Patient-reported outcome (PRO) instruments for disease severity and quality of life in patients with atopic dermatitis: a systematic review of English and Chinese literature

Aihua Li¹, Minlu Zhang¹, Yating Yang², Joshua Zhang¹, Xiaoping Xie¹, Xiaoian Li², Hui Zhang³

¹Sanofi (China) Investment, Shanghai, China; ²Shanghai Jusure Health Technology, Shanghai, China; ³Department of Dermatology, Xinhua Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China

Contributions: (I) Conception and design: A Li, H Zhang; (II) Administrative support: M Zhang, X Li, X Xie; (III) Provision of study materials or patients: A Li, M Zhang, Y Yang, J Zhang; (IV) Collection and assembly of data: Y Yang, A Li, X Li, J Zhang; (V) Data analysis and interpretation: Y Yang, A Li, H Zhang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Dr. Hui Zhang. Xinhua Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai 200092, China. Email: zhanghui@xinhuamed.com.cn.

Background: Many patient-reported outcome (PRO) on disease severity quality of life (QOL) have been developed for atopic dermatitis (AD) patients. However, none of them on the reliability and validity of the instruments was sufficient for their application in clinical studies. The objective of this study is to identify and assess the quality of recently developed PROs for disease severity and QOL in English and Chinese in AD patients.

Methods: We conducted a systematic review of PROs for disease severity and QOL for AD from PubMed, Web of Science, PsycINFO and ERIC (English literatures), and CNKI and Wanfang Data (Chinese literatures) from September 2010 to December 2021 with string including “atopic dermatitis” and “scaling”. All studies were screened by 2 reviewers. After being removed duplications, the studies developed the instruments for the AD patients, were reported by patients, and assessing the disease severity or QOL were included.

Results: Twenty-six instruments were retrieved. Three single-item Numeric Rating Scale (NRS) and 8 multidimensional instruments assessing disease severity and 15 assessing QOL were found to be originally developed in English (n=23) or Chinese (n=3). After full assessment on the reliability and validity, 3 NRS and 9 multidimensional instruments were recommended. The 3 NRS were Peak Pruritus/Itch NRS, Skin Pain NRS and Sleep Disturbance (SD) NRS. The multidimensional instruments for disease severity included the Patient-Oriented Eczema Measure (POEM), the patient oriented-SCORAD (PO-SCORAD), and Atopic Dermatitis Symptom Score (ADSS), and the instruments for QOLs included Infant’s Dermatology Quality of Life Index (IDQOL), Children’s Dermatology Life Quality Index (CDLQI), Atopic Dermatitis Control Tool (ADCT), PROMIS® Itch Questionnaire Mood and Sleep (PIQ-MS), PROMIS-Sleep Disturbance (PROMIS-SD), and PROMIS-Sleep-Related Impairment (PROMIS-SRI) for QOL. However, none of the Chinese PROs either originally developed or adapted were fully validated.

Discussion: Single-item NRS is a complement to multidimensional PROs in assessing the disease severity of AD. Quality of these instruments vary greatly and only a few instruments that meet the Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) standards are recommended. Therefore, standardization of PROs is essential for developing new instruments, and for adapting a PRO in other populations with different culture and languages.

Keywords: Atopic dermatitis (AD); patient report outcome (PRO); quality of life (QOL)

Submitted May 19, 2022. Accepted for publication Aug 17, 2022.
doi: 10.21037/atm-22-3164

View this article at: https://dx.doi.org/10.21037/atm-22-3164
Introduction

Atopic dermatitis (AD) is a common major pruritic skin condition that leads to a substantial burden on patients, their families, and society (1-3). AD affects ~20% of children and 1–10% of adults with an increasing prevalence worldwide (4). AD most often begins in infancy or early childhood, with ~90% of cases appearing within the first 5 years of life (5). The course of AD can be long-lasting, relapsing, and often significantly affects the quality of life (QOL) of patients and their families. In the past a few years, increasing attention has been given to AD, with the development of various treatments and therapies, including biological therapies (6). However, the outcome measurements for assessing AD are usually based on clinical signs, as currently few satisfactory objective markers of disease activity are adequate and reliable to be used as a golden standard (7). Patients report outcome (PRO) is able to reflect any status of a patient's health condition that immediately provided by the patient. And PRO avoids the secondary changes and misunderstanding of the patient’s responses by any third party (8). Therefore, the disease severity related symptoms that are directly reported by patients are essential for assessing the efficacy of the treatment in patients with AD.

The Food and Drug Administration (FDA) Guidelines for Industry recommend the primary endpoint of AD treatment success to be based on the Investigator’s Global Assessment (IGA) score difference, which is a clinical assessment scale that depends on physical symptoms assessed by physicians (9). This clinician-rated scale is intended to be objective measures of highly visible symptoms (e.g., redness, flaking, bleeding from scratching) and measurable functional impairment. Unfortunately, this approach has limited patient input on treatment outcomes in AD. Furthermore, the IGA was defined by a particular study sponsor in a particular context, resulting in variation in IGA versions and IGA has not been adequately validated until a standardized vIGA is published (10). Importantly, AD often causes constant, intensive itching, and impaired psychosocial and working functioning (1). Psychiatric comorbidities, including depression, anxiety, and suicidal ideation, are more common in AD patients than general population, even among patients with clinically mild or moderate conditions (11). Therefore, despite the physical burden of AD, healthcare providers may underestimate the psychological effect of the disease (1). On the other hand, key symptoms and impacts of AD, such as pruritus, sleep disturbance, and interference with activities, are more difficult or impossible for clinicians to assess. Furthermore, the meaningfulness of clinical improvement can only be assessed by patients (12). Unfortunately, PROs have been used in some forms in only a small portion of clinical trials (13). As the principle of patient-centered care is becoming increasingly recognized and valued, instruments that directly assess the impact of AD on the QOL of patients are needed to determine the effectiveness of treatments as stated in the 21st century Cures Act (14). Therefore, PROs, which collects disease-related information directly from the patient without any interpretation, are an important complement to the clinician-reported outcomes and are increasingly expected and considered for the evaluation of treatment outcomes (1). Meanwhile, unlike a survey questionnaire, PROs are conducted by experts and their content covers every aspect of the patient's experience, including symptom burden, mood, physical function, QOL, and distress (15). Using the information extracted from the PROs, healthcare providers can provide more patient-focused and specific therapies, which meet the requirement of the Patient-Focused Drug Development guideline issued by the US Food and Drug Administration in June, 2020 (16).

