Assessment of clinical signs of atopic dermatitis: A systematic review and recommendation

Jochen Schmitt, MD, MPH,a,b Sinéad Langan, MD, PhD,c,e Stefanie Deckert, MPH,a,e Ake Svensson, MD, PhD,d Laura von Kobyletzki, MD, MSc,e Kim Thomas, PhD,f Phyllis Spuls, MD, PhD,g on the behalf of the Harmonising Outcome Measures for Atopic Dermatitis (HOME) Initiative Dresden, Germany, London and Nottingham, United Kingdom, Malmö and Karlstad, Sweden, and Amsterdam, The Netherlands

Background: Clinical signs are a core outcome domain for atopic dermatitis (AD) trials. The current lack of standardization of outcome measures in AD trials hampers evidence-based communication. Objective: We sought to provide evidence-based recommendations for the measurement of clinical signs in AD trials and to inform the Harmonising Outcome Measures for Atopic Dermatitis Initiative.

Methods: We conducted a systematic review on measurement properties of outcome measurements for clinical signs of AD. We systematically searched MEDLINE and Embase (until October 1, 2012) for validation studies on instruments measuring the clinical signs of AD. Grading of the truth, discrimination, and feasibility of scales; methodological study quality; and recommendations were based on predefined criteria.

Results: Sixteen eligible instruments were identified, of which 2 were best validated. The Eczema Area and Severity Index has adequate validity, responsiveness, internal consistency, intraobserver reliability, and intermediate interobserver reliability but unclear interpretability and feasibility. The Severity Scoring of Atopic Dermatitis Index (SCORAD) has adequate validity, responsiveness, interobserver reliability, and interpretability and unclear intraobserver reliability. Only the objective SCORAD (ie, the clinical signs domain of the SCORAD) is internally consistent. The Six Area, Six Sign Atopic Dermatitis Index severity score and Three Item Severity Score performed inadequately.

Conclusions: The Eczema Area and Severity Index and SCORAD are the best instruments to assess the clinical signs of AD. The other 14 instruments identified are (currently) not recommended because of unclear or inadequate measurement properties. (J Allergy Clin Immunol 2013;132:1337-47.)

Key words: Atopic dermatitis, evidence-based medicine, systematic review, severity of illness index, reliability, validity, responsiveness

Atopic dermatitis (AD) is a major medical condition that causes substantial burden to patients, their families, and society.1-4 Various interventions exist, many of which have been assessed in randomized controlled trials.5 However, clinical trials are only as credible as their end points.6

A previous systematic review7 published in 2007 indicated that most of the published and applied outcome measures for AD had either been inadequately validated or had performed poorly when validated. None of the outcome measures identified in our 2007 article met all the requirements for measurement instruments to be “highly recommended” for use.7

Because of substantial variation in AD outcome measure8 trials, interventions are not comparable. The lack of standardization of AD outcome measures currently makes evidence-based decision making impossible and hampers scientific communication. The international, multiprofessional Harmonising Outcome Measures for Eczema (HOME) Initiative aims to standardize and validate outcome measurements for AD.9 By using Delphi Study methodology,10 the following 4 criteria of AD have been defined as core outcome domains to be assessed in all future AD trials11,12: (1) clinical signs measured with a physician-assessed instrument; (2) symptoms measured with a patient-assessed instrument; (3) health-related quality of life; and (4) long-term control of flares.

The next crucial step in the process of standardization of AD outcome measurements is to identify appropriate instruments to measure each of the 4 core outcome domains of AD. There was broad international consensus among clinicians, patients, and
methodologists that the quality criteria “truth, discrimination, and feasibility” need to be met for AD outcome measures to be recommended by the HOME initiative.12

The objectives of this systematic review were to (1) systematically assess measurement properties of outcome measurements for AD signs, (2) provide evidence-based recommendations for the measurement of clinical signs in future AD trials, and (3) prioritize future validation studies concerning instruments to assess the clinical signs of AD.

METHODS

We conducted a systematic review on the measurement properties of outcome measures for clinical signs of AD according to an a priori study protocol. The protocol of our review was registered in the international prospective register of systematic reviews (PROSPERO; registration no. CRD42013003935; http://www.crd.york.ac.uk/PROSPERO/display_record.asp?ID=CRD42013003935).

Eligibility criteria

On the basis of the research question, inclusion and exclusion criteria were defined for literature search and study selection (Table I). Only articles reporting on AD (population) and validation studies (study design) that investigated clinical signs (outcome) were considered. Articles in which an eligible scale was published first (“inauguration articles”) and articles that subsequently explicitly investigated the measurement properties of eligible scales were eligible for this review. Articles that reported an eligible scale (eg, outcome in a clinical trial without any explicit validation) were not considered eligible.

