Outcome Measures of Disease Severity in Atopic Eczema

Carolyn Charman, BM, BCh, MRCP; Hywel Williams, PhD, FRCP

**Background:** An essential component of evidence-based medicine is the use of valid and reliable outcome measures in clinical trials. There is much confusion in the field of atopic eczema regarding how to best measure disease severity objectively.

**Objective:** To establish the extent to which existing objective clinical scales for atopic eczema have been tested for validity, reliability, sensitivity to change, and acceptability.

**Design:** An electronic bibliographic search was performed for published data on all currently available named atopic eczema scales.

**Results:** Thirteen scales were identified in total. Data on construct or criterion validity were available for 10 scales. Only 5 scales had been tested for reliability (interobserver variability, intraobserver variability, or internal consistency). Data on responsiveness to change were available for 8 scales. An estimated time to administer the measure had been given for 3 scales. The only severity scale for which published data could be found on validity, reliability, sensitivity, and acceptability testing was the Severity Scoring of Atopic Dermatitis index, although problems occurred with interobserver variation of the index.

**Conclusion:** The rapidly increasing number of severity scales for atopic eczema, many of which have been inadequately tested, has made the interpretation of patient outcomes confusing, and comparison of results between studies almost impossible. Consensus among clinicians and researchers on the use of severity scales for atopic eczema should be based on evidence of adequate validity, reliability, sensitivity to change, and ease of use.

Arch Dermatol. 2000;136:763-769

**Accurate and appropriate measurements of health outcome form the basis of good evidence-based practice.** Methods of assessing the effects of health care intervention are continually being sought, both for use in clinical trials and medical audit, and for guiding the allocation of health service resources.\(^1\) Assessing disease severity in dermatological disorders can present problems because laboratory tests of disease severity often do not exist. In chronic inflammatory dermatological conditions such as atopic eczema, subjective criteria such as itching and sleep disturbance may be the most useful indicators of disease severity and patient morbidity. Other measures such as quality of life measurements, which assess social well-being in addition to personal health status, are being increasingly recognized as important outcome measures, but at present they are largely used to supplement existing measurement tools in dermatology.\(^2\) In addition to these patient-assessed outcome measures, objective clinical scoring of disease severity for features such as erythema, exudation, and excoriations may provide important additional information that is less subject to social or cultural factors. Such objective clinical scoring forms the basis of many existing severity scales in atopic eczema.

**For editorial comment see page 788**

Although atopic eczema is a common inflammatory skin disease that physicians must frequently assess, both in everyday clinical practice and in clinical trials, there remains a lack of standardization of objective severity scales for the disease. One problem in the field of therapy assessment is the rapidly increasing profusion of scales, many of which have been developed without proper testing. This has made the interpretation of patient outcome confusing and the comparison of results between different studies almost impossible. Many different clinical features have been pro-
posited as important indicators of disease intensity in atopic eczema, and deciding which of these to include in any measurement scale is difficult. Hence, there is wide variation in the number and type of clinical features included in different scales. Some clinical features such as erythema, edema, papulation, and vesiculation occur primarily in the acute phase of the disease, whereas lichenification, dryness, and desquamation may be better appreciated in chronic disease. Many other features such as cracking, excoriations, and crusting are frequently included in measurement scales, along with assessments of disease extent or percentage of body surface area affected. Deciding whether to assess the most severely affected area of eczema, or whether to grade a representative body site for each clinical feature is also a problem. Assessing eczema intensity at several body sites may give a better representation of overall disease intensity, although the optimum number of body sites for assessment remains unclear.