The rapid development of new treatments for AD, such as alefacept, dupilumab, and upadacitinib, has also highlighted the need for a generalizable scale to assess the severity and progress of AD to evaluate and compare treatment effectiveness (17-19). At present, the Eczema Area and Severity Index (EASI) and IGA instruments are the primary endpoints in most clinical studies of AD, but neither was included in this review because they are completed by healthcare providers. Increasingly, more clinical trials are utilizing PROs as an endpoint for testing drug efficacy and safety. There are a variety of instruments aiming at quantifying AD outcomes. A patient-reported outcome measures (PROM) is a questionnaire used to elicit information directly from patients, covering the measures of symptoms, activity limitations, health status, health-related quality of life (HRQOL), QOL, etc. (20). As physiological and psychological burdens often occur concurrently for patients with AD, both disease severity and QOL measures are fundamental to patient evaluation and care (21). Therefore, this study focused on two aspects of PRO: disease severity and QOL.

A previous systematic review published in 2011 identified a total of 20 disease severity scales and 14 QOL instruments in English used in randomized controlled trials (RCTs) of AD treatment from 1985 to 2010 (22). In
Table 1: Search strategy

<table>
<thead>
<tr>
<th>#</th>
<th>Search strategy</th>
<th>No. of records</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Atopic Dermatitis*) OR (Atopic Neurodermatitis)</td>
<td>16,206 (PubMed); 9,922 (WoS); 167 (PsycInfo); 167 (ERIC); 112,111 (CNKI); 15,104 (Wanfang)</td>
</tr>
<tr>
<td>2</td>
<td>(Animals) OR (canine OR dog OR dogs OR cat OR cats)</td>
<td>1,928,982 (PubMed); 611,382 (WoS); 58,559 (PsycInfo); 58,559 (ERIC); 4,455,372 (CNKI); 690,783 (Wanfang)</td>
</tr>
<tr>
<td>3</td>
<td>1 NOT 2</td>
<td>9,893 (PubMed); 9,034 (WoS); 162 (PsycInfo); 162 (ERIC); 36,602 (CNKI); 14,584 (Wanfang)</td>
</tr>
<tr>
<td>4</td>
<td>(questionnaire*) OR (scal*) OR (assessment) OR (indicato*) OR (measur*) OR (score*)</td>
<td>3,838,315 (PubMed); 5,936,191 (WoS); 627,265 (PsycInfo); 627,265 (ERIC); 2,596,416 (CNKI); 2,581,210 (Wanfang)</td>
</tr>
<tr>
<td>5</td>
<td>3 AND 4</td>
<td>4,372 (PubMed); 3,297 (WoS); 69 (PsycInfo); 69 (ERIC); 1,332 (CNKI); 1,293 (Wanfang)</td>
</tr>
</tbody>
</table>

Database(s): PubMed, Web of Science, PsycINFO and ERIC (English); CNKI and Wanfang Data (Chinese). Publication period: September 2010 to 31 December 2021.

A more recent review published in 2016, 62 disease severity measures and 28 QOL instruments in English were identified and analyzed (23). However, in such a rapid developing field, there are numerous new instruments and follow-up studies assessing the validity and reliability of existing instruments, but none has delivered a complete and precise evaluation of the quality of the instruments. Therefore, the identification of new instruments and reassessment of the quality of the existing ones are crucial for choosing the appropriate instrument for use in clinical studies and practice.

There are significantly fewer AD assessment instruments in Chinese and they often lack validation and reliability testing after translation. In China, the prevalence of AD has increased from 0.7% (age 6–20 years) in 2000 to 8.3% in Shanghai in 2012 (age 3–6 years) (24). The incidence of AD in outpatients has also dramatically increased from 2.3% in 2008 to 7.8% in 2016 (24, 25). In addition to changes in environmental factors and lifestyle, the increased AD incidence rate is largely contributed to a correction in diagnostic methods, which historically have caused overdiagnosed eczema and underdiagnosed AD (21). The lack of an accurate gold standard to differentiate the two diseases further challenges comparability between AD trials in China. The treatment guideline of AD in China has been a combination of topical corticosteroid (TCS) and traditional Chinese medicine (TCM) (26). TCM has been used especially to treat children aged 0–12 years in clinical practice, mostly aiming to reduce the use for TCS (27). As 31–36% of TCM users are combining it with TCS, unique instruments that offers insights to the prescription of both TCM and TCS would be a good reference for clinicians and also good candidates for further studies (28).

We aimed to assess the quality of existing PROs in English and Chinese in patients with AD which measure the disease severity and QOL instruments by systematically reviewing the instrument development and validation literature published between September 2010 and December 2021. Specifically, we sought to (I) evaluate the measurement properties of the outcome measurements of commonly used instruments in both Chinese and English; (II) identify the gaps in instrument translation, adaptation and validation; and (III) prioritize future validation studies of instruments to assess disease severity and QOL of AD. We present the following article in accordance with the PRISMA reporting checklist (available at https://atm.amegroups.com/article/view/10.21037/atm-22-3164/rc).

Methods

Search strategy and eligibility criteria

A comprehensive systematic literature search was carried out in PubMed, Web of Science, PsycINFO and ERIC (for literature in English), and in CNKI and Wanfang Data (for literature in Chinese). In order to capture recently used instruments, the search was limited to studies published from September 2010 to December 31, 2021. A specific search string including search terms of “atopic dermatitis” and “scaling” was developed (Table 1). Included studies were full-text papers with human subjects and with the aim of developing or validating an instrument to measure symptoms of AD and the QOL of patients with AD. Instruments that had not been validated were ineligible.
Study selection

Inclusion criteria
(I) AD or eczema was included in the target population of the instrument.
(II) Instruments should be patient-reported outcomes or contain at least one domain that is self-reported.

Exclusion criteria
(I) Patient was not an accessor of the instrument (i.e., instrument was not self-reported).
(II) Instrument was not developed for measurement purposes.
(III) No psychometric validation was available for the instrument or the instrument had poor psychometric properties.
(IV) Instruments not for assessing the disease severity or the QOL of patients with AD.

Two reviewers (YTY and XXL) screened the titles and abstracts, and then assessed the full-text for eligibility.

Data extraction
All instruments that met the inclusion criteria were extracted and two types of information were recorded: basic information and instrument properties.

For the basic information, name, symptoms assessed, target population, assessor, number of items and components, rating method, score algorithm and available translation were listed. For the instrument properties, internal consistency, reliability measurement error, content validity, construct validity, cross-cultural validity, and responsiveness were recorded.

Assessment of measurement properties of instruments of AD

The measurement properties assessed in this study were selected based on the recommendations of the Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) group, which included reliability (internal consistency, reliability and measurement error), validity (content and construct validity) and responsiveness (29) (Table 2).