Literature search

A comprehensive systematic literature search was carried out in MEDLINE (through the PubMed interface) and Embase (through the Ovid interface) for all articles published until October 1, 2012. A specific search string was developed and included search terms for AD, severity of illness, and measurement properties using a combination of medical subject headings developed and included search terms for AD, severity of illness, and all articles published until October 1, 2012. A specific search string was extracted by 4 review teams (J.S./A.S., S.L./L.vK., K.T./S.D., and P.S./J.S.) using standardized and pilot-tested evidence tables. These 4 review teams also extracted data on measurement properties of outcome measures for clinical signs of AD and data on the methodological quality of the studies included. Disagreements among the 4 review teams were resolved by consensus within each team.

Assessment of measurement properties of outcome measures for clinical signs of AD

Measurement properties assessed in this review were content validity, construct validity, internal consistency, interobserver reliability, intraobserver reliability, responsiveness/sensitivity to change, floor and ceiling effects, interpretability, and acceptability/ease of use.15 The predefined criteria of rating properties of outcome measures are in accordance with the recommendations of the Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) group16; the OMERACT filter,13 which has been adopted by the HOME initiative12; and the criteria applied in our previous review on AD outcome measures. Table II provides details concerning definitions of the measurement properties assessed and the criteria for an adequate (+), intermediate (+/−), and inadequate (−) rating. Criterion validity refers to the extent to which a measure relates to a gold standard. The absence of a gold standard to measure the clinical signs of AD was the main reason to perform this review. Therefore the aspect of criterion validity was not considered in this review.

Assessment of the methodological quality of included studies

The COSMIN checklist17,18 was used to evaluate the methodological quality of included studies. On the basis of the extended version of this checklist,19 an overall methodological quality score was calculated by using a 4-point scale (ie, poor, fair, good, or excellent) for each item of the COSMIN checklist. As recommended, the overall score for each item was obtained by taking the lowest score for any question within the item.19 The COSMIN checklist was not applicable for the psychometric properties of floor or ceiling effect, interpretability, and ease of use.

Data synthesis

For each instrument, the evidence concerning each of the abovementioned measurement properties was summarized separately. Considering all evidence

<table>
<thead>
<tr>
<th>TABLE I. Inclusion and exclusion criteria of study selection</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inclusion</strong></td>
</tr>
<tr>
<td>Population AD (synonyms: atopic eczema, eczema, neurodermatitis)</td>
</tr>
<tr>
<td>Outcome</td>
</tr>
<tr>
<td>Clinical signs (eg, edema, erythema, redness, lichenification) and severity by means of multiple items</td>
</tr>
<tr>
<td>Study design</td>
</tr>
</tbody>
</table>

Abbreviations used

AD: Atopic dermatitis
BSA: Body surface area
COSMIN: Consensus-based Standards for the Selection of Health Measurement Instruments
EASI: Eczema Area and Severity Index
HOME: The Harmonising Outcome Measures for Eczema Initiative
oSCORAD: Objective Scoring Atopic Dermatitis Index
POEM: Patient-oriented Eczema Measure
PO-SCORAD: Patient-oriented Scoring Atopic Dermatitis Index
SA-EASI: Self-administered Eczema Area and Severity Index
SASSAD: Six Area, Six Sign Atopic Dermatitis Index
SCORAD: Scoring Atopic Dermatitis Index
TIS: Three Item Severity Score

