

EASI, (objective) SCORAD and POEM for atopic eczema: responsiveness and minimal clinically important difference

M. E. Schram¹, Ph. I. Spuls¹, M. M. G. Leeflang², R. Lindeboom², J. D. Bos¹ & J. Schmitt³

¹Department of Dermatology, Academic Medical Center, University of Amsterdam; ²Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands; ³Institute of Occupational and Social Medicine, Medizinische Fakultät Carl Gustav Carus, Technische Universität, Dresden, Germany

To cite this article: Schram ME, Spuls PI, Leeflang MMG, Lindeboom R, Bos JD, Schmitt J. EASI, (objective) SCORAD and POEM for atopic eczema: responsiveness and minimal clinically important difference. *Allergy* 2011; DOI: 10.1111/j.1398-9995.2011.02719.x.

Keywords

atopic eczema; Eczema Area and Severity Index; minimal clinically important difference; outcome measures; Patient-Oriented Eczema Measurement and (objective) Severity Scoring of Atopic Dermatitis Index; responsiveness; sensitivity to change.

Correspondence

M. E. Schram, MD, Department of Dermatology, A0-229, Academic Medical Center, University of Amsterdam, PO Box 22700, 1100 DD Amsterdam, the Netherlands.
Tel.: +31 20 566 2581
Fax: +31 20 566 9152
E-mail: M.E.Schram@amc.uva.nl

Accepted for publication 25 August 2011

DOI:10.1111/j.1398-9995.2011.02719.x

Edited by: Thomas Bieber

Abstract

Background: Demonstration of adequate reliability and validity is sufficient for concluding that an instrument is applicable for descriptive and predictive purposes, but before we can confidently use an outcome measure in clinical trials, the responsiveness (synonymous with sensitivity to change) and minimal clinically important difference (MCID) should be known. With this study, we aimed to assess responsiveness and MCID of four outcome measures used in atopic eczema: the Severity Scoring of Atopic Dermatitis (SCORAD), the objective SCORAD, Eczema Area and Severity Index (EASI), and the Patient-Oriented Eczema Measure (POEM).

Methods: Data of three randomized controlled trials were used. To demonstrate responsiveness, we plotted receiver operating characteristic (ROC) curves. MCID was estimated using mean change scores of patients that showed a relevant improvement. Bland and Altman methods were used to quantify the limits of agreement.

Results: Area under the ROC curve for the SCORAD was 0.70 [95% confidence interval (CI): 0.61–0.78], for the objective SCORAD, 0.73 (95% CI: 0.70–0.77), for the EASI, 0.67 (95% CI: 0.60–0.76), and for the POEM, 0.67 (95% CI: 0.59–0.75). Scores above 0.70 represent a fair responsiveness. The MCID was 8.7 points for the SCORAD, 8.2 for the objective SCORAD, 6.6 for the EASI, and 3.4 for the POEM.

Conclusion: The objective SCORAD and SCORAD showed a fair responsiveness. The MCIDs are an important prerequisite for the interpretation of published eczema trials and for the planning/sample size estimation of future trials.

Background and rationale

Atopic eczema (AE) is a highly pruritic and chronic inflammatory disease with a high prevalence (0.3–20.5% worldwide) that can lead to considerable social burden (1, 2). As reliable laboratory tests to assess and monitor disease severity of AE are not available yet, clinical outcome measures were proposed. Until now, 20 clinical outcome measures have been inaugurated to measure the severity of AE. However, a recently published systematic review (3) indicated that only three of these measures have shown adequate validity and reliability: the Eczema Area and Severity Index (EASI), the Severity Scoring of Atopic Dermatitis (SCORAD), of which there is a variant called the objective SCORAD, and the Patient-Oriented Eczema Measure (POEM) (4–6).

One significant limitation of these outcome measures, however, is the lack of evidence concerning responsiveness. Also, it is still unclear what change in scores represents a clinical relevant change. Demonstration of adequate validity and reliability is sufficient for concluding that an instrument is applicable for descriptive and predictive purposes, but before we can confidently use an outcome measure in clinical trials, the responsiveness and minimal clinically important difference (MCID) should be known (7, 8).

Responsiveness has been demonstrated in one study for the EASI and SCORAD but not for the objective SCORAD and POEM (9). The MCID is unknown for all the outcome measures. Therefore, we aimed to assess and compare the responsiveness of the (objective) SCORAD, EASI,

and POEM with data of three recently performed clinical trials and to estimate their MCID.