The instruments assessed were then placed in one of the three recommendation categories based on the sufficiency of their measurement properties as suggested by the COSMIN group (29). An instrument was placed in category A if there was evidence for sufficient content validity (any level) and at least low-quality evidence of sufficient internal consistency. An instrument was placed in category C if there was high-quality evidence of an insufficient measurement property. If an instrument could not be categorized as A or C, it was placed in category B. Instruments categorized as ‘A’ are recommended for use, and results obtained with these instruments can be seen as trustworthy. Instruments categorized as ‘B’ have the potential to be recommended for use, but require further validation. Instruments categorized as ‘C’ are not recommended for use. If only PROMs categorized as ‘B’ are found in a review, the one with the best evidence for content validity is the one to be provisionally recommended for use, until further evidence is found (29).

Results

Study characteristics

After filtering out all database a total number of 10,432 studies were evaluated, of which 9,349 were non-duplicate records according to the PRISMA statement (Figure 1) (30). After screening, 38 studies in English and 2 in Chinese were included in this systematic review. The 40 articles covered 26 instruments, which comprised 11 different instruments for disease severity and 15 for QOL that met the inclusion and exclusion criteria. All studies used at least 1 disease severity scale. Only 13 (37%) studies used QOL instruments and 9 (25%) studies used ≥1 QOL measurement. We also found 1 article for 1 instrument by manually searching the relevant reference.

Content of the English instruments identified

A total of 11 disease severity scales, including 3 single-item (unidimensional) Numeric Rating Scales (NRS) and 8 multidimensional scales, were utilized in the included studies. The most commonly used disease severity instrument in English was Peak Pruritus/Itch NRS which was used in more than 10 clinical studies (14,31,32). The second most common disease severity instrument was the Scoring Atopic Dermatitis Index (SCORAD), which was used in 7 studies (33). The next two commonly used disease severity instruments were the Patient-Oriented Eczema Measure (POEM) and the Objective SCORAD, both of which were used in 5 studies (33,34). These were closely followed by the Three Item Severity score (n=4) (35,36).
<table>
<thead>
<tr>
<th>Measurement property</th>
<th>Measurement property name</th>
<th>Criteria for adequate rating (+)</th>
<th>Criteria for intermediate rating (?)</th>
<th>Criteria for inadequate rating (−)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>Internal consistency</td>
<td>Cronbach’s α calculated per dimension AND Cronbach α 0.70–0.95</td>
<td>Unclear whether the instrument is unidimensional OR Doubtful design or method</td>
<td>Cronbach’s α not calculated per dimension despite being a unidimensional instrument OR Cronbach α &lt;0.70 or &gt;0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pearson’s R &gt;0.80 OR Weighted κ &gt;0.60 OR Coefficient of variation &lt;20% OR ANOVA &lt;10%</td>
<td>Reliability not evaluated OR (Pearson’s R 0.60–0.80 OR Weighted κ 0.40–0.60 OR Coefficient of variation 20–30% OR ANOVA 10–20%)</td>
<td>Pearson’s R &lt;0.60 OR Weighted κ &lt;0.40 OR Coefficient of variation &gt;30% OR ANOVA &gt;20%</td>
</tr>
<tr>
<td>Measurement error</td>
<td>Standard error of measurement (SEM), smallest detectable change (SDC) or limits of agreement was calculated OR Both positive and negative percentage agreement (PA) was calculated</td>
<td>SEM, SDC or LoA can be calculated from the given data OR PA was calculated</td>
<td>SEM was calculated based on Cronbach’s α or SD from another population AND PA was not calculated</td>
<td></td>
</tr>
<tr>
<td>Validity</td>
<td>Content validity</td>
<td>Professionals OR patients were involved in item selection AND Professionals AND patients considered &gt;90% of items to be relevant, comprehensive and understandable</td>
<td>Professionals OR patients were involved in item selection AND Professionals AND patients considered 70–89% of items to be relevant, comprehensive and understandable</td>
<td>NEITHER professionals NOR patients were involved in item selection OR professionals OR patients considered &lt;70% of items to be relevant, comprehensive and understandable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Factor analysis performed with adequate sample size (&gt;7-fold the number of items AND &gt;100) AND Two different instruments that aims to measure signs of AD show high correlation (correlation coefficient &gt;0.70)</td>
<td>No factor analysis OR factor analysis performed with intermediate sample size (&gt;5-fold the number of items AND &lt;100) OR Two different instruments that measure signs of AD show low correlation (correlation coefficient 0.60–0.69)</td>
<td>Factor analysis performed with inadequate sample size (&lt;5-fold the number of items) OR Two different instruments that measure signs of AD do not show correlation (correlation coefficient &lt;0.50)</td>
</tr>
<tr>
<td>Cross-cultural validity</td>
<td>Instrument functions in the same way in different translated versions across different samples of respondents</td>
<td>No translated versions</td>
<td>Instrument does not function in the same way in different translated versions across different samples of respondents</td>
<td></td>
</tr>
<tr>
<td>Responsiveness</td>
<td>The correlation between changes from baseline in a PRO score with changes from baseline in other PROs or outcomes</td>
<td>Moderate to strong correlation</td>
<td>Weak correlation</td>
<td>No correlation</td>
</tr>
<tr>
<td>Meaningful change estimation</td>
<td>Threshold</td>
<td>Receiver operating characteristic (ROC) curve was plotted OR [minimal important change (MIC) defined AND MIC &gt; smallest detectable change (SDC)]</td>
<td>MIC undefined</td>
<td>ROC curve was not plotted AND (MIC defined AND MIC ≤ SDC)</td>
</tr>
</tbody>
</table>

+, sufficient; −, insufficient (−); ?, indeterminate. PRO, patient-reported outcome; ANOVA, analysis of variance; AD, atopic dermatitis.
Identification of studies via database

Records identified from:
- Databases (PubMed 4,372; WoS 3,297; PsycINFO 69; ERIC 69; CNKI 1,332; Wanfang Data 1,293; total n=10,432)
- Manual searching (n=1)

Records removed before screening:
- Duplicate records removed (n=1,084)

Records screened (n=9,349)

Records excluded (n=8,256)

Reports excluded:
- Reason 1 (n=7)
- Reason 2 (n=2)
- Reason 3 (n=5)
- Reason 4 (n=3)
- (Duplicate n=2)*

Reports sought for retrieval (n=1,093)

Reports not retrieved (n=1,038)

Reports assessed for eligibility (n=55)

Studies included in review (n=40)

**Figure 1** Flow chart of the search procedure adapted from the 2020 PRISMA statement. Reason 1 refers to ‘Instruments were not patient reported outcomes or contain at least one domain that is self-reported’; Reason 2 refers to ‘Instrument was not developed for measurement purposes’; Reason 3 refers to ‘No psychometric validation was available for the instrument or the instrument had poor psychometric properties’; Reason 4 refers to ‘Instruments not for assessing the QOL or the disease severity or the QOL of patients with AD’. *, duplicate refers to overlapped studies across reason 1 to 4. QOL, quality of life; AD, atopic dermatitis.