J ALLERGY CLIN IMMUNOL
DECEMBER 2013
<table>
<thead>
<tr>
<th>Measurement property (name)</th>
<th>Measurement property (description)</th>
<th>Inclusion in OMERACT filter</th>
<th>Criteria for adequate rating (+)</th>
<th>Criteria for intermediate rating (+/−)</th>
<th>Criteria for inadequate rating (−)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content validity</td>
<td>Extent to which the domain “signs” is comprehensively represented by the items in the instrument</td>
<td>Truth</td>
<td>Clinicians OR patients were involved in item selection; clinicians AND patients consider &gt;90% of items as representative for the domain atopic eczema signs</td>
<td>Clinicians OR patients were involved in item selection; clinicians AND patients consider at 70% to 89% of items as representative for the domain atopic eczema signs</td>
<td>NEITHER clinicians NOR patients were involved in item selection; clinicians OR patients consider &lt;70% of items as representative for the domain atopic eczema signs</td>
</tr>
<tr>
<td>Construct validity</td>
<td>Extent to which an instrument truly measures the construct “atopic eczema signs”</td>
<td>Truth</td>
<td>Two different instruments that aim to measure signs of atopic eczema show high correlation (ie, factor loading/correlation coefficient &gt;0.70); OR specific hypotheses were formulated AND ≥75% of the results are in accordance with these hypotheses</td>
<td>Two different instruments that aim to measure signs of atopic eczema show high correlation (ie, factor loading/correlation coefficient 0.60-0.69)</td>
<td>Two different instruments that aim to measure signs of atopic eczema do not show high correlation (ie, factor loading/ correlation coefficient &lt;0.60); AND/OR specific hypotheses were formulated BUT &lt;75% of hypotheses were confirmed, despite adequate design and methods</td>
</tr>
<tr>
<td>Internal consistency</td>
<td>Extent to which items in the scale “signs” are interrelated (measure the same construct)</td>
<td>Discrimination</td>
<td>Factor analyses performed on adequate sample size (7 times the number of items and &gt;100) AND Cronbach α calculated per dimension AND Cronbach α of 0.70-0.95</td>
<td>No factor analysis OR doubtful design or method</td>
<td>Cronbach α &lt;0.70 or &gt;0.95, despite adequate design and methods</td>
</tr>
<tr>
<td>Test-retest reliability/ intrarater variability</td>
<td>Extent to which repeated measurement by the same investigator provides identical results</td>
<td>Discrimination</td>
<td>Correlation coefficient &gt;0.90 OR percentage variation &lt;5% OR coefficient of variation &lt;10%</td>
<td>Correlation coefficient 0.80-0.90 OR percentage variation 5% to 10% OR coefficient of variation 10% to 20%</td>
<td>Correlation coefficient &lt;0.80 OR percentage variation &gt;10% OR coefficient of variation &gt;20%</td>
</tr>
<tr>
<td>Interobserver reliability</td>
<td>Extent to which application of the instrument by different investigators provides identical results</td>
<td>Discrimination</td>
<td>Intraclass correlation &gt;0.80 OR weighted κ &gt;0.60 OR coefficient of variation &lt;20% OR ANOVA (percentage variance explained by observer) &lt;10%</td>
<td>Intraclass correlation 0.60-0.80 OR weighted κ 0.40-0.60 OR coefficient of variation 20% to 30% OR ANOVA (percentage variance explained by observer) 10% to 20%</td>
<td>Intraclass correlation &lt;0.60 OR weighted κ &lt;0.40 OR coefficient of variation &gt;30% OR ANOVA (percentage variance explained by observer) &gt;20%</td>
</tr>
<tr>
<td>Sensitivity to change/ responsiveness</td>
<td>Ability of questionnaire to detect clinically important changes over time</td>
<td>Discrimination</td>
<td>MIC defined; MIC &gt; SDC</td>
<td>MIC defined, but doubtful design or method; MIC &gt; SDC</td>
<td>MIC not defined; MIC ≤ SDC</td>
</tr>
<tr>
<td>Interpretability</td>
<td>Degree to which one can assign qualitative meaning to quantitative scores (ie, ranges for clear/ almost clear, mild, moderate, and severe eczema)</td>
<td>Feasibility</td>
<td>Ranges of the scale have been defined that represent “clear/almost clear,” “mild,” “moderate,” and “severe” atopic eczema atopic eczema signs AND MIC defined</td>
<td>Mean and SD scores presented of ≥3 relevant subgroups of patients but no clear ranges in scale for qualitative meaning of severity of signs defined OR no MIC defined</td>
<td>No information found on interpretation</td>
</tr>
<tr>
<td>Floor or ceiling effects</td>
<td>Number of respondents who achieved the lowest or highest possible score</td>
<td>Not required by OMERACT filter</td>
<td>≤15% of the respondents (of validation study but not in a given RCT) achieved the highest or lowest possible scores</td>
<td>Doubtful design or method</td>
<td>&gt;15% of the respondents (of validation study but not in a given RCT) achieved the highest or lowest possible scores despite adequate design and method</td>
</tr>
<tr>
<td>Acceptability/ ease of use</td>
<td>Degree to which the score can be applied easily, given constraints of time and money</td>
<td>Feasibility</td>
<td>Time to administer &lt;7 min AND score can be used without charge AND no specific tools needed</td>
<td>Time to administer 7-10 min AND score can be used without charge AND no specific tools needed</td>
<td>Time to administer &gt;10 min OR score cannot be used without charge OR specific tool needed</td>
</tr>
</tbody>
</table>

MIC, Minimal important change; RCT, randomized controlled trial; SDC, smallest detectable change.
available, an overall scale quality concerning content validity, construct validity, internal consistency, interobserver reliability, intraobserver reliability, responsiveness/sensitivity to change, floor or ceiling effects, interpretability, and acceptability/ease of use of each outcome measure was rated as adequate (+), intermediate (+/−), or inadequate (−). If 2 studies came to different conclusions concerning a specific measurement property of a specific scale, the study with the higher methodological quality according to the COSMIN checklist17-19 informed the overall rating.

Predefined criteria for recommendations

For each instrument identified in the review, a standardized recommendation for use or required future validation work was prepared, depending on the measurement properties.

According to the results of the HOME II meeting,12 all 3 criteria of the OMERACT filter (ie truth, discrimination, and feasibility)13 have to be met by an outcome measure to be recommended by the HOME initiative.