Patients and methods

For this study, we used data of three randomized controlled trials (RCTs) on treatments for AE (Table 1) (10–12). Responsiveness is the ability of an outcome measure to detect change over time (13). First, the mean scores of the global assessments were examined in relation to the mean scores of the outcome measure of interest per time-point in each treatment group, thereby globally illustrating responsiveness. Second, the receiver operating characteristic (ROC) curves – area under the curve (AUC) – was used (14, 15). We calculated only improvements *vs* no improvement. Patient global assessment (PGA) and investigator global assessment (IGA) were used as anchors/global ratings of change in our study. A clinically relevant change was predefined as an ‘improvement’ or ‘decline’ of ≥ 1 in PGA and IGA. If area under the ROC curve is at least 0.70, the responsiveness is considered fair, 0.8 indicates good, and 0.9 excellent responsiveness (16).

The MCID is the smallest change in an outcome measure that represents a clinically relevant outcome. We estimated MCID as absolute changes observed within individuals during treatment using the global assessment as anchors for clinical change (17). Sensitivity analysis was performed by calculating the cutoff point of the ROC curve at which correct classification (improved *vs* not improved) was optimal. Finally, the limits of agreement (LOA) were calculated to judge the absolute magnitude of the measurement error (18). When the measurement error is random, 95% of the deviation scores are expected to fall within 1.96 standard deviation (SD). In the absence of bias, the mean of deviation scores should approximate zero and have equal variance over the

total range of possible scores. A full-detailed version of the methods is available in Supporting Information.

Results

Table 1 shows the baseline characteristics of the included patients ($n = 143$). Mean age of the patients was approximately 40 years in the MAcAD trial and 30 years in the PROVE trial. In the TASCO trial, 25 children (< 18 years) and 38 adults participated, with a mean age of 9.4 and 30.9 years, respectively. Sex was generally equally distributed in all trials. In the MAcAD trial, 39 patients (93%) showed a gradual improvement over 6 months. Three patients (7%) showed no change. The PROVE trial was terminated preliminarily because of exacerbations after the termination of prednisolone (11). In the prednisolone group, most patients experienced rapid improvement within the first 3 weeks and tended to relapse thereafter. In the cyclosporine group, there was improvement until week 6. Overall, 20 patients (53%) improved, 12 patients (32%) showed no change, and six patients (16%) declined. In the TASCO trial, improvement was seen in the first 4 weeks in both groups, after which a stable period was seen. In general, 47 (75%) of the patients improved, 1 (2%) declined, and 15 (24%) showed no change. For the assessment of responsiveness and the MCID of the SCORAD, 239 observations were used, for the objective SCORAD 1067, for the EASI 239, and for the POEM 385.

Responsiveness

To illustrate responsiveness, Fig. 1 shows the mean scores of the outcome measures in relation to either IGA or PGA score during the trials. It is exemplified that all outcome measures reacted to improvements and declines in global severity measurement.

Table 1 Demographics and baseline characteristics of the included patients

	MAcAD		PROVE		TASCO	
	Methotrexate ($n = 20$)	Azathioprine ($n = 22$)	Prednisolone ($n = 21$)	Cyclosporine ($n = 17$)	Algorithm ($n = 32$)	Individual ($n = 31$)
Age in years	43.0 (14.7)	37.0 (14.1)	28.8 (9.9)	30.1 (8.9)	24 (17)	21 (13)
Male sex	10 (50%)	12 (54.5%)	9 (52.4%)	10 (58.8%)	14 (44%)	16 (45%)
Presence of asthma or allergic rhinitis	19 (95%)	18 (82%)	16 (76.2%)	12 (70.6%)	10 (31%)A 19 (59%)R	9 (29%)A 21 (68%)R
Presence of allergen-specific IgE	20 (100%)	21 (95%)	–	–	28 (88%)	28 (90%)
Duration of eczema in years	39.8 (16.2)	33.1 (6.8)	–	–	–	–
Disease activity						
SCORAD	57.2 (11.8)	58.4 (10.4)	–	–	–	–
Objective SCORAD	48.9 (13.0)	47.4 (9.2)	57.6 (10.2)	54.7 (10.9)	46.1 (12.1)	47.1 (19.4)
EASI	27.9 (12.3)	30.4 (14.2)	–	–	–	–
POEM	19.8 (5.4)	19.6 (4.0)	22.1 (7.1)	21.4 (5.4)	–	–

Data are mean (SD) or number (%).

AE, Atopic eczema; A, asthma; MAcAD, trial comparing methotrexate with azathioprine in adult patients with severe AE; PROVE, trial comparing prednisolone with cyclosporine in adult patients with severe AE; TASCO, trial comparing an evidence-based treatment algorithm with individualized symptom-oriented treatment in children and adults with moderate to severe AE; n , number; R, allergic rhinitis; EASI, Eczema Area and Severity Index; POEM, Patient-Oriented Eczema Measure; SCORAD, Severity Scoring of Atopic Dermatitis; –, not provided.