No discernable trend was found in the use of the top 5 instruments and their publication years.

Twelve QOL instruments in the included studies were analysed. The most frequently used English QOL instrument was the Dermatology Life Quality Index (DLQI), which was used in 6 of the included studies (37). Following were the Children’s Dermatology Life Quality Index (CDLQI), Infant’s Dermatology Quality of Life Index (IDQOL) and Patient-Reported Outcomes Measurement Information System (PROMIS®) Itch Questionnaire (PIQ), all of which were used in 3 studies (19,38,39).

Intensive itch, skin pain and related sleeping disturbance are highly prevalent symptoms in the patients with AD that impact both physical and mental functioning. Three single-item NRS (Peak Pruritus/Itch NRS, Skin Pain NRS, and Sleep Disturbance NRS) are developed and validated to measure a specific aspect of AD severity that are meaningful to patients (31,32,40). The multidimensional disease severity instruments included 3–10 clinical signs. Among them erythema was the most frequently included item (mentioned in 7/8 disease severity instruments), followed by oedema (6/8) and lichenification (5/8). Table 3 presents the characteristics of the included instruments.

**Content of the Chinese instruments identified**

The 2 included studies in Chinese reported 13 different instruments to assess the clinical signs of AD. Of 13 instruments, 3 were originally developed in Chinese for disease severity in patients with AD, and 10 were Chinese versions of the Quality of Life Index for Atopic Dermatitis (QoLIAD), IDQOL. None of the self-developed original
<table>
<thead>
<tr>
<th>Instrument</th>
<th>Symptoms assessed</th>
<th>Target population</th>
<th>Assessor</th>
<th>No. of components (c) and/or items (i)</th>
<th>Rating method</th>
<th>Scoring algorithm</th>
<th>Available translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak Pruritus/Itch NRS</td>
<td>“on a scale of 0 to 10, with 0 being “no itch“ and 10 being “worst itch imaginable”, how would you rate your itch at the worst moment during the previous 24 hours?”</td>
<td>Moderate-to-severe AD</td>
<td>Patient</td>
<td>1 (i)</td>
<td>11-point NRS</td>
<td>Daily scores are averaged over 1-week interval</td>
<td>NA</td>
</tr>
<tr>
<td>Skin Pain NRS</td>
<td>The patients were asked to select a number from 0 (“no pain”) to 10 (“worst pain imaginable”) that best described the worst level of skin pain in the past 24 hours</td>
<td>Moderate-to-severe AD</td>
<td>Patient</td>
<td>1 (i)</td>
<td>11-point NRS</td>
<td>Daily scores are averaged over 1-week interval</td>
<td>NA</td>
</tr>
<tr>
<td>DS NRS</td>
<td>“On a scale of 0-10, with 0 being “no sleep loss related to the symptoms of AD” and 10 being “I did not sleep at all due to the symptoms of AD”, how would you rate your sleep last night?</td>
<td>Moderate-to-severe AD</td>
<td>Patient</td>
<td>1 (i)</td>
<td>11-point NRS</td>
<td>Daily scores are averaged over 1-week interval</td>
<td>NA</td>
</tr>
<tr>
<td>POEM</td>
<td>Frequency of pruritus, sleep disturbance, bleeding, weeping or oozing, cracking, flaking, dryness or roughness</td>
<td>Eczema</td>
<td>Patient or parent/caregiver</td>
<td>7 (i)</td>
<td>5-point VAS</td>
<td>Sum score (range, 0–28)</td>
<td>38 (including simplified and traditional Chinese)</td>
</tr>
<tr>
<td>TIS</td>
<td>Scoring of erythema, oedema and excoriation</td>
<td>AD</td>
<td>Patient</td>
<td>3 (i)</td>
<td>3-point NRS</td>
<td>Sum score (range, 0–9)</td>
<td>NA</td>
</tr>
<tr>
<td>PO-SCORAD</td>
<td>Affected surface area; severity of dryness, redness, swelling, crusting/oozing, scratching, thickening; severity of itching, sleep disturbance</td>
<td>AD</td>
<td>Patient or parent/caregiver</td>
<td>3 (c), 9 (i)</td>
<td>Component 1: shading and describing, physician estimates %; component 2: 4-point NRS; component 3: VAS; 100-mm VAS</td>
<td>NA</td>
<td>36 (including simplified and traditional Chinese)</td>
</tr>
<tr>
<td>ADIS</td>
<td>Pruritus (severity, timing); sleep disturbance</td>
<td>AD</td>
<td>Patient</td>
<td>2 (c), 8 (i)</td>
<td>10-point NRS and VRS</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>ADerm-SS</td>
<td>Itch during sleep hours, itch during awake hours, skin pain; intensity of skin cracking, pain caused by skin cracking, dry skin, skin flaking rash, skin thickening, bleeding, skin oozing</td>
<td>AD</td>
<td>Patient</td>
<td>2 (c), 11 (i)</td>
<td>11-point NRS</td>
<td>Sum score</td>
<td>NA</td>
</tr>
<tr>
<td>ADSS</td>
<td>Subjective symptoms (itching, sleep disturbance); objective signs (erythema, dryness, oozing, edema)</td>
<td>AD</td>
<td>Patient or parent/ caregiver</td>
<td>6 (i)</td>
<td>VAS</td>
<td>Sum score (range, 0–24)</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 3 (continued)
<table>
<thead>
<tr>
<th>Instrument</th>
<th>Symptoms assessed</th>
<th>Target population</th>
<th>Assessor</th>
<th>No. of components (c) and/or items (i)</th>
<th>Rating method</th>
<th>Scoring algorithm</th>
<th>Available translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rajka-Langeland severity score</td>
<td>Eczema extent, course, and intensity (sleeplessness)</td>
<td>AD</td>
<td>Patient</td>
<td>3 (c), 3 (i)</td>
<td>3-point VRS; 11-point NRS; VAS</td>
<td>Sum score</td>
<td>NA</td>
</tr>
<tr>
<td>ZRADSQ</td>
<td>AD symptoms (itching aggravated at night, itching, dry skin, itching accompanied by pain, burning skin, insomnia), heat (thirst, mouth dryness, constipation, dark urine), mood (fidgeting, irascibility)</td>
<td>AD</td>
<td>Patient or parent/caregiver</td>
<td>15 (i)</td>
<td>4-point VRS, 10 cm VAS</td>
<td>Sum score</td>
<td>Chinese, English</td>
</tr>
<tr>
<td>DLQI</td>
<td>Symptoms and feelings, daily activities, leisure, work and school, personal relationships, treatment</td>
<td>Patients with pruritus</td>
<td>Patient</td>