Four categories of recommendation were specified as follows:

A. The outcome measure meets the requirements of truth (content validity and construct validity), discrimination (reliability, internal consistency, and sensitivity to change), and feasibility (interpretability and ease of use) and is recommended for use.

B. The outcome measure meets 2 or more requirements, but performance in other required measurement properties is unclear, so that the outcome measure has the potential to be recommended in the future depending on the results of further validation studies.

C. The outcome measure has inadequate quality in at least 1 required measurement property (≥1 rating of "minus") and therefore is not recommended to be used.

D. The outcome measure has (almost) not been validated. Its performance in all or most relevant quality requirements is unclear, so that it is not recommended to be used until further validation studies clarify its measurement properties.

RESULTS

We identified 45 eligible articles.7,20-63 Of these, 43 studies were identified by means of electronic search, and 2 studies47,57 were identified through hand search. The flow diagram (Fig 1)64 provides information on study identification and the selection process. Table E2 in this article’s Online Repository at www.jacionline.org presents a list of the articles that were excluded based on full-text screening, including the reason for exclusion. Interrater reliability among the 3 reviewers was substantial65 for title and abstract screening (Cohen κ = 0.75-0.79; agreement = 98.9% to 99.3%) and poor to almost perfect65 for full-text inclusion (Cohen κ = 0.31-0.84; agreement = 65.5% to 92.3%).

Study characteristics

The 45 articles reported on 16 different instruments to assess the clinical signs of AD. The characteristics of these 45 articles are summarized by outcome measurement in Table III.7,20-63 Most validation studies investigated the Scoring Atopic Dermatitis Index (SCORAD)32 and the Eczema Area and Severity Index (EASI).54 For both the SCORAD and EASI, more than 2000 patients with a broad spectrum of age groups and degrees of severity of AD were included in the identified validation studies (Table III).

Content of the instruments identified

Thirteen different clinical signs are included in the identified instruments, with each instrument including 3 to 8 clinical signs. Erythema is the most frequently included (14/16 instruments). Lichenification is included in 10 of 16 instruments. The items edema/induration/papulation, oozing/crusting/weeping/exudation,
and excoriation are included in 9 of 16 instruments. Dryness and scaling are assessed in 8 and 6 instruments, respectively. Cracking/fissuring, vesicles, (de)pigmentation, flaking, bleeding, and erosions are included in 3 or less instruments to assess the clinical signs of AD (Table IV). All instruments except the Patient-oriented Eczema Measure (POEM)\(^{36}\) and the Three Item Severity Score (TIS)\(^{60}\) include an assessment of intensity and extent of clinical signs. Charman et al\(^{35}\) assessed clinical signs and symptoms of AD, a unidimensional version only including the intensity and extent of clinical signs, the objective Scoring Atopic Dermatitis Index (oSCORAD),\(^{22}\) is available.

### Studies investigating content validity

Two studies\(^{35}\) on the content validity of measures to assess the intensity and extent of clinical signs were identified. Charman et al\(^{35}\) assessed clinical signs and BSA involvement in 180 consecutive patients with AD and analyzed their relationship with a patient global bother score by means of bivariate and multivariate regression. Excoriations, erythema, and edema/population were independent predictors of patient-rated disease severity. BSA involvement, which might not be reliably assessed,\(^{41}\) was not linearly related with the global bother score of greater than 30% BSA involvement. Schmitt et al\(^{7}\) asked 12 consumers and 6 dermatology experts to rate the content validity of all items and domains included in the identified outcome measures for AD on a 5-point Likert scale. Both experts and consumers considered the intensity