Although the precise outcome measures chosen will depend on many factors such as patient group, body site, and trial duration, any measure should ideally be tested for validity and reliability (Table 1), responsiveness to change, and acceptability before being adopted into clinical practice.¹ There is varying terminology surrounding the assessment of validity, although it is traditionally subdivided into content, criterion, and construct validity.² However, all types of validity address the degree of confidence that can be placed on inferences drawn from the measurement scale. Testing for construct validity in atopic eczema may involve evaluating how closely the new scale is related to requirements for topical corticosteroids, days off work, or visits to the physician. Assessing criterion validity involves correlating the new scale with other existing measures of severity, and in the absence of a “gold standard” measure, other established indices or global assessments of disease severity may be used. When considering reliability testing, analysis of chance-corrected agreement can be used to examine variations in scoring between and within observers, and methods such as Cronbach α or factor analysis can give information about internal consistency.³ The ability of the instrument to detect clinically meaningful changes in disease severity in response to health care intervention is also essential, particularly in short-term clinical trials. Acceptability can be assessed in terms of both patient and observer acceptability. Clinical eczema scoring systems in current use are generally acceptable to patients as they involve a simple examination only, but observer acceptability in terms of ease of administration is also important, especially for measures intended to be used in a busy clinic setting.

The various measurement scales for assessing atopic eczema severity have been subject to detailed descriptive review.⁴ This review described the various scales in use at the time (up to 1996) and made some qualitative recommendations about their use. More recently, 4 further scales have become available (the Six-Area, Six-Sign Atopic Dermatitis Severity Score,⁵ the Nottingham Eczema Severity Score,⁶ the Eczema Area and Severity Index [EASI],⁷ and the Assessment Measure for Atopic Dermatitis [ADAM]).⁸⁹

The aim of this article is to establish the extent to which existing disease-specific objective skin examination scales for atopic eczema satisfy the desirable attributes of a “good scale.” We quantitatively examine the various disease-specific atopic eczema severity scoring systems currently available with respect to validity, reliability, responsiveness, and acceptability testing. We also discuss the quality of such testing when information is available.

### RESULTS

#### QUANTITATIVE SUMMARY

Thirteen scales were identified in total over the last 10 years, an average of 1.3 new scales per year (Table 2).¹⁰-⁴⁵

### METHODS

An electronic bibliographic search was carried out for all published data relating to objective atopic eczema severity scales in MEDLINE from January 1966 to April 1999 and EMBASE from January 1988 to April 1999. Specific search terms used were “atopic dermatitis,” “atopic eczema,” “severity of illness index,” “severity,” “outcome measures,” and “scores.” All “named” scales specifically designed for atopic dermatitis were included in the review. Identified scale names were used to search for further data on scale use and testing. Review articles and references to other scales were also pursued. Outcome measures were based on the extent to which each scale satisfied a quality rating assessment based on validity (content, construct, criterion), reliability (internal consistency and interobserver and intraobserver variability), sensitivity to change, and ease of use/acceptability.

### TABLE 1. Properties of Measurement Scales

<table>
<thead>
<tr>
<th>Property of Scale</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>Does the scale measure what it is intended to measure?</td>
</tr>
<tr>
<td>Content</td>
<td>Does the scale appear to be assessing all the relevant content or domains, based on judgment by 1 or more experts?</td>
</tr>
<tr>
<td>Construct</td>
<td>Does the scale agree with other related variables and measures of the same construct with which, in theory, it ought to agree?</td>
</tr>
<tr>
<td>Criterion</td>
<td>Does the scale correlate with some other measure of the disease, ideally a “gold standard” that has been used and accepted in the field?</td>
</tr>
<tr>
<td>Reliability</td>
<td>Does the scale measure what it is intended to measure in a reproducible fashion?</td>
</tr>
<tr>
<td>Interobserver</td>
<td>Do measurements made by 2 or more observers produce the same or similar results?</td>
</tr>
<tr>
<td>Intraobserver</td>
<td>Do measurements made by the same observer on 2 or more occasions produce the same or similar results?</td>
</tr>
<tr>
<td>Internal</td>
<td>Do the scores from different items in the scale correlate with each other and with the total scale score, ie, are all items in the scale measuring the same attribute?</td>
</tr>
<tr>
<td>Consistency</td>
<td>Internal consistency</td>
</tr>
</tbody>
</table>