<td>10 (i)</td>
<td>4-point VRS</td>
<td>Sum score</td>
<td>range, 0–30</td>
</tr>
<tr>
<td>IDQOL</td>
<td>Itching and scratching, mood, time to sleep, sleep disturbances, disturbed playing, disturbed family activities, problems during meal times, problems from treatment, dressing problems, problems at bath time</td>
<td>AD</td>
<td>Parent/caregiver</td>
<td>10 (i)</td>
<td>3-point NRS</td>
<td>Sum score</td>
<td>range, 0–30</td>
</tr>
<tr>
<td>CDLQI</td>
<td>Symptoms and feelings, leisure, school or holidays, leisure, personal relationships, sleep, treatments</td>
<td>AD</td>
<td>Patient with the help of parent/caregiver</td>
<td>10 (i)</td>
<td>3-point NRS</td>
<td>Sum score</td>
<td>range, 0–30</td>
</tr>
<tr>
<td>PIQ-MS</td>
<td>General concerns, mood and sleep, clothing and physical activity, scratching behaviours</td>
<td>Patients with pruritus</td>
<td>Patient</td>
<td>4 (c), 63 (i)</td>
<td>4-point VRS</td>
<td>Sum score</td>
<td>NA</td>
</tr>
<tr>
<td>5-D itch</td>
<td>Duration, degree, direction, disability (sleep, leisure/social, housework/errands, work/school), distribution</td>
<td>Patients with pruritus</td>
<td>Patient</td>
<td>5 (c), 9 (i)</td>
<td>5-point VRS, check-box</td>
<td>Sum score</td>
<td>range, 5–25</td>
</tr>
<tr>
<td>ItchyQoL</td>
<td>Symptoms, functional limitations, emotions</td>
<td>Patients with pruritus</td>
<td>Patient</td>
<td>3 (c), 22 (i)</td>
<td>5-point VRS</td>
<td>Sum score</td>
<td>range, 22–110</td>
</tr>
<tr>
<td>SF-12</td>
<td>Limitations in physical activities, social activities, and usual role activities; bodily pain, general mental health, vitality, general health perceptions</td>
<td>General patients</td>
<td>Patient</td>
<td>12 (i)</td>
<td>VRS</td>
<td>Sum score</td>
<td>range, 123–220</td>
</tr>
<tr>
<td>Instrument</td>
<td>Symptoms assessed</td>
<td>Target population</td>
<td>Assessor</td>
<td>No. of components (c) and/or items (i)</td>
<td>Rating method</td>
<td>Scoring algorithm</td>
<td>Available translations</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------</td>
<td>----------------</td>
<td>----------------------------------------</td>
<td>-----------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>DFI</td>
<td>Housework, food, sleep, leisure, time shopping, expenditure, tiredness, distress, relationships, treatment</td>
<td>Eczema, AD</td>
<td>Parent</td>
<td>10 (i)</td>
<td>4-point VRS</td>
<td>Sum score (range, 0–30)</td>
<td>30 (including Chinese)</td>
</tr>
<tr>
<td>ABS</td>
<td>Family life, budget &amp; work, daily life, treatment</td>
<td>AD</td>
<td>Parent</td>
<td>14 (i)</td>
<td>6-point VRS</td>
<td>Sum score</td>
<td>NA</td>
</tr>
<tr>
<td>ABS-A</td>
<td>Daily life, economic constraints, care &amp; management of disease, work and stress</td>
<td>AD</td>
<td>Patient</td>
<td>18 (i)</td>
<td>6-point VRS</td>
<td>Sum score</td>
<td>NA</td>
</tr>
<tr>
<td>A Derm-IS</td>
<td>Sleep impact of AD (difficulty falling asleep, effect on sleep, bothersomeness of waking up at night); daily impacts of AD (limitation in household activities, physical activities, and social activities; difficulty concentrating, feeling self-conscious, feeling embarrassed, feeling sad)</td>
<td>AD</td>
<td>Patient</td>
<td>2 (c), 10 (i)</td>
<td>11-point NRS</td>
<td>Sum score</td>
<td>NA</td>
</tr>
<tr>
<td>ESS</td>
<td>Erythema; edema, induration, or papules; excoriation; oozing, weeping, or crusting; scaling; and lichenification</td>
<td>AD</td>
<td>Patient</td>
<td>6 (i)</td>
<td>3-point VRS, 6-point NRS</td>
<td>Proportional score (range, 0–6) multiply by sum score of clinical signs (range, 0–3); total range, 0–108</td>
<td>NA</td>
</tr>
<tr>
<td>PROMIS SD</td>
<td>Sleep disturbance</td>
<td>General patients</td>
<td>Patient</td>
<td>8 (i)</td>
<td>NRS</td>
<td>Sum score</td>
<td>NA</td>
</tr>
<tr>
<td>PROMIS SRI</td>
<td>Sleep-related impairment</td>
<td>General patients</td>
<td>Patient</td>
<td>8 (i)</td>
<td>NRS</td>
<td>Sum score</td>
<td>NA</td>
</tr>
<tr>
<td>ADCT</td>
<td>Symptom severity, itch, bother, impacts on AD sleep, daily activities, mood and emotions</td>
<td>Patient</td>
<td>6 (c), 6 (i)</td>
<td>5-point VRS</td>
<td>Sum score</td>
<td>50 (including simplified Mandarin)</td>
<td></td>
</tr>
</tbody>
</table>

AD, atopic dermatitis; VAS, Visual Analogur Scale; NRS, numeric rating scale; VRS, visual rating scale; NA, not applicable.
instruments had been properly tested for reliability, validity or responsiveness.

It’s worth noting that, the Quality of Life Scale for Chronic Eczema Patients-Prior Test Version (EQOLS) was the only instrument based on TCM features. Although it was initially tested for test-retest reliability, construct validity and criterion-related validity, no other studies providing reliable assessment of this instrument were found. The quality of this instrument was therefore not rated in this study.

**Psychometric properties of the instruments identified**

We summarized the psychometric properties of the 26 included instruments for AD and the recommendations were made based on the COSMIN checklist in Table 4. Three single-item NRS were reliable, valid, and responsive with a meaningful threshold (31,32,40). They were recommended to be in category A. Among the 23 multidimensional instruments, 9 (39.1%) were category A (34,35,38,39,41-57), 6 (26.1%) category B (36,48,58-62), and 8 (34.8%) were category C (37,39,46,51,54,63-78). None of the included multidimensional instruments has been proved to be sufficient in all of the assessed measurement properties. There was evidence for all of the assessed instruments for sufficient content validity in patients with AD. The most common measurement property considered to be insufficient was measurement error (found in 3 instruments), following by internal consistency and responsiveness (each found in 2 instruments).