---

**TABLE III. Study characteristics of studies included by outcome measurement**

<table>
<thead>
<tr>
<th>Outcome measurement</th>
<th>No. of validation studies</th>
<th>Setting: Primary care Secondary/tertiary care</th>
<th>Severity of AD</th>
<th>Study population</th>
<th>No. of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCORAD(^{32})</td>
<td>26(^{1,3,11,12,15,16,20,22-24,26,28-30,32,34,35,37,45,46,48,49,51-53,55,56,58-63})</td>
<td>Secondary/tertiary care, n = 2; secondary/tertiary care, n = 20; all settings, n = 2; NR, n = 2</td>
<td>Mild to severe, n = 14; moderate to severe, n = 5; severe, n = 1; NR, n = 6</td>
<td>Primary care, n = 2</td>
<td>4127 15-823 0-74</td>
</tr>
<tr>
<td>PO-SCORAD(^{58})</td>
<td>2(^{1,52,58})</td>
<td>All secondary/tertiary care</td>
<td>Mild to severe, n = 1; NR, n = 1</td>
<td>Primary care, n = 2</td>
<td>504 33-471 2-74</td>
</tr>
<tr>
<td>EASI(^{54})</td>
<td>9(^{20,21,29,35,40,47,49,55,61})</td>
<td>Secondary/tertiary care, n = 6; all settings, n = 1; NR, n = 2</td>
<td>Mild to moderate, n = 1; mild to severe, n = 5; severe, n = 1; NR, n = 2</td>
<td>Secondary/tertiary care, n = 1</td>
<td>2062 20-1550 0-70</td>
</tr>
<tr>
<td>SA-EASI(^{41})</td>
<td>4(^{21,23,24,56})</td>
<td>All secondary/tertiary care</td>
<td>Moderate to severe, n = 3; NR, n = 1</td>
<td>Secondary/tertiary care, n = 1</td>
<td>240 35-98 1-17</td>
</tr>
<tr>
<td>SASSAD(^{53})</td>
<td>5(^{33,35,42,56,61})</td>
<td>Secondary/tertiary care, n = 3; all settings, n = 1; NR, n = 1</td>
<td>Mild to severe, n = 3; moderate to severe, n = 1; mild to moderate, n = 1</td>
<td>Secondary/tertiary care, n = 1</td>
<td>340 6-180 0-67</td>
</tr>
<tr>
<td>POEM(^{56})</td>
<td>2(^{56,49})</td>
<td>All settings, n = 1, NR, n = 1</td>
<td>Mild to severe, n = 1</td>
<td>Secondary/tertiary care, n = 2</td>
<td>120 40-80 1-58</td>
</tr>
<tr>
<td>BCSS(^{57})</td>
<td>5(^{51})</td>
<td>Secondary/tertiary care</td>
<td>Mild to severe, n = 1; moderate to severe, n = 1</td>
<td>Secondary/tertiary care, n = 1</td>
<td>82 NA 0-67</td>
</tr>
<tr>
<td>ADAM(^{43})</td>
<td>2(^{51,44})</td>
<td>Secondary/tertiary care</td>
<td>Mild to severe, n = 1; moderate to severe, n = 1</td>
<td>Secondary/tertiary care, n = 1</td>
<td>224 53-171 0-16</td>
</tr>
<tr>
<td>ADASI(^{38,39})</td>
<td>5(^{39})</td>
<td>Secondary/tertiary care</td>
<td>NR</td>
<td>Secondary/tertiary care, n = 2</td>
<td>100 NA 0-70</td>
</tr>
<tr>
<td>ADQ(^{34})</td>
<td>1(^{34})</td>
<td>Secondary/tertiary care</td>
<td>Mild to severe</td>
<td>Secondary/tertiary care, n = 2</td>
<td>68 NA 0-12</td>
</tr>
<tr>
<td>OSAAD(^{53})</td>
<td>3(^{37,45,53})</td>
<td>All secondary/tertiary care</td>
<td>Mild to severe, n = 2; moderate to severe, n = 1</td>
<td>Secondary/tertiary care, n = 1</td>
<td>92 22-38 0-57</td>
</tr>
<tr>
<td>SSS(^{31})</td>
<td>2(^{51,51})</td>
<td>All secondary/tertiary care</td>
<td>All mild to severe</td>
<td>Secondary/tertiary care, n = 4</td>
<td>96 7-34 0-67</td>
</tr>
<tr>
<td>TIS(^{60})</td>
<td>7(^{50,32,35,51,59,60,62})</td>
<td>Primary care, n = 2; secondary/tertiary care, n = 4; all settings, n = 1</td>
<td>Mild to severe, n = 6; NR, n = 1</td>
<td>Secondary/tertiary care, n = 1</td>
<td>926 66-180 0-67</td>
</tr>
<tr>
<td>W-AZS(^{50})</td>
<td>0(^{50})</td>
<td>NA</td>
<td>NA</td>
<td>Secondary/tertiary care, n = 2</td>
<td>28 NA 0-8</td>
</tr>
<tr>
<td>Unnamed scale(^{125})</td>
<td>1(^{25})</td>
<td>Secondary/tertiary care</td>
<td>NR</td>
<td>Secondary/tertiary care, n = 1</td>
<td>28 NA 0-8</td>
</tr>
<tr>
<td>Unnamed scale(^{225})</td>
<td>1(^{25})</td>
<td>Secondary/tertiary care</td>
<td>NR</td>
<td>Secondary/tertiary care, n = 1</td>
<td>28 NA 0-8</td>
</tr>
</tbody>
</table>

---

\(^{a}\)References 21, 25, 31-34, 36, 38, 39, 43, 50, 53, 54, 57, 58, and 60.

\(^{b}\)References 21, 25, 31, 32, 34, 36, 38, 43, 50, and 58.
of lesions and the extent of AD as “very important” criteria for the assessment of the severity of AD. Excoriations, erythema, edema/papulation, dryness, lichenification, and oozing/crusting/exudation were considered “important” or “very important” by consumers and experts. 

