist, Executive Committee HOME</td>
<td>The Netherlands</td>
</tr>
<tr>
<td>Jochen Schmitt</td>
<td>Dermatologist, Executive Committee HOME, co-founder CSG-COUSIN</td>
<td>Germany</td>
</tr>
<tr>
<td>Stefanie Deckert</td>
<td>Scientific researcher, member HOME research group, member VAPAIN</td>
<td>Germany</td>
</tr>
<tr>
<td>Maarten Boers</td>
<td>Rheumatologist, clinical epidemiologist, co-founder OMERACT</td>
<td>The Netherlands</td>
</tr>
<tr>
<td>Marcel Bekkenk</td>
<td>Dermatologist, expertise in melanoma</td>
<td>The Netherlands</td>
</tr>
<tr>
<td>Robert Stern</td>
<td>Dermatologist</td>
<td>USA</td>
</tr>
<tr>
<td>Alexander van Akkooi</td>
<td>Surgical oncologist, expertise in melanoma</td>
<td>The Netherlands</td>
</tr>
<tr>
<td>Astrid Nollen</td>
<td>Patient research partner, chair Dutch Melanoma Foundation</td>
<td>The Netherlands</td>
</tr>
<tr>
<td>Cynthia Chauhan</td>
<td>Patient research partner</td>
<td>USA</td>
</tr>
<tr>
<td>Oncologist (vacancy)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhD student (vacancy)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How are core outcomes agreed upon?

The HOME roadmap to develop core sets of outcome measurements

Step 1: Define scope and applicability
- Population (condition)
- Intervention
- Setting (e.g., trial, registry)
- Geographical / regional scope
- Stakeholders

Step 2: Develop Core Set of Outcome Domains
- Consensus study involving representatives of relevant stakeholders.

Step 3: Develop Core Set of Outcome Measurements
- Identification and recommendation of adequate measurement instrument(s) for each core outcome domain by a 5-stage process

Stage 1 → Stage 2 → Stage 3 → Stage 4 → Stage 5

- Task
  - Identify all instruments previously used to measure the domain.
  - Establish the extent and quality of testing of the identified instruments.
  - Systematic review of validation studies of the long list of identified instruments.
  - Highlight any gaps in validation.
  - Apply OMERACT filter: Truth, discrimination, and feasibility.
  - Consensus discussion and voting to determine what validation studies will be conducted on short-listed instruments. Gaps in testing were highlighted in stage 2 (systematic review). Appropriate methods used to fill the gaps in validation.
  - Re-apply the OMERACT filter with the results of the component validation studies. Consensus discussion and voting on core outcome to be recommended.

- Methodology
  - Systematic review of outcome instruments used.
  - Systematic review of validation studies of the long list of identified instruments.
  - Highlight any gaps in validation.
  - Consensus discussion and voting on truth.
  - Consensus discussion and voting on discrimination.
  - Consensus discussion and voting on feasibility.

- Output
  - Long list of all instruments have previously been used to measure the domain.
  - Summary of which instruments have been tested and the quality, extent and results of any testing.
  - Short list of potential instruments that meet the requirements of the OMERACT filter.
  - Short list of fully tested instruments.
  - Recommended core outcome instrument for the domain.

Step 4: Disseminate, review, and possibly revise Core Set of Outcome Measurements

Schmitt et al. (2015)
1. Define scope and applicability

**Scope and applicability:**

◊ **Population** (melanoma)
◊ **Setting** (clinical trials)
◊ **Geographical scope** (global)
◊ **Stakeholders*** (all relevant)
1. Define scope and applicability

**Stakeholders:**

◇ Patient representatives
◇ Researchers (incl methodologists)
◇ Healthcare providers (incl dermatologists, oncologists, surgeons, nurses)
◇ Policy makers (incl regulators, payers)
◇ Representatives from pharmaceutical industry
◇ Representatives from drug regulatory authorities
◇ Journal editors
Consensus-based method

Delphi study:
A structured, iterative process to achieve consensus among a group of stakeholders about a given issue

The group of stakeholders does not need to meet which confers anonymity; opinions are to be expressed free from group pressure; and possible dominance of individuals in face-to-face group meetings is being avoided.
2. Define core set of outcomes

◊ **Literature review:** to identify all outcomes that have been measured and reported in melanoma clinical trials*

◊ **Questionnaire survey:** to reach consensus on core outcomes

How important do you consider the assessment of <outcome> of melanoma in clinical trials?