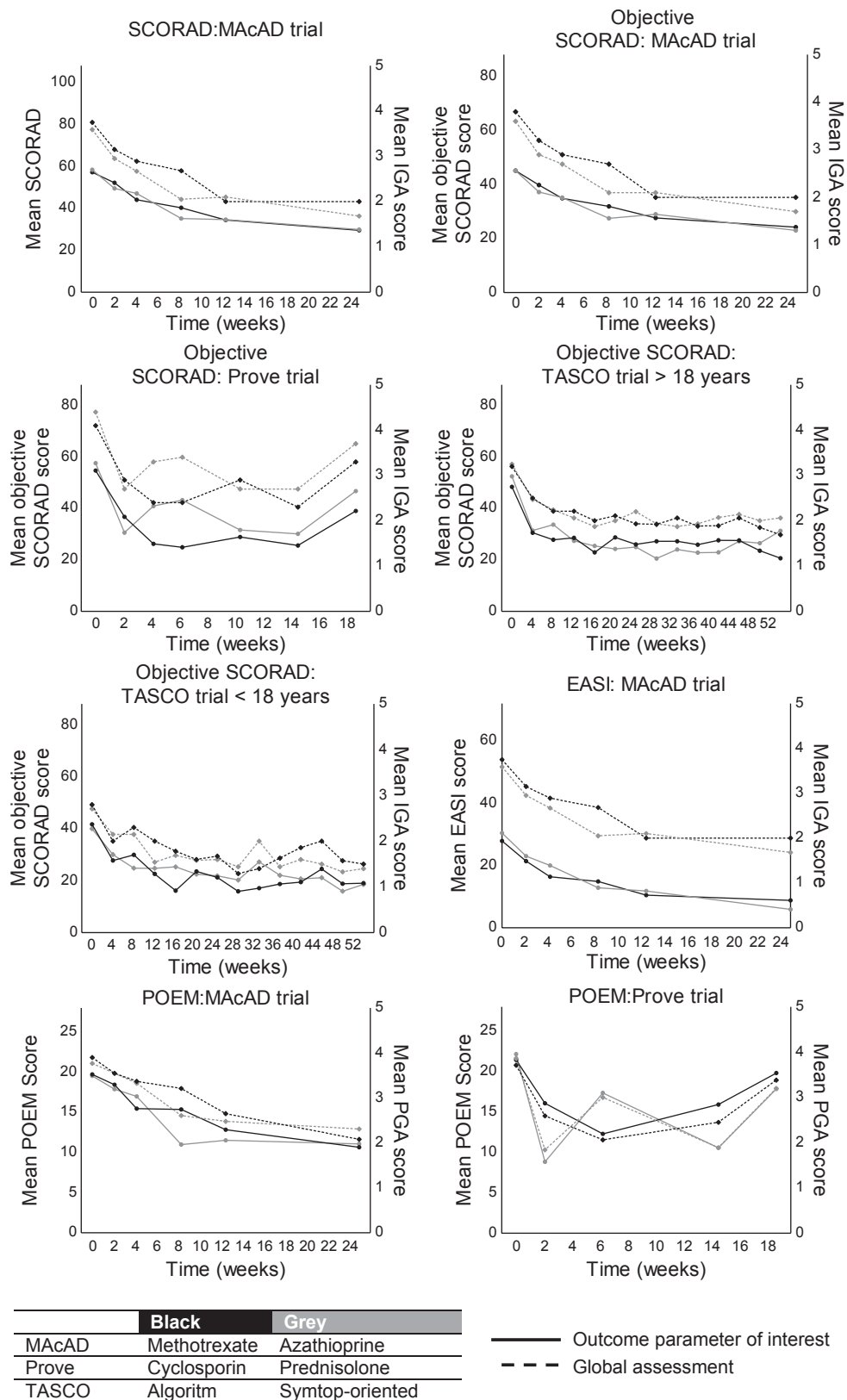


Figure 1 Responsiveness: (objective) Severity Scoring of Atopic Dermatitis, Eczema Area and Severity Index, and Patient-Oriented Eczema Measure against global assessment.

The SCORAD showed an AUC of 0.70 [95% confidence interval (CI): 0.61–0.78] (Table 2). The AUC of objective SCORAD was 0.73 (95% CI: 0.70–0.77). The AUC of the EASI is 0.67 (95% CI: 0.60–0.76) and of the POEM is 0.67 (95% CI: 0.59–0.75). We performed separate subgroup analysis for the objective SCORAD, in children and adults. The AUC of the objective SCORAD for adults was 0.73 (95% CI: 0.68–0.77) and for children 0.76 (95% CI: 0.68–0.84).