**Table V.** Number of studies assessing the measurement properties of outcome measures for clinical signs of AD

<table>
<thead>
<tr>
<th>Measurement property (name)</th>
<th>ADAM 43</th>
<th>ADASI 38,39</th>
<th>ADQ 34</th>
<th>BCSS 57</th>
<th>EASI 64</th>
<th>OSAAD 53</th>
<th>POEM 40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content validity</td>
<td>27,35</td>
<td>27,35</td>
<td>27,35</td>
<td>27,35</td>
<td>27,35</td>
<td>27,35</td>
<td>37,35,36</td>
</tr>
<tr>
<td>Construct validity</td>
<td>—</td>
<td>—</td>
<td>134</td>
<td>151</td>
<td>5*</td>
<td>37,45,53</td>
<td>—</td>
</tr>
<tr>
<td>Internal consistency</td>
<td>144</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>29,40</td>
<td>—</td>
<td>136</td>
</tr>
<tr>
<td>Intraobserver reliability</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>147</td>
<td>—</td>
<td>137</td>
<td>—</td>
</tr>
<tr>
<td>Interobserver reliability</td>
<td>143</td>
<td>—</td>
<td>—</td>
<td>151</td>
<td>29,47</td>
<td>137</td>
<td>—</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>240,49</td>
<td>—</td>
<td>137</td>
<td>2,36,49</td>
</tr>
<tr>
<td>Floor or ceiling effects</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>237,53</td>
<td>—</td>
</tr>
<tr>
<td>Interpretability</td>
<td>—</td>
<td>134</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>137</td>
<td>136</td>
</tr>
<tr>
<td>Acceptability</td>
<td>—</td>
<td>139</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>137</td>
<td>136</td>
</tr>
</tbody>
</table>

ADAM, Atopic Dermatitis Assessment Measure; ADASI, Atopic Dermatitis Area and Severity Index; ADQ, Atopic Dermatitis QuickScore; BCSS, Basic Clinical Scoring System; OSAAD, Objective Severity Assessment of Atopic Dermatitis; SSS, Simple Scoring System; W-AZS, New Scoring System for Assessment of the Extent and Severity of Skin Inflammation Index in Atopic Dermatitis Patients.

*References 20, 21, 29, 55, and 61.
†References 21, 23, 24, and 56.
‡References 23, 24, 29, 30, 34, 37, 45, 51-53, 55, 56, 58-61, and 63.
§References 28, 29, 32, 45, 46, 48, and 62.
¶References 22, 23, 26, 29, 32, 34, 37, 51, 59, 60, and 62.
#References 32, 51, 59, 60, and 62.

**Measurement properties of outcome measures for clinical signs of AD**

Table V summarizes the number of studies assessing the different measurement properties of each outcome measure identified. From the total of 146 assessed (sub)studies included, the COSMIN checklist was applicable for 123 validation (sub)studies. Of these, 17 (14%) were rated as having excellent, 8 (6%) as having good, 54 (44%) as having fair, and 44 (36%) as having poor quality. The most frequent reason for fair or poor quality was low sample size. Table E3 in this article’s Online Repository at www.jacionline.org provides detailed information about the study base, methods, results, quality, and conclusions of each of these validation (sub)studies. The SCORAD is the only instrument for which each measurement property of interest has been investigated according to our predefined criteria (Tables II and V).

Table VI provides a summary of measurement properties of each outcome measure for clinical signs of AD and recommendations. None of the studies met all of our predefined requirements of truth, discrimination, and feasibility. Currently, the EASI and SCORAD are the most valid and reliable instruments to assess clinical signs of AD.
The EASI and SCORAD both have adequate content validity. High correlation between the EASI and SCORAD \((r = 0.84-0.93)\) indicates adequate construct validity of both instruments. Correlation between the EASI and the Self-administered Eczema Area and Severity Index (SA-EASI; \(r = 0.17-0.62\)) and between the SCORAD and the Patient-oriented Scoring Atopic Dermatitis Index (PO-SCORAD; \(r = 0.27-0.79\)) was weaker, suggesting that the patient-assessed modifications of the EASI and SCORAD measure different constructs than the original investigator-assessed instruments. For an overview of the strength of correlation between the different instruments included in this review, please refer to Table E4 in this article’s Online Repository at www.jacionline.org. The EASI is internally consistent (Cronbach’s \(\alpha = 0.94\)). The SCORAD is a composite score, and therefore internal consistency within the constructs assessed (signs and symptoms) is relevant. The objective parts of the SCORAD that assess the intensity and extent of clinical signs are strongly interrelated \((r = 0.64-0.86)\), suggesting internal consistency of the objective SCORAD. The EASI had adequate intraobserver reliability and intermediate interobserver reliability in a population of 20 children and adults with different degrees of severity of AD assessed by 15 trained investigators. Interobserver reliability of the SCORAD was adequate in different populations of children and adults with different degrees of disease severity (see Table E3 for details). Intraobserver reliability of the SCORAD was investigated in one study, but grading of intraobserver reliability according to other domains.
TABLE VI. Summary of measurement properties of outcome measures for clinical signs of AD and recommendations