- Consensus: ≥70% scoring 7-9 and ≤15% scoring 1-3
- Approx. 3 rounds

◊ **Group discussions and voting:** to achieve consensus on final core set of outcomes
3. Define core set of OMIs

To identify, validate, or develop an appropriate outcome measurement instrument (OMI) **for each core outcome**

5 stages*

Guideline on instrument selection: recommended to include **only one** outcome measurement instrument for each outcome in the COS

*Prinsen et al. (manuscript in preparation)*
3. Define core set of OMIs

For each core outcome:

Stage 1: Systematic review to identify all OMIs used to assess the construct of interest → long list

Stage 2: Systematic review for each OMI to investigate the quality of the OMIs (COSMIN, qual criteria)

Stage 3: Determine whether OMIs are suitable for the assessment of the core outcome (reliability, validity, and feasibility) → short list

Stage 4: Additional validation studies may be needed

Stage 5: Delphi study to reach consensus on the core outcome measurement instrument → voting
COSMIN taxonomy

Reliability
- Internal Consistency
- Reliability (test-retest, inter-rater, intra-rater)
- Measurement error (test-retest, inter-rater, intra-rater)

Validity
- Content validity
  - Face validity
- Criterion validity (concurrent validity, predictive validity)
- Construct validity
  - Structural validity
  - Hypotheses-testing
  - Cross-cultural validity

Responsiveness
- Responsiveness

Interpretability
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g. using different sets of items from the same health related-patient reported outcomes (HR-PRO) (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons (i.e. raters or responders) on different occasions (intra-rater)</td>
</tr>
<tr>
<td>Internal consistency</td>
<td>The degree of the interrelatedness among the items</td>
</tr>
<tr>
<td>Reliability (extended definition)</td>
<td>The proportion of the total variance in the measurements which is due to ‘true’ differences between patients</td>
</tr>
<tr>
<td>Measurement error</td>
<td>The systematic and random error of a patient’s score that is not attributed to true changes in the construct to be measured</td>
</tr>
<tr>
<td>Validity</td>
<td>The degree to which an HR-PRO instrument measures the construct(s) it purports to measure</td>
</tr>
<tr>
<td>Content validity</td>
<td>The degree to which the content of an HR-PRO instrument is an adequate reflection of the construct to be measured</td>
</tr>
<tr>
<td>Face validity</td>
<td>The degree to which the items of an HR-PRO instrument indeed looks as though they are an adequate reflection of the construct to be measured</td>
</tr>
<tr>
<td>Construct validity</td>
<td>The degree to which the scores of an HR-PRO instrument are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the HR-PRO instrument validity measures the construct to be measured</td>
</tr>
<tr>
<td>Structural validity</td>
<td>The degree to which the scores of an HR-PRO instrument are an adequate reflection of the dimensionality of the construct to be measured</td>
</tr>
<tr>
<td>Hypotheses testing</td>
<td>Idem construct validity</td>
</tr>
<tr>
<td>Cross-cultural validity</td>
<td>The degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument are an adequate reflection of the performance of the items of the original version of the HR-PRO instrument</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>The degree to which the scores of an HR PRO instrument are an adequate reflection of a ‘gold standard’</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>The ability of an HR-PRO instrument to detect change over time in the construct to be measured</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>Idem responsiveness</td>
</tr>
<tr>
<td>Interpretability*</td>
<td>Interpretability is the degree to which one can assign qualitative meaning - that is, clinical or commonly understood connotations – to an instrument’s quantitative scores or change in scores.</td>
</tr>
</tbody>
</table>

* The word ‘true’ must be seen in the context of the CTT, which states that any observation is composed of two components – a true score and error associated with the observation. ‘True’ is the average score that would be obtained if the scale were given an infinite number of times. It refers only to the consistency of the score, and not to its accuracy (ref Steiner & Norman)
<table>
<thead>
<tr>
<th>Domain</th>
<th>Measurement property</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td></td>
<td>The degree to which the measurement is free from measurement error</td>
</tr>
<tr>
<td></td>
<td>Internal consistency (Box A)</td>
<td>The degree of the interrelatedness among the items</td>
</tr>
</tbody>
</table>
For each measurement property standards were developed for how this property should be evaluated
– Design requirements
– Requirements for the statistical methods

The standards refer to the **quality of a study** on measurement properties, not the quality of the instruments that are being evaluated.
### Box B. Reliability: relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability)