MCID

The anchor-based longitudinal mean difference in change scores for patients with a one-point improvement in global severity score is shown in Table 3. For the SCORAD, the overall mean difference was 8.7 points (SD 7.8) for an improvement. In other words, a mean improvement of 8.7 points on the SCORAD corresponds with a clinically relevant change. The mean difference varied from 6.8 to 12.6 points, when each individual step of improvement on IGA was specified (e.g., from 5 to 4 points and from 4 to 3 points). Analysis of the ROC cutoff points for greatest correct classification showed that 4.1 points on the SCORAD was best for identifying a clinically significant improvement with a sensitivity of 64.7% and a specificity of 64.2% (Table 2). The difference in mean score for stable patients on the SCORAD (measurement error) was 2.2 points (SD 7.8). Limits of agreement ranged from –13.1 to 17.5 points (Fig. 2). Ninety four percent of our observations were within this range.

When pooling all the available data from the different trials, we found an overall mean difference of the objective SCORAD of 8.2 points (SD 8.7) when an improvement in global severity score (IGA) was observed (Table 3). Taking into account all the observations in adult patients, the mean difference was 8.0 points (SD 8.6). For observations from children, the mean difference was 9.0 (SD 9.2). This was only based on IGA anchor scores ranging from 3 to 0, owing to the milder disease of the included children. Overall, analysis of the ROC cutoff points for greatest correct classification

showed that 6.5 points on the objective SCORAD was best for identifying a clinically significant improvement with a sensitivity of 54.8% and a specificity of 78.8%. Cutoff points for the different adult populations were similar. The cutoff point for the observations of children was 3.45 points. For all patients, the measurement error was 1.2 (SD 8.0) points. The LOA ranged from –14.5 to 16.8 (93.6% of the observations within the range). For all adult patients, the same measurement error was found. The LOA ranged from –14.4 to 16.9 (93.2% of the observations within the range). For observations in children, the measurement error was 0.7 points (SD 7.1), with a limit of agreement ranging from –13.1 to 14.5 subsequently (92.9% of the observations within the range).

For the EASI, the overall mean difference was 6.6 points (SD 5.9) when IGA improved with one point. The MCID varied from 1.0 (anchor from 1 to 0) to 8.6 (anchor from 5 to 4). This range might be due to the small number of observations in both categories. The ROC cutoff point was 2.75 points with 73.8% sensitivity and 57.4% specificity. Measurement error subtracted from that of the Bland and Altman analysis was 3.03 points (SD 5.8). Limits of agreement ranged from –8.7 to 14.4 points (97.3% of the observations within the range).

The overall mean difference of the POEM was 3.4 points (SD 4.8). The mean difference ranged from 0.1 (anchor from 5 to 4) to 4.5 (anchor from 4 to 3). An improvement of 1.5 points on the POEM was the ROC cutoff point (sensitivity 62.1%, specificity 66.1%). The mean measurement error was 0.8 points (SD 5.3). Therefore, LOA ranged from –9.6 to 11.2 points (91.8% of the observations within the range).

Discussion

Main findings

In this study, we assessed the responsiveness and the MCID of four outcome measures for AE. The results indicate that these measures are all fairly responsive but that their ability

Table 2 Responsiveness: area under the ROC curve

	AUC	95% CI	Cutoff	Sensitivity (%)*	Specificity (%)*
SCORAD	0.70	0.61–0.78	4.05	64.7 (51.0–76.4)	64.2 (55.8–71.8)
Objective SCORAD					
MAcAD & PROVE	0.73	0.67–0.79	7.00	55.5 (46.5–64.1)	80.4 (73.5–85.8)
TASCO <18 years	0.76	0.68–0.84	3.45	51.3 (37.7–64.1)	82.1 (73.7–88.8)
TASCO >18 years	0.72	0.65–0.79	6.55	51.3 (40.4–62.1)	82.1 (75.9–86.9)
All adult patients	0.73	0.68–0.77	6.85	54.8 (47.8–61.6)	79.5 (74.6–83.2)
All patients	0.73	0.70–0.77	6.45	54.8 (48.6–60.9)	78.8 (74.8–82.4)
EASI	0.67	0.60–0.76	2.75	73.8 (61.2–83.6)	57.4 (46.8–67.5)
POEM	0.67	0.59–0.75	1.50	62.1 (50.1–72.9)	66.1 (57.4–73.9)

AUC, area under the curve; CI, confidence interval; EASI, Eczema Area and Severity Index; MAcAD, trial comparing methotrexate with azathioprine in adult patients with severe AE; POEM, Patient-Oriented Eczema Measure; PROVE, trial comparing prednisolone with cyclosporine in adult patients with severe AE; ROC, receiver operating characteristic; SCORAD, Severity Scoring of Atopic Dermatitis; TASCO, trial comparing an evidence-based treatment algorithm with individualized symptom-oriented treatment in children and adults with moderate to severe AE.