<table>
<thead>
<tr>
<th>Measurement property (name)</th>
<th>ADAM*35</th>
<th>ADASI**38,39</th>
<th>ADQ**44</th>
<th>BCSS*57</th>
<th>EASI**44</th>
<th>OSAAD*53</th>
<th>POEM**60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content validity</td>
<td>−</td>
<td>+/−</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Construct validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internal consistency</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intraobserver reliability</td>
<td></td>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>+/−</td>
<td></td>
</tr>
<tr>
<td>Interobserver reliability</td>
<td>+/−</td>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Sensitivity to change</td>
<td></td>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Floor or ceiling effects</td>
<td></td>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Interpretability</td>
<td></td>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Acceptability</td>
<td></td>
<td></td>
<td></td>
<td>−</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Recommendation</td>
<td>C</td>
<td>D</td>
<td>C</td>
<td>C</td>
<td>B</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

Recommendations are defined as follows: A, outcome measure meets the requirements of truth (content validity and construct validity), discrimination (reliability, internal consistency, and sensitivity to change), and feasibility (interpretability and ease of use) and is recommended for use; B, outcome measure meets 2 or more requirements, but performance in other required measurement properties is unclear, so that the outcome measure has the potential to be recommended in the future depending on the results of further validation studies; C, outcome measure has inadequate quality in at least 1 required measurement properties (≥1 rating of "minus") and therefore is not recommended to be used; and D, outcome measure has (almost) not been validated. Its performance in all or most relevant quality requirements is unclear, so that it is not recommended to be used until further validation studies clarify its measurement properties.

ADAM, Atopic Dermatitis Assessment Measure; ADASI, Atopic Dermatitis Area and Severity Index; ADQ, Atopic Dermatitis Quickscore; BCSS, Basic Clinical Scoring System; NA, not applicable; NR, not reported, indicating measurement property unclear; OSAAD, Objective Severity Assessment of Atopic Dermatitis; W-AZS, New Scoring System for Assessment of the Extent and Severity of Skin Inflammation Index in Atopic Dermatitis Patients.

+ Positive rating indicating "adequate" measurement property; +/−, intermediate rating indicating "intermediate" measurement property; −, inadequate rating indicating "inadequate" measurement property (please refer to Table II for definitions).

*Compared with other clinical sign measures.
†Further research is necessary to clarify whether an outcome measure with intermediate content validity is sufficient to capture clinical signs of AD in clinical trials.

our predefined criteria was not possible. In a study that used an investigator global assessment as the anchor, the minimal clinical important difference was 8.7 points for the SCORAD, 8.2 for the objective SCORAD, and 6.6 for the EASI, indicating adequate sensitivity to change. Reports on the feasibility of the EASI (ie, floor or ceiling effects), interpretability, and ease of use were not identified. The SCORAD had no floor or ceiling effects in 2 studies. Score ranges to define mild, moderate, and severe AD have been defined by expert opinion for the SCORAD and the oSCORAD. The time to perform the SCORAD has been reported to be “about 10 minutes.”

The SASSAD and TIS fulfill some quality criteria (Table VI), but the performance in other required measurement properties is unclear. These outcome measures might have the potential to be recommended in the future depending on the results of further validation studies. However, both the SASSAD and TIS only have intermediate content validity, which is a disadvantage compared with the EASI and SCORAD.

Seven outcome measures (the Atopic Dermatitis Assessment Measure, Atopic Dermatitis Quickscore, Basic Clinical Scoring System, Objective Severity Assessment of Atopic Dermatitis, POEM, Simple Scoring System, and New Scoring System for Assessment of the Extent and Severity of Skin Inflammation Index in Atopic Dermatitis Patients) have inadequate quality in at least 1 of the required measurement properties and are therefore not recommended to be used to assess clinical signs of AD. The POEM is internally consistent, reliable, responsive, and feasible but is primarily a symptoms scale that does not capture the content required of a clinical signs instrument.

Five outcome measures (unnamed scale 1, unnamed scale 2, PO-SCORAD, SA-EASI, and the Atopic Dermatitis Area and Severity Index) have almost not been validated, and their measurement properties are unclear. These outcome measures are currently not recommended (Table VI). Table E3 provides a summary of all validation studies on the 16 instruments to assess clinical signs of AD.

DISCUSSION

This systematic review identified, summarized, and critically appraised 16 different outcome measures to assess the severity of clinical signs of AD. Although none of these instruments completely met all the predefined criteria for adequate truth, discrimination, and feasibility, substantial validation work has been undertaken since our previous review.

The EASI and SCORAD are the best instruments available to measure clinical signs of AD.