<table>
<thead>
<tr>
<th>Design requirements</th>
<th>excellent</th>
<th>good</th>
<th>fair</th>
<th>poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Was the percentage of missing items given?</td>
<td>Percentage of missing items described</td>
<td>Percentage of missing items NOT described</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Was there a description of how missing items were handled?</td>
<td>Described how missing items were handled</td>
<td>Not described but it can be deduced how missing items were handled</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Was the sample size included in the analysis adequate?</td>
<td>Adequate sample size (≥100)</td>
<td>Good sample size (50-99)</td>
<td>Moderate sample size (30-49)</td>
<td>Small sample size (&lt;30)</td>
</tr>
<tr>
<td>4 Were at least two measurements available?</td>
<td>At least two measurements</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Were the administrations independent?</td>
<td>Independent measurements</td>
<td>Assumable that the measurements were independent</td>
<td>Doubtful whether the measurements were independent</td>
<td>measurements NOT independent</td>
</tr>
<tr>
<td>6 Was the time interval stated?</td>
<td>Time interval stated</td>
<td>Time interval NOT stated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Were patients stable in the interim period on the construct to be measured?</td>
<td>Patients were stable (evidence provided)</td>
<td>Assumable that patients were stable</td>
<td>Unclear if patients were stable</td>
<td>Patients were NOT stable</td>
</tr>
<tr>
<td>8 Was the time interval appropriate?</td>
<td>Time interval appropriate</td>
<td>Doubtful whether time interval was appropriate</td>
<td>Time interval NOT appropriate</td>
<td></td>
</tr>
</tbody>
</table>
Use criteria for good measurement properties:

Quality criteria were proposed for measurement properties of health status questionnaires

Caroline B. Terwee\textsuperscript{a,\ast}, Sandra D.M. Bot\textsuperscript{a}, Michael R. de Boer\textsuperscript{a,\textsuperscript{b}}, Daniëlle A.W.M. van der Windt\textsuperscript{a,\textsuperscript{c}}, Dirk L. Knol\textsuperscript{a,\textsuperscript{d}}, Joost Dekker\textsuperscript{a,\textsuperscript{b}}, Lex M. Bouter\textsuperscript{a}, Henrica C.W. de Vet\textsuperscript{a}

\textsuperscript{a}EMGO Institute, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands
\textsuperscript{b}Department of Ophthalmology, VU University Medical Center, Amsterdam, The Netherlands
\textsuperscript{c}Department of General Practice, VU University Medical Center, Amsterdam, The Netherlands
\textsuperscript{d}Department of Clinical Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands
\textsuperscript{e}Department of Rehabilitation Medicine, VU University Medical Center, Amsterdam, The Netherlands

Accepted 29 March 2006
<table>
<thead>
<tr>
<th>Level</th>
<th>Rating</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>strong</td>
<td>+++ or ---</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>++ or --</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>+ or -</td>
<td>One study of fair methodological quality</td>
</tr>
<tr>
<td>conflicting</td>
<td>+/-</td>
<td>Conflicting findings</td>
</tr>
<tr>
<td>unknown</td>
<td>?</td>
<td>Only studies of poor methodological quality</td>
</tr>
</tbody>
</table>
3. Define core set of OMIs

For each core outcome:

Stage 1: Systematic review to identify all OMIs used to assess the construct of interest → long list

Stage 2: Systematic review for each OMI to investigate the quality of the OMIs (COSMIN, qual criteria)

Stage 3: Determine whether OMIs are suitable for the assessment of the core outcome (reliability, validity, and feasibility) → short list

Stage 4: Additional validation studies may be needed

Stage 5: Delphi study to reach consensus on the core outcome measurement instrument → voting
4. Dissemination

Dissemination and implementation:
◊ To enhance the use of the core set in clinical trials
◊ Involvement of all relevant stakeholders
◊ Publications in leading journals
◊ Presentations at relevant meetings
◊ Dissemination to journal editors and reviewers
◊ Dissemination to other stakeholders
→ Guidance materials
→ Monitoring to detect barriers
Potential importance of the results

Standardization in outcome reporting will:
- Allow comparisons across clinical trials
- Improve the usefulness of clinical trial evidence to inform healthcare providers in decision making
- Limit outcome reporting bias

Idzerda et al. (2014)
If we will be awarded...

Announcement: April/May 2015

Proposed project will be embedded within Cochrane Skin Group Outcomes Research Initiative (CSG-COUSIN)

Projected study time lines:
01-Oct-2015 to 30-Sep-2018
- COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN): a guideline for systematic reviews of outcome measurement instruments
CAC Prinsen, LB Mokkink, HCW de Vet, CB Terwee

- How to select outcome measurement instruments for outcomes included in a ‘Core Outcome Set’ – a practical guideline
CAC Prinsen, S Vohra, MR Rose, M Boers, P Tugwell, M Clark, PR Williamson, CB Terwee

www.cosmin.nl    www.comet-initiative.org
Thank you

c.prinsen@vumc.nl

Developing a core outcome set for (stage IV) melanoma trials – results of a systematic review

Stefanie Deckert, Melanie Schubert, Sanna Prinsen, Mariene Garzaroli, Jochen Schmitt