*Sensitivity and specificity reflect the highest correct classification for the cutoff value.

Table 3 Longitudinal mean difference of changes scores

Outcome measurement	RCT	Anchor (IGA/PGA)	N of obs.	Mean difference	Min	Max	SD
SCORAD	MAcAD	5 → 4	4	10.8	-1.1	20.9	9.6
		4 → 3	29	6.8	-9.0	28.3	8.9
		3 → 2	30	9.8	-3.0	25.6	7.1
		2 → 1	13	9.5	-0.8	16.3	6.5
		1 → 0	1	12.6	-12.6	12.6	-
		Total	77	8.7	-9.0	28.3	7.8
Objective SCORAD	MAcAD and PROVE	5 → 4	8	11.9	2.5	31.0	9.8
		4 → 3	50	7.7	-11.4	29.0	9.1
		3 → 2	43	7.9	-7.0	19.5	6.0
		2 → 1	17	7.5	-3.9	19.5	6.2
		1 → 0	1	10.6	10.6	10.6	-
		Total	119	8.0	-11.4	31.0	7.7
	TASCO <18 years	5 → 4	0	-	-	-	-
		4 → 3	0	-	-	-	-
		3 → 2	16	12.4	-5.0	37.9	11.1
		2 → 1	28	7.4	-7.6	25.5	8.5
		1 → 0	7	7.5	1.3	12.2	4.5
		Total	51	9.0	-7.6	37.9	9.2
	TASCO >18 years	5 → 4	0	-	-	-	-
		4 → 3	11	9.9	-6.0	31.6	12.2
		3 → 2	27	10.4	-22.6	33.0	11.3
		2 → 1	35	5.8	-9.0	28.8	7.6
		1 → 0	5	6.0	-4.4	17.9	9.2
		Total	78	8.0	-22.6	33.0	9.9
	All adult patients	5 → 4	8	11.9	2.5	31.0	9.8
		4 → 3	61	8.1	-11.4	31.6	9.7
		3 → 2	70	8.8	-22.6	33.0	8.4
		2 → 1	52	6.4	-9.0	28.8	7.2
		1 → 0	6	6.8	-4.4	17.9	8.4
		Total	197	8.0	-22.6	33.0	8.6
	All patients	5 → 4	8	11.9	2.5	31.0	9.8
		4 → 3	61	8.1	-11.4	31.6	9.7
		3 → 2	86	9.5	-22.6	37.9	9.0
		2 → 1	80	6.7	-9.0	28.8	7.7
		1 → 0	13	7.2	-4.4	17.9	6.3
		Total	248	8.2	-22.6	37.9	8.7
EASI	MAcAD	5 → 4	3	8.6	2.9	18.7	8.8
		4 → 3	21	7.8	-3.7	24.3	6.9
		3 → 2	27	7.3	-1.7	21.7	5.7
		2 → 1	13	3.3	-0.6	6.6	2.4
		1 → 0	1	1.0	1	1	-
		Total	65	6.6	-3.7	24.3	5.9
POEM	MAcAD and PROVE*	5 → 4	7	0.1	-5	3	3.1
		4 → 3	29	4.5	-1	21	5.1
		3 → 2	25	3.4	-8	18	5.0
		2 → 1	5	2.0	-2	7	3.2
		1 → 0	2	4.0	3	5	1.4
		Total	68	3.4	-8	21	4.8

EASI, Eczema Area and Severity Index; IGA, investigator global assessment; max, maximal change score; min, minimal change score; obs., observations; PGA, patient global assessment; SD, standard deviation; MAcAD, trial comparing methotrexate with azathioprine in adult patients with severe AE; POEM, Patient-Oriented Eczema Measure; PROVE, trial comparing prednisolone with cyclosporine in adult patients with severe AE; SCORAD, Severity Scoring of Atopic Dermatitis; TASCO, trial comparing an evidence-based treatment algorithm with individualized symptom-oriented treatment in children and adults with moderate to severe AE.

*Anchor by PGA.