The SCORAD is a valid, internally consistent, responsive, interpretable composite score that includes the intensity and extent of clinical signs of AD and the severity of AD symptoms. The SCORAD has adequate interobserver reliability and no floor or ceiling effects. The oSCORAD is internally consistent. Intraobserver reliability of the SCORAD is unclear and requires further investigation. Disadvantages of the SCORAD include the difficulty of identifying an “average representative area” for the assessment of the intensity of clinical signs and the underlying assumption that the extent of AD is linearly related to severity, which has been shown to be only true in patients with up to 30% BSA affected.

The EASI is valid and internally consistent and has adequate intraobserver reliability, intermediate interobserver reliability, and adequate responsiveness. One advantage of the EASI is its unidimensionality; that is, it only measures clinical signs of AD and does not include symptoms. Another advantage of the EASI is that it measures the intensity of the lesions at multiple body parts (rather than relying on a representative lesion). Ease of use and floor or ceiling effects of the EASI have not yet been reported in a validation study. However, from clinical trials that applied the EASI, we have some evidence concerning the absence of floor or
ceiling effects, and personal experience indicates that the time needed to perform the EASI and oSCORAD is similar. The main disadvantage of the EASI is that interpretability data are lacking (ie, definition of ranges of the EASI that represent mild, moderate, and severe clinical signs of AD). Ideally, interpretability of the EASI should be investigated with a global (gestalt) assessment of the severity of AD signs as a reference measure.

Clinical signs are typically assessed by an investigator. However, for the EASI and SCORAD, patient-assessed variants have been introduced. As highlighted in this review, content and measurement properties differ substantially between the EASI and SA-EASI and between the SCORAD and PO-SCORAD, so that these patient-assessed instruments cannot be used interchangeably with the original investigator-assessed outcome measures. On the basis of the existing evidence, the SA-EASI and PO-SCORAD are not recommended to assess clinical signs of AD.

The conceptual framework (ie, the underlying relationship between the items of a measurement instrument and the construct [domain] “clinical signs of AD” to be measured) follows a formative model. This means that the items determine or cause the construct to be assessed. In such a situation, content validity is of utmost importance. Content validity has been defined as “the degree to which the content of a measurement instrument is an adequate reflection of the construct to be measured.” Two studies included in the review informed content validity and suggest that the extent of AD and the intensity of excoriations, erythema, edema/papulation, dryness, lichenification, and oozing/crusting/exudation were considered as “important” or “very important” by consumers and experts. The EASI and SCORAD are the only outcome measures that have adequate content validity according to our predefined criteria. Although the SASSAD and TIS fulfill some important quality criteria and therefore received recommendation B (ie, possibly recommendable if validation gaps are closed), a consensus involving the relevant stakeholders is necessary to clarify whether an outcome measure with intermediate content validity is sufficient to capture the clinical signs of AD in clinical trials.

Strengths and limitations of this systematic review

We applied predefined criteria to rate the measurement properties of instruments and make recommendations. The protocol was preregistered. We used a validated, highly sensitive search strategy for validation studies, which identified all articles included in our previous review listed in PubMed, Embase, or both. We did not consider unpublished data or studies only published as an abstract because we believed that relevant information would be missing. We applied the COSMIN checklist methodology to rate study quality and define the criteria for adequate measurement properties. Even though this checklist was originally developed for the assessment of the methodological quality of patient-reported outcome measures, we believe that it is also applicable for non–patient-reported outcome measures.

Reliability is substantially influenced by the variability between patients and cannot be considered a fixed characteristic of an instrument, such as content validity. The degree of variability of the severity of clinical signs of AD was heterogeneous across the populations assessed in the reliability studies, so that the reliability results might not be directly comparable between the different instruments.

Recommendations for researchers, clinicians, and decision makers

Clinical sign measures have been the primary end point in the vast majority of all clinical trials undertaken to investigate the efficacy of interventions for AD. Clinical signs assessed by a physician have been included in the core set of outcome domains
for AD trials based on a 100% consensus between the 43 panel members participating in our previous multiperspective, international Delphi consensus study on outcome measurement in patients with AD. This review indicates that either the EASI or SCORAD should be used to assess the clinical signs of AD in future clinical trials. If the SCORAD is used, it should be reported as a profile (ie, the scores for the intensity and extent of AD signs and the severity of AD symptoms should be reported separately).

The ultimate goal from the perspective of evidence-based medicine is to achieve worldwide consensus to consistently apply a single valid, reliable, and feasible instrument to measure the clinical signs of AD in all future clinical trials. The HOME Initiative (www.homeforeczema.org/) is dedicated to achieving this goal.

Key messages
- Valid, reliable, and feasible outcome measures are a prerequisite to evidence-based health care.
- This systematic review on outcome measures for clinical signs of AD identified and critically appraised 16 different instruments.
- The EASI and SCORAD have the best measurement properties and are recommended for use in future clinical trials.

REFERENCES