©2000 American Medical Association. All rights reserved.
Of these 13, the only severity scale for which published data could be found on validity, reliability, sensitivity, and acceptability testing was the Severity Scoring of Atopic Dermatitis (SCORAD) index.35 Most scales have been validated for content by expert dermatologists. Seven scales show some evidence of criterion validity against investigator or patient global assessments of disease severity.5,6,8,17,22,23,44 Evidence of construct validity, usually against topical steroid use or soluble markers of disease activity have been demonstrated for 9 scales.3,6,9,19,22,24,31,41,44, Published data on interobserver variability were available for 5 scales,7,8,17,19,31 but only 2 scales were tested for intraobserver variability.7,31 Two scales had some data on variability.7,31 Although many scales were reported as being quickly and easily used, a guiding estimate of the time required to carry out the measurement by either trained or untrained observers was given for only 3 scales.5,30,31.

In addition to the 13 named scales identified, many other studies had used a variety of clinical parameters such as erythema, crusting, excoriations, lichenification, scaling, edema, induration, dryness, and infiltration. Parameters were usually scored on a scale of 0 to 3, and various combinations of parameters were added together to form a number of unnamed severity scores, which will not be discussed further in this article.

### QUALITATIVE SUMMARY

The SCORAD Index

The most extensively tested severity index in our search was the SCORAD index. This composite scoring index was developed by the European Task Force on Atopic Dermatitis in 1993.36 It has undergone testing for validity and reliability and has shown sensitivity to change in trials of cyclosporin,35 topical steroids,39 and UV-A therapy.40 It combines an assessment of disease extent using the rule of nines with 6 clinical features of disease intensity (assessed at a single representative site), plus a visual analogue score for itch and sleep loss, which may be excluded for clinical trial work.31 The index has shown agreement with global assessments of disease severity35 as well as with various circulatory factors thought to reflect disease activity in atopic dermatitis.22-24 However, problems with interobserver variation have occurred. In the original description of the scoring system, reliability was tested using 10 trained observers and 10 slides.35 Significant interobserver variability occurred for edema, oozing, lichenification, and total score. Intraobserver variation for disease intensity showed a minimum of 70% probability that 2 slides with the same severity for 1 particular item would be scored identically by the same physician. Significant interobserver variability has also been shown for recording lichenification and excoriations us-

### Table 2. Published Quality Criteria for Severity Scoring Systems in Atopic Dermatitis

<table>
<thead>
<tr>
<th>Scale and Year of Publication</th>
<th>Validity</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAM 19959</td>
<td>(8,9)</td>
<td>(8)</td>
</tr>
<tr>
<td>ADASI 199110,11</td>
<td>(10)</td>
<td>(12-14)</td>
</tr>
<tr>
<td>ADSI 199515</td>
<td>(15,16)</td>
<td></td>
</tr>
<tr>
<td>BCSS 199517</td>
<td>(17)</td>
<td>(17)</td>
</tr>
<tr>
<td>Costa’s SSS 198919</td>
<td>(19)</td>
<td>(18,19)</td>
</tr>
<tr>
<td>EASI 19987</td>
<td>(7)</td>
<td>(21)</td>
</tr>
<tr>
<td>Leicester 199322</td>
<td>(22,23)</td>
<td>(7)</td>
</tr>
<tr>
<td>Nottingham Eczema Severity Score 199636</td>
<td>(6)</td>
<td>(7)</td>
</tr>
<tr>
<td>Rajka and Langeland 198924</td>
<td>(24)</td>
<td>(7)</td>
</tr>
<tr>
<td>SASSAD 19963</td>
<td>(5)</td>
<td>(5,26-28,30)</td>
</tr>
<tr>
<td>SCORAD 199331</td>
<td>(31)</td>
<td>(35,39,40)</td>
</tr>
<tr>
<td>SIS 199282</td>
<td>(41-43)</td>
<td></td>
</tr>
<tr>
<td>TBSA 199244</td>
<td>(44-45)</td>
<td></td>
</tr>
</tbody>
</table>

* indicates data available on quality criteria (see “Comment” section for further information); ○, trends suggesting construct or criterion validity (eg, in graph format or figures for before and after treatment) but no correlation statistics given. Numbers in parentheses are reference numbers. For full study names see indicated references.