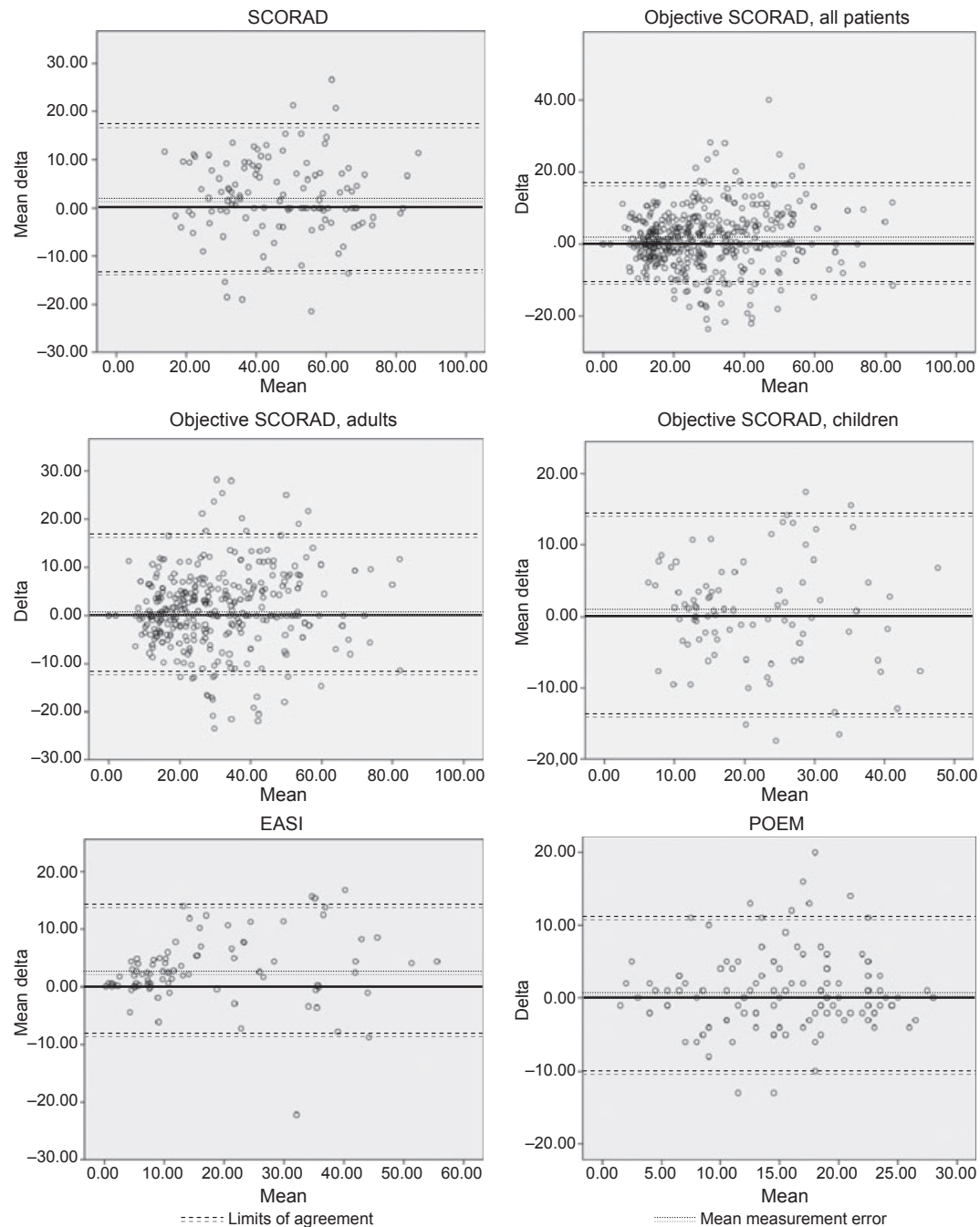


Figure 2 Measurement error by Bland and Altman.

to discriminate between patients who improved and who did not varies. The responsiveness, or sensitivity to change, of the objective SCORAD and SCORAD was fair, meaning that the area under their ROC curve was larger than 0.70. Although the EASI and the POEM had an AUC below 0.70, the differences in responsiveness were not significant as the CIs overlapped almost entirely.

To calculate the MCID, we used two different analytic methods. For the SCORAD, we found a mean difference in

SCORAD scores between two subsequent categories of IGA (i.e., one-point improvement on the IGA) of 8.7 points. As the SCORAD ranges from 0 to 103, and the IGA ranges from 0 to 5, this means a change of about 20% on the IGA only translates to a change of <10% on the SCORAD. The optimum cutoff on the ROC curve of the SCORAD was 4.1 points, meaning that when considering people with a change of 4.1 SCORAD points or more as being changed, this will result in the highest combination of sensitivity and specificity

for detecting change in IGA. For the SCORAD, the sensitivity at this optimal cutoff point was 65% (95% CI: 51–76%) and the specificity was 64% (95% CI: 56–72%).