Although further validation was needed for the assessed multidimensional instruments, it is still possible for the instruments in category A to be recommended, namely POEM, patient-oriented SCORAD (PO-SCORAD), Atopic Dermatitis Symptom Score (ADSS), IDQOL, CDLQI, PROMIS® Itch Questionnaire Mood and Sleep (PIQ-MS), PROMIS-Sleep Disturbance (PROMIS-SD) and PROMIS-Sleep-Related Impairment (PROMIS-SRI). Among those, POEM and PO-SCORAD were the most valid and reliable instruments to assess the disease severity of AD, and IDQOL and CDQOL were the most valid and reliable QOL instruments.

Both POEM and PO-SCORAD had adequate content validity (34,46,64) and were highly correlated (r=0.75–0.79) (43), so both are adequate in terms of construct validity. POEM and PO-SCORAD showed evidence for sufficient internal consistency, with the Cronbach α being 0.86–0.88 and 0.84, respectively (34,46). POEM showed evidence for sufficient responsiveness, as the area under the receiver operating characteristic (ROC) curve was 0.67 and the minimal clinically important difference (MCID) was 3.4 (44). Neither instrument was examined for reliability or measurement error (42).

The IDQOL and CDLQI showed high-quality evidence for adequate content validity and construct validity, and both were considered to be internally consistent, because the Cronbach α of IDQOL was 0.89 and that of CDLQI was 0.83–0.87 (49,52). Both instruments had adequate test-retest reliability, with the Spearman’s rank order correlation coefficient ranging from 0.74 to 0.97 for IDQOL and from 0.73 to 0.92 for CDQOL (49,53).

The quality of evidence for sufficient content validity, reliability and responsiveness was low for 4 instruments (ADSS, PIQ-MS, PROMIS-SD and PROMIS-SRI). Although these instruments were still placed in category A and therefore recommended, their measurement properties were all assessed in only one study per instrument.

Seven instruments [Objective SCORAD, TIS, Atopic Dermatitis Itch Scale (ADIS), Zheng-Related Atopic Dermatitis Symptom Questionnaire (ZRADSO), Atopic Dermatitis Symptom Scale (ADerm-SS), Atopic Dermatitis Impact Scale (ADerm-IS) and 5-dimensions itch scale (5-D itch)] were adequate in some measurement properties but their overall performance was unclear. It is still possible that these instruments could be recommended when more validation studies are available.

Nine instruments [SCORAD, Rajka-Langeland severity score, DLQI, Itchy Quality of Life (ItchyQoL), Short-Form 12 items (SF-12), Dermatitis Family Impact questionnaire (DFI), Atopic dermatitis Burden Scale (ABS), Atopic Dermatitis Burden Scale for Adults (ABS-A) and Epworth Sleepiness Scale (79)] had inadequate quality in at least one of the assessed measurement properties and therefore are not recommended for use in clinical settings. Note that in a study conducted by Liu et al. in 2016 (66), the Chinese translated version of DLQI showed poor fit to the Rasch model in Chinese patients, indicating insufficient structural validity for the translated instrument.

**Description of the recommended instruments**

The Peak Pruritus/Itch NRS is a single self-reported item designed to measure peak pruritus, or worst itch, over the past 24 hours based on the following question: “on a scale of 0 to 10, with 0 being “no itch” and 10 being “worst itch
Table 4 Measure properties of the instruments included and degrees of recommendation

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Internal consistency</th>
<th>Reliability</th>
<th>Measurement error</th>
<th>Content validity</th>
<th>Construct validity</th>
<th>Responsiveness</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak Pruritus/Itch NRS</td>
<td>NA</td>
<td>+ (Yosipovitch et al., 2019)</td>
<td>+ (Yosipovitch et al., 2019)</td>
<td>+ (Yosipovitch et al., 2019)</td>
<td>+ (Yosipovitch et al., 2019)</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>Skin Pain NRS</td>
<td>NA</td>
<td>+ (Silverberg et al., 2021)</td>
<td>+ (Silverberg et al., 2021)</td>
<td>+ (Newton et al., 2019)</td>
<td>+ (Silverberg et al., 2021)</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>DS NRS</td>
<td>NA</td>
<td>+ (Puelles, et al., 2022)</td>
<td>+ (Puelles, et al., 2022)</td>
<td>+ (Dias-Barbosa C et al., 2020)</td>
<td>+ (Puelles, et al., 2022)</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>POEM</td>
<td>+ (Charman et al., 2004; Gerbens et al., 2017; Silverberg et al., 2020)</td>
<td>? (Charman et al., 2004; Silverberg et al., 2017)</td>
<td>+ (Charman et al., 2004)</td>
<td>+ (Charman et al., 2004; Coutanceau &amp; Stalder, 2014)</td>
<td>+ (Charman et al., 2004)</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>TIS</td>
<td>−</td>
<td>? (Wolkerstorfer et al., 1999)</td>
<td></td>
<td></td>
<td>+ (Charman et al., 2005)</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>PO-SCORAD</td>
<td>+ (Stalder et al., 2011; Coutanceau &amp; Stalder, 2014; Silverberg et al., 2020)</td>
<td>? (Stalder et al., 2011; Silverberg et al., 2020)</td>
<td>+ (Stalder et al., 2011)</td>
<td>+ (Stalder et al., 2011; Silverberg et al., 2020)</td>
<td>+ (Stalder et al., 2011)</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>ADIS</td>
<td>+ (Martin et al., 2020)</td>
<td>+ (Martin et al., 2020)</td>
<td></td>
<td></td>
<td></td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>ADerm-SS</td>
<td>+ (Foley et al., 2019)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>ADSS</td>
<td>+ (Lee et al., 2018)</td>
<td>+ (Lee et al., 2018)</td>
<td></td>
<td>+ (Lee et al., 2018)</td>
<td>+ (Lee et al., 2018)</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>Rajka-Langeland severity score</td>
<td>− (Gånemo et al., 2016)</td>
<td>+ (Gånemo et al., 2016; Silverberg et al., 2020a)</td>
<td></td>
<td>+ (Rajka &amp; Langeland, 1989; Gånemo et al., 2016)</td>
<td>+ (Silverberg et al., 2020a)</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>ZRADSQ</td>
<td>+ (Wu et al., 2013)</td>
<td>? (Wu et al., 2013)</td>
<td></td>
<td>? (Wu et al., 2013)</td>
<td></td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>DLQI</td>
<td>+ (Wang et al., 2004)</td>
<td>+ (Finlay &amp; Khan, 1994)</td>
<td></td>
<td>(Finlay &amp; Khan, 1994; Liu et al., 2016)</td>
<td>+ (Shikiar et al., 2005; Wang et al., 2004)</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>IDQOL</td>
<td>+ (Neri et al., 2012)</td>
<td>+ (Lewis-Jones et al., 2001; van Valburg et al., 2011; Neri et al., 2012)</td>
<td></td>
<td>+ (Lewis-Jones et al., 2001; Neri et al., 2012; Wu et al., 2013)</td>
<td>+ (Lewis-Jones et al., 2001; Neri et al., 2012; Wu et al., 2013)</td>
<td>A</td>
<td></td>
</tr>
</tbody>
</table>