©2000 American Medical Association. All rights reserved.
ing SCORAD in an epidemiological study and in a multinational randomized trial. In the latter study, reliability was tested using 3 members of the European Task Force and 98 observers using selected photographs of patients. It showed approximately 30% of observers' scores to be outside the range of the 3 experts. Surface area assessments ranged from 20% to 100% in 1 of 3 sets of photographs assessed for disease extent. Other studies have also demonstrated the difficulties of estimating body surface area involvement in atopic eczema. Further results of reliability testing using SCORAD have been published more recently using 19 patients (rather than photographs) and 12 observers. Again, large interobserver variations in assessing disease extent (especially for patients with 20% to 60% body surface area involvement) and intensity items (especially lichenification) occurred. Variations in the choice of a representative site for assessing disease intensity were thought to have contributed to these results. A beneficial effect of observer training was suggested in this study (but remains to be confirmed in controlled studies), and hence an instructive CD-ROM has been developed for training purposes. However, a recent study using 34 patients and 2 trained observers (who had practiced SCORAD and reviewed a SCORAD training atlas) still showed statistically significant interobserver variation for edema and/or infiltration and/or papulation assessed for disease extent on a scale of 0 to 6 in 4 defined body sites on a scale of 0 to 3, giving a maximum score of 108. All 6 recorded signs have been shown to change in parallel during treatment of childhood atopic dermatitis with cyclosporin. The score has shown agreement with patient and observer global assessments of disease severity, patient-assessed itch and sleep loss, and topical steroid requirements, although it has correlated poorly with quality of life parameters. It avoids an assessment of surface area involvement, the poor reliability of which has been demonstrated. Although it is said to be simply and quickly performed without previous training, there are no published data on reliability testing. Because the index assesses many of the clinical features used in the SCORAD index, but at more body sites, one might expect interobserver variation to be substantial.

The SASSAD Index

The Six-Area, Six-Sign Atopic Dermatitis (SASSAD) severity index involves assessing 6 clinical features of disease intensity (erythema, exudation, excoriation, dryness, cracking, and lichenification) at 6 defined body sites on a scale of 0 to 3, giving a maximum score of 108. All 6 recorded signs have been shown to change in parallel during treatment of childhood atopic dermatitis with cyclosporin. The score has shown agreement with patient and observer global assessments of disease severity, patient-assessed itch and sleep loss, and topical steroid requirements, although it has correlated poorly with quality of life parameters. It avoids an assessment of surface area involvement, the poor reliability of which has been demonstrated. Although it is said to be simply and quickly performed without previous training, there are no published data on reliability testing. Because the index assesses many of the clinical features used in the SCORAD index, but at more body sites, one might expect interobserver variation to be substantial.

The Leicester score is an earlier version of the SASSAD index that originally involved assessing 10 body zones for erythema, excoriation, dryness, cracking, and lichenification, giving a maximum score of 150. Although the Leicester score has shown some evidence of construct validity against topical steroid requirements and patient symptom diaries, the score has been superseded by the SASSAD index.

A modified version of the SASSAD known as the 6-Area Total Body Severity Assessment (TBSA) has also been described. The TBSA, which has a maximum score of 108, differs from SASSAD in that it assesses infiltration and vesicles and/or papules, and excludes lichenification. It has shown evidence of construct validity against itch, sleep loss, and histological changes, as well as criterion validity against patient and investigator global measures of disease severity, but has been less widely tested than the SASSAD. Confusingly, the name 6-Area TBSA has also been used in a study using the SASSAD.