The POEM showed similar sensitivity and specificity at its optimal cutoff point (1.5), and its mean difference was 3.4. Although the objective SCORAD showed higher specificity (79%) at its optimal cutoff (6.5), its sensitivity was lower than that of the other measures (55%). The EASI, on the other hand, showed a higher sensitivity (74%) but a lower specificity (57%) at an optimal cutoff of 2.8. The mean difference for the objective SCORAD was 8.2 points and for the EASI 6.6 points.

Differences in outcomes between the primary and sensitivity analysis could be due to the fact that we measured two slightly different MCIDs. It can be argued that the longitudinal anchor-based mean difference is technically not the most *minimal* difference that is clinically important but is more a clinically important difference. This explains the trend that the longitudinal mean difference is higher than the optimal cutoff point for best discrimination. We would advise to use the highest approximation of the MCID, because this would be less likely to overestimate clinical meaningfulness of a trial.

When performing power calculations for a parallel-group RCT based on these MCIDs, a group size of 14, 19, 14 or 34 is needed for rejecting the null hypothesis based on the SCORAD, objective SCORAD, EASI, or POEM, respectively (80% power and 5% significance).

All estimated clinically important differences and optimal cutoffs were within the range of the LOA, meaning that individual change scores that are similar to the MCID cannot solely be attributed to a real clinical meaningful change but could also be attributed to measurement error. This is another argument to use the highest approximation of the MCID.

Strengths and weaknesses

This study was conducted following the guidelines for performing clinimetric studies by Consensus-based Standards for the selection of health Measurement INstruments (COSMIN) (19, 20). As the COSMIN panel reached consensus that effect sizes and related measures are inappropriate measures for responsiveness, we have chosen to use the ROC-AUC method. The COSMIN does not provide a consensus on how to calculate the MCID. Therefore, we have estimated the MCID using different analytical methods.

We have only performed ROCs on improvement. Improvement was defined as improving one point on global assessment; thus, improvements of >1 point on global assessment were excluded. This will have resulted in slightly lower outcome on AUC. As most of the patients improved during the trials, there were too little observations on worsening.

We have chosen to determine the MCID within patients using an anchor-based longitudinal approach. The advantage is that change in an outcome measure is linked to a meaningful external anchor. Disadvantages are that a longitudinal approach can be biased by response shift and that the anchor relies on global ratings (21). Global ratings are broad or not precise and might have a nonlinear distribution. For instance,

the difference between ‘severe disease’ and ‘very severe disease’ is smaller than the difference between ‘mild’ and ‘moderate’. Therefore, some outcomes can be attributed to characteristics of the global assessment rather than to the outcome measure. Also, we used the improvement on the global assessment as anchor for change instead of a transition scale, which is more generally accepted.

Absolute changes to calculate the MCID were used. These changes have the advantage that they are more easily applicable in studies and in practice. Relative changes might be more methodologically sound, but they are not easy to handle on a routine basis.

The LOA showed a large variation. Although we assumed that the patients had stable disease when global assessment remained stable at consecutive visits, chance variation can never be excluded. Besides, some degree of variation in disease severity could not be excluded owing to the time between visits (4 weeks generally) and because the global assessment allows some variation in disease activity within a point on the scale. Furthermore, inter- and intrarater variability will have played a role. Therefore, we believe that the range of the LOA is an overestimation to some degree.

Generalizability

We think that the data from the MACAD and the PROVE trial are representative for all adult patients with chronic severe AE. Data from the TASCO trial also included children. Only the objective SCORAD was calculated with data from children. Those children who had generally milder disease than the adult patients were included. Responsiveness and MCID did not substantially differ between the children and adults on the objective SCORAD. This might be a positive argument for generalizing the outcomes of other outcome measures to children.

It is known that the MCID may vary by population and context, and this is the first approximation of the MCID. Confidence in a specific MCID evolves over time and is confirmed by additional evidence.

Clinical implications

Currently, the harmonizing outcome measures in eczema (HOME) initiative is undertaken by international specialists on eczema. Its main goal is to define core outcome domains that should be used in every clinical trial concerning AE to enable meta-analysis and thereby improving the strength of evidence. In this perspective, it was of great significance to establish the responsiveness of the outcome measures. The quantification of the MCID offers advantages to analyze and design clinical trials. Based on the MCID, accurate power calculations can be made and it can be decided whether statistically significant differences reflect a real change in disease severity.