Table 4 (continued)
## Table 4 (continued)

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Internal consistency</th>
<th>Reliability</th>
<th>Measurement error</th>
<th>Content validity</th>
<th>Construct validity</th>
<th>Responsiveness</th>
<th>Recommendation*</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDLQI</td>
<td>+ (Ramírez-Anaya et al., 2010; Neri et al., 2012; Salek et al., 2013)</td>
<td>+ (Ramírez-Anaya et al., 2010; Salek et al., 2013)</td>
<td>+ (Lewis-Jones &amp; Finlay, 1995; Gabes &amp; Apfelbacher, 2021)</td>
<td>+ (Lewis-Jones &amp; Finlay, 1995; Ramírez-Anaya et al., 2010)</td>
<td>+ (Lewis-Jones &amp; Finlay, 1995; Ramírez-Anaya et al., 2010)</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>PIQ-MS</td>
<td>+ (Lei et al., 2020)</td>
<td>+ (Lei et al., 2020)</td>
<td>+ (Silverberg et al., 2020b)</td>
<td>+ (Lei et al., 2020)</td>
<td>+ (Lei et al., 2020)</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>5-D itch</td>
<td>? (Lin et al., 2019)</td>
<td>+ (Elman et al., 2010)</td>
<td>+ (Elman et al., 2010)</td>
<td>+ (Elman et al., 2010)</td>
<td>+ (Elman et al., 2010)</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>SF-12</td>
<td>+ (Huo et al., 2018)</td>
<td>+ (Cheak-Zamora et al., 2009)</td>
<td>+ (Cheak-Zamora et al., 2009)</td>
<td>+ (Cheak-Zamora et al., 2009)</td>
<td>+ (Cheak-Zamora et al., 2009)</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>ABS</td>
<td>+ (Méni et al., 2013)</td>
<td>- (Méni et al., 2013)</td>
<td>- (Méni et al., 2013)</td>
<td>+ (Méni et al., 2013)</td>
<td>+ (Méni et al., 2013)</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>ADerm-IS</td>
<td></td>
<td></td>
<td></td>
<td>+ (Foley et al., 2019)</td>
<td>+ (Foley et al., 2019)</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>ESS</td>
<td>+ (Manzar et al., 2019; Lei et al., 2020b)</td>
<td>? (Lei et al., 2020b)</td>
<td>+ (Crook et al., 2019)</td>
<td>+ (Johns et al., 1991)</td>
<td>+ (Johns et al., 1991)</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>PROMIS SD</td>
<td>+ (Lei et al., 2020a)</td>
<td>+ (Lei et al., 2020a)</td>
<td>+ (Li et al., 2018)</td>
<td>+ (Li et al., 2018; Lei et al., 2020a)</td>
<td>+ (Li et al., 2018; Lei et al., 2020a)</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>PROMIS SRI</td>
<td>+ (Lei et al., 2020a)</td>
<td>+ (Lei et al., 2020a)</td>
<td>+ (Li et al., 2018)</td>
<td>+ (Li et al., 2018; Lei et al., 2020a)</td>
<td>+ (Li et al., 2018; Lei et al., 2020a)</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>ADCT</td>
<td>+ (Simpson et al., 2019)</td>
<td>+ (Simpson et al., 2019)</td>
<td>+ (Pariser et al., 2020)</td>
<td>+ (Simpson et al., 2019)</td>
<td>+ (Simpson et al., 2019)</td>
<td>A</td>
<td></td>
</tr>
</tbody>
</table>

* category of recommendation: A, recommended; B, will recommended with more evidence; C, not recommended; +, sufficient; -, insufficient (-); ?, indeterminate; NA, not applicable
imaginary”, how would you rate your itch at the worst moment during the previous 24 hours?” (31,32,35). The single-item Skin Pain NRS assesses self-reported severity of worst skin pain each day. The patients are asked to select a number from 0 (“no pain”) to 10 (“worst pain imaginable”) that best described the worst level of skin pain in the past 24 hours (32,35). The current versions of Peak Pruritus/Itch NRS and Skin Pain NRS are easy for moderate-to-severe AD patients to understand and answer compared to earlier versions of assessing average itch (rather than worst itch) or other types of pain (rather than skin pain), or other point scale (rather than 11-point scale), or longer recall (rather than the past 24 hours). Similarly, Sleep Disturbance (SD) NRS asks the AD patients: “On a scale of 0-10, with 0 being “no sleep loss related to the symptoms of AD” and 10 being “I did not sleep at all due to the symptoms of AD”, how would you rate your sleep last night (32,41)? All the three NRS are validated in multiple languages. It can be completed by most patients in less than 2 minutes while also providing a more comprehensive view of disease severity than measurement of itch and/or sleep disturbance alone (34). However, the reliability and measurement error for POEM still require further investigation. POEM is commonly used in clinical trials and everyday practice as a subjective measurement to describe the disease outcome. Often together with physician-assessed objective outcome measurements such as EASI, POEM has been used by thousands of patients among all age groups in clinical trials of drug development for AD (79,80,81).

PO-SCORAD is a patient-oriented instrument derived from SCORAD, published in 2010 (52). It assesses 3 components of AD: the affected body surface area (BSA), the severity of clinical signs, and other clinical symptoms. The affected BSA is calculated as a percentage of each defined body area and reported as the sum of all areas. It is validated and internally consistent. One advantage of PO-SCORAD is that it provides visual explanations that are understandable by patients regardless of their age, which therefore improves the accuracy of the instrument (45). Also, PO-SCORAD measures both subjectively and objectively, which may minimize the bias caused by any misunderstanding between patients and physicians (45). The disadvantages of PO-SCORAD include lack of evidence supporting its reliability, and assessment of the measurement error and responsiveness. PO-SCORAD is also widely used in clinical trials for drug development and therapy evaluation in various populations including Chinese (82-84).