The ADASI

The Atopic Dermatitis Area and Severity Index (ADASI) uses a color-coding system of body charts and point counting to produce a scale that has shown sensitivity to change in trials of various dietary treatments in atopic eczema. However, drawing the extent of the disease onto body charts using 3 different colors may be difficult to reproduce among different observers and with a single observer at different times. Formal reliability testing of the index is not available. The name ADASI has also been applied to a modified version of the Psoriasis Area and Severity Index (PASI), although the ADASI score differs from that described above.

The EASI

The Eczema Area and Severity Index (EASI) involves an assessment of disease extent on a scale of 0 to 6 in 4 defined body regions plus an assessment of erythema, infiltration and/or papulation, excoriation, and lichenification each on a scale of 0 to 3. A formula is then used to calculate the total score for each of the 4 regions, which are then added together. In a recently published abstract, overall interobserver consistency was good using 15 observers (with 30 minutes of training) and 20 patients. However, the chance-corrected correlation coefficient (κ statistic) for individual variables showed only fair agreement for the assessment of infiltration and/or papulation (κ = 0.23-0.27). Assessments took 5 minutes per patient. A modified version of the score has shown evidence of sensitivity to change and criterion validity against physician and patient global severity ratings, although patient assessment of pruritus was included in the total score in this case.

The Rajka and Langeland Scoring System

The Rajka and Langeland scoring system is a simple scale measuring clinical course, intensity, and extent of atopic eczema that was published in abstract form in 1989. It was originally presented as a broad, free-standing indicator of atopic eczema severity status, both long-term and at present. It is probably most suitable for baseline categorization of patients rather than to monitor severity changes in trials, because its broad categories (maximum score, 9) mean that only large changes in clinical condition can be detected. Published data on the system's responsiveness to change are not available. The scale involves an assessment of body surface area involvement, albeit into 1 of 3 categories only, and lacks any formal reliability testing. The score has shown correlation with a soluble immunological marker thought to reflect disease activity in atopic eczema, although the categori-
zation of total scores in this study was slightly different from that originally described.23 The recently described refined version of the index (Nottingham Eczema Severity Score) uses a 5-point rather than a 3-point grading system for clinical course and intensity, giving the potential for increased sensitivity to change while still being easy to administer.8 It still includes a measure of disease extent but uses a tick-box system corresponding to sites commonly affected by atopic eczema to simplify the assessment. The modified scale has shown criterion and construct validity in a community sample of 290 children with respect to clinical severity assessed by a dermatologist, parental severity assessment, potency of topical steroid use, and impairment of quality of life.9 However, reliability testing has yet to be carried out.

The SSS

Costa’s19 Simple Scoring System (SSS) scores 10 severity criteria (0-7) and 10 topographic sites (0-3) giving a maximum score of 100. In the original description of the scale, no significant interobserver variability occurred using 2 observers and 7 patients. However, a further study using 2 observers and 34 patients showed significant interobserver variation in assessing excoriations and scales.18 Sensitivity to long-term change has been demonstrated.21 The system has shown poor agreement with the SCORAD index, possibly because it evaluates disease severity in the most severely affected site rather than in a representative area as is used in SCORAD.18 Various modified versions of Costa’s scale have been described but not formally tested.19-31 Although it is described as a “simple” scoring system, the recording of 100 pieces of information on each patient suggests that it is anything but simple to administer.

The BCSS

The Basic Clinical Scoring System (BCSS) is a simple scale that assesses the presence or absence of disease in 5 body sites giving a total score of 5.17 It has been used in outpatient clinics and in primary health care settings and has shown excellent agreement between observers.18 In a study testing 1 observer it showed poor agreement with both the SCORAD index and Costa’s SSS.18 Responsiveness to change over 10 to 12 weeks has been demonstrated in 1 field study,17 but its ability to detect clinical meaningful change in short-term clinical trials has not been shown.

The ADSI

The Atopic Dermatitis Severity Index (ADSI) comprises an assessment of erythema, pruritus, exudation, excoration, and lichenification, each on a scale of 0 to 3 to give a maximum score of 15.15 Content validity of these parameters was originally proposed by Hanifin,16 and sensitivity to change has been demonstrated in a double-blind, placebo-controlled, right-and-left comparison study using 34 patients with atopic dermatitis.15 However, we could find no data on criterion or construct validity or reliability testing of the index.

The SIS

The Skin Intensity Score (SIS) assesses itching, erythema, and lichenification (0-10; maximum score, 30), and has shown construct validity against soluble markers thought to reflect disease activity.41-43 Although sensitivity to change has been demonstrated, further quality testing data were not available.

The ADAM

The Assessment Measure for Atopic Dermatitis (ADAM) is the most recently developed eczema-specific outcome measure to be described.44-46 It comprises an assessment of 6 body areas for scale and/or dryness, lichenification, erythema, and excoriations (0-3), and 4 further body areas for the presence or absence of eczema. A log transformation and complex mathematical model have been used in the derivation of the scoring system, but these are not required to be used in administration of the scale. The score has also generated “word pictures,” which have been recommended as operational definitions of grades of severity. Reliability, tested using 2 physicians and 51 patients, has shown variable agreement for individual elements of the score, with better agreement for mild cases than severe. Criterion validity against physicians’ global severity ratings showed marginal agreement (κ=0.4). Data on the sensitivity of the index to change have not yet been published.

COMMENT

Thirteen atopic eczema severity scales have been described, but nearly all have not been adequately tested. It is also important to point out that just because a measure has been tested for the various attributes of a good scale, this does not mean that it is a good scale. For instance in the SCORAD index, serious interobserver variability has been demonstrated for many aspects of the scale. Although intraobserver variation may be more important in small single-center trials, in practice many clinical trials involve more than 1 center and hence consistency between observers is extremely important. It is therefore important to assess how well individual scales have performed as well as simply whether the scale has been tested.

The large number of severity scales available for atopic eczema partly reflects the varying requirements of scales for different clinical situations. For example, indices such as SCORAD and SASSAD, which use an assessment of various combinations of clinical signs of atopic eczema in several body sites, result in wide scales that are probably best at detecting small changes in disease activity that might be useful in clinical trials. Other scales such as the Rajka and Langeland24 proposal and its refinement or the BCSS17 might be more suitable for broadly categorizing patients or for rapid assessments of disease severity for use in epidemiological studies or in health services research.

In addition to the concept of different scales for different purposes, the properties of such scales require some consideration. It is unlikely that most scales are linear.
The profusion of atopic eczema scales described in this article probably represents only the tip of the iceberg; there are many other unnamed scales used in clinical trials that seem to simply shuffle around the various combinations of signs and sites. The array of scales reflects the fact that no measurement tool is able to perfectly reflect the disease state of a patient. In most of the scoring systems discussed, content validity has been judged by dermatologists, whereas the clinical relevance to patients of many of the factors included in these scores remains largely unknown.

The development of indices that accurately reflect the morbidity that skin diseases cause in a way that clinicians and health care users can readily understand should remain an important cornerstone for future advances in dermatological health care. Consensus on the use of 1 particular scale should be related to the scientific evidence supporting its development rather than on “expert” opinion. Clinicians, researchers, and licensing authorities should encourage the use of valid and reliable scoring systems, and, to provide standardization, should use 1 of the most extensively developed scales for comparison purposes, accepting that all of the scales discussed have potential problems associated with them. The use of new scales should be discouraged until they have been fully tested and shown to be superior in terms of validity, reliability, sensitivity, and acceptability. Future research should be directed toward identifying which symptoms and signs best measure the impact of atopic eczema on patients, and whether such measurements have global validity.

Of the severity scoring systems currently available for atopic eczema, the SCORAD index has been the most extensively tested for the quality criteria of a good scale. However, in view of the potential problems with interobserver variation, a single observer should be used whenever possible in clinical trial situations. The SASSAD has also been extensively used in clinical trials, but data on reliability testing have yet to be published. Other measures require further quality testing to confirm their usefulness.

Michael Bigby, MD
Editor