IDQOL is an instrument designed to assess the QOL in children (<4 years old) with AD from the parental view, and was published in 2001 (39). It has two parts: dermatitis severity and quality of life index. It is been proved to have adequate internal consistency. IDQOL can be easily used by parents with or without another assessment of clinical severity. Although adequate, more research needs to be done to further evaluate the content validity and responsiveness to increase confidence in using IDQOL in clinical settings. IDQOL and CDLQI are often used together to evaluate the effects of treatment in children in clinical trials (85,86). These instruments were also used in several studies investigating the effect of AD on QOL of the children and their caregivers in various languages and populations (87-89).

CDLQI is an instrument designed to measure the effect of skin disease on children's QOL, published in 1995 (38). Unlike the abovementioned instruments, CDLQI is a generic instrument for skin and connective tissue diseases.
There is also a DLQI version for adults and a family version (FDLQI). Similar to those instruments, the CDLQI is a 10-item questionnaire that assesses 6 different aspects (symptoms and feelings, leisure, school or holidays, personal relationships, sleep, treatment) that may affect a child’s QOL. It has cartoon illustrations based on the theme of every question, aiming to be more user-friendly for younger children. CDLQI has 131 translated versions and has been verified for reliability, interpretability and cross-cultural validity multiple times, so it can be easily adapted to different culture groups using different languages. However, as a generic instrument, the wording of the questions in CDLQI may lack precision for the effects of AD.

Discussion

This systematic review identified, summarized, and assessed the measurement properties of 26 different instruments used in the literature since 2010 to assess the disease severity and QOL of AD. Three single-item NRS (Peak Pruritus/Itch, Skin Pain and SD) demonstrate good reliability, validity and responsiveness and can measure day-to-day fluctuations in a specific aspect of disease severity related to AD (31,32,40). With determined minimal important changes, these single-item NRS are easy-to-interpret in clinical trials and are comparative among studies. However, AD is a characteristic of a variety of symptoms. Single-item NRS is an important complement to multidimensional scales as an outcome in clinical studies. Among multidimensional scales, only 4 instruments, namely POEM, PO-SCORAD, IDQOL and CDLQI, had adequate content and construct validity, internal consistency and reliability, but unclear measurement error, responsiveness, interpretability and feasibility. These are recommended to be used in clinical settings according to the COSMIN guidelines. Nine of the assessed instruments reported insufficiency in at least 1 measurement property and therefore are not recommended until future studies are available.

Nowadays, the diagnosis of AD is mostly based on clinical criteria, namely the historical features, morphology and distribution of skin lesions, and associated clinical signs (90,91). One of the earliest and most recognized sets of diagnostic criteria, the Rudzki criteria, has been used and validated in clinical practice for more than 40 years (92). With systematic modifications intended to provide a tool for researchers who are not dermatologists, the UK Working Party diagnostics were developed and are widely used and assessed. Both of these diagnostic schemes have been validated in a wide range of age groups, languages and populations (93). However, neither disease severity nor QOL measurement scales are commonly used, or even recommended, for routine clinical practice, as suggested by Eichenfield et al. in 2014 (94).

However, in current clinical trials and drug development procedures, disease severity and QOL measurement scales are getting increasing attention and are often used as endpoint measures (90). Some disease severity scales, such as EASI, have been used extensively as the primary or secondary endpoint for large clinical drug trials (e.g., dupilumab, nemolizumab and tezepelumab) (18,95). In those studies, most of the patients had a 50% reduction in the EASI score after 12 weeks of treatment (18,95). Other endpoints used in clinical trials include the SCORAD and 5-D itch scales (56,96).

Recognizing the lack of generality, uniformity and accessibility of AD measurement scales, the development, validation and standardization of the outcome methodology requires greater attention for a golden standard to be established. Few studies have covered different AD patient populations, especially with respect to cross-cultural equivalence, age groups and sex. In terms of content uniformity, a systematic review conducted by Schmitt et al. in 2007 (97) demonstrated substantial heterogeneity in the domains included in the different outcomes, the items used to measure the domains, the relative weights of the domains in the summary score, the scales used to measure the items, and the person performing the assessment. This phenomenon could lead to unfeasible comparison between clinical trials, and therefore misunderstanding and biased results when used in clinical settings. The accessibility and acceptance of AD measurement scales, especially for young patients, also requires attention. Although many measures were originally developed to be plain text-based for paper-and-pencil administration, some of them (e.g., CDLQI) do have illustrations to aid in understanding and could be easily adapted to electronic format. Moreover, electronic adaptation of existing measure scales may bring less administrative burden, higher patient acceptance rate, less secondary data entry errors, and more accurate and complete data (98). Migration of existing validated AD measurement scales to electronic platforms may help to improve the quality and accessibility of future clinical investigations, considering that some AD scales may involve assessment of affected BSA, severity of dryness, redness, swelling, crusting and oozing. It can be difficult for the patients to understand text descriptions, so if the electronic
version of these instruments includes illustrations, it will help patients to determine the progression of their disease more accurately and help researchers make judgments about the effectiveness of the drugs more precisely.

This study has several strengths and limitations. We used a highly sensitive search strategy for study collection instead of regular search strings. We also took advantage of 6 databases in English and Chinese, minimizing relevant information from being missed. Furthermore, we applied the validated COSMIN checklist methodology to rate the studies’ quality and classified them in a systematic way based on predefined criteria.

In conclusion, as the treatment of AD is developing rapidly, the requirement for instruments to measure both disease severity and QOL for patients with AD is urgent. Therefore, in this review, we systematically analyzed 23 instruments utilized in the clinical trials from 2010 to 2021 (8 for disease severity, 15 for QOL). Of them, 9 instruments with significant reality and validity properties were recommended for further application. We also searched for the Chinese adapted versions of these 23 instruments and for instruments originally developed in Chinese that satisfied the inclusion criteria. However, there were too few Chinese AD instruments and their validation was inadequate. Therefore, further studies are needed to develop more original Chinese instruments for patients with AD.

Acknowledgments

Funding: None.

Footnote

Reporting Checklist: The authors have completed the PRISMA reporting checklist. Available at https://atm.amegroups.com/article/view/10.21037/atm-22-3164/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://atm.amegroups.com/article/view/10.21037/atm-22-3164/coif). AL, MZ, JZ, and XX are from Sanofi (China) Investment. YY and XL are from Shanghai Jusure Health Technology. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

10. Simpson E, Bissonnette R, Eichenfield LF, et al. The


58. Charnan CR, Venn AJ, Williams H. Measuring atopic eczema severity visually: which variables are most important to patients? Arch Dermatol 2005;141;1146-51; discussion 1151.

Dermatolog Treat 2020;31:484-90.