Author contributions

ME Schram designed the study; collected, analyzed, and interpreted the data, drafted the article; and contributed to

the final approval of the version. PhI Spuls and JD Bos conceived the study question and contributed to editing, revision of the manuscript, response to reviewers, and the final approval of the version. MMG Leeftang analyzed and interpreted the data, and contributed to editing, revision of the manuscript, response to reviewers, and the final approval of the version. R Lindeboom designed the study; analyzed and interpreted the data; and contributed to editing, revision of the manuscript, response to reviewers, and the final approval of the version. J Schmitt conceived the study question and contributed to designing the study, acquisition of data, editing, revision of the manuscript, response to reviewers, and the final approval of the version.

Funding

None.

References

- Worldwide variation in prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and atopic eczema: ISAAC. The International Study of Asthma and Allergies in Childhood (ISAAC) Steering Committee. *Lancet* 1998;**351**:1225–1232.
- Carroll CL, Balkrishnan R, Feldman SR, Fleischer AB Jr, Manuel JC. The burden of atopic dermatitis: impact on the patient, family, and society. *Pediatr Dermatol* 2005;**22**:192–199.
- Schmitt J, Langan S, Williams HC. What are the best outcome measurements for atopic eczema? A systematic review. *J Allergy Clin Immunol* 2007;**120**:1389–1398.
- Severity scoring of atopic dermatitis: the SCORAD index. Consensus Report of the European Task Force on Atopic Dermatitis. *Dermatology* 1993;**186**:23–31.
- Charman CR, Venn AJ, Williams HC. The patient-oriented eczema measure: development and initial validation of a new tool for measuring atopic eczema severity from the patients' perspective. *Arch Dermatol* 2004;**140**:1513–1519.
- Tofte SJ, Graeber M, Cherill R, Omoto M, Thurston M, Hanifin JM. Eczema area and severity index (EASI): a new tool to evaluate atopic dermatitis. *J Eur Acad Dermatol Venereol* 1998;**11**:S197.
- Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. *J Clin Epidemiol* 1989;**42**:403–408.
- Fletcher A. Quality-of-life measurements in the evaluation of treatment: proposed guidelines. *Br J Clin Pharmacol* 1995;**39**:217–222.
- Breuer K, Braeutigam M, Kapp A, Werfel T. Influence of pimecrolimus cream 1% on different morphological signs of eczema in infants with atopic dermatitis. *Dermatology* 2004;**209**:314–320.
- Schmitt J, Meurer M, Schwanebeck U, Grahler X, Schakel K. Treatment following an evidence-based algorithm versus individualised symptom-oriented treatment for atopic eczema. A randomised controlled trial. *Dermatology* 2008;**217**:299–308.
- Schmitt J, Schakel K, Folster-Holst R, Bauer A, Oertel R, Augustin M et al. Prednisolone vs. ciclosporin for severe adult eczema. An investigator-initiated double-blind placebo-controlled multicentre trial. *Br J Dermatol* 2010;**162**:661–668.
- Schram ME, Roekevisch E, Leeftang MM, Bos JD, Schmitt J, Spuls PI. A randomized trial of methotrexate versus azathioprine for severe atopic eczema. *J Allergy Clin Immunol* 2011;**128**:353–359.
- Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008;**61**:102–109.
- Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. *J Clin Epidemiol* 2001;**54**:1204–1217.
- Beaton DE, Katz JN, Fossel AH, Wright JG, Tarasuk V, Bombardier C. Measuring the whole or the parts? Validity, reliability, and responsiveness of the disabilities of the arm, shoulder and hand outcome measure in different regions of the upper extremity. *J Hand Ther* 2001;**14**:128–146.
- Hosmer DW. Applied Logistic Regression. New York: John Wiley and Sons, 2000.
- Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;**10**:407–415.
- de Winter AF, Heemskerk MA, Terwee CB, Jans MP, Deville W, van Schaardenburg DJ et al. Inter-observer reproducibility of measurements of range of motion in patients with shoulder pain using a digital inclinometer. *BMC Musculoskelet Disord* 2004;**5**:18.
- Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol* 2010;**10**:22.
- Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;**63**:737–745.
- Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol* 2003;**56**:395–407.

Conflict of interest

The authors declare that there are no conflicts of interest.

Supporting information

Additional Supporting Information may be found in the online version of this article found at: <http://www.wileyonlinelibrary.com>

Data S1. Detailed version of the methods section.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Data

For this study we used the clinical outcome data of three investigator-initiated randomized controlled trials (RCT's) on the efficacy of treatments for AE.

The first RCT (further specified as MAcAD) compared methotrexate versus azathioprine in adult patients with severe AE and was performed in the Academic Medical Center, University of Amsterdam, the Netherlands.¹ It concerned a single blind, parallel-group trial in which 42 subjects were recruited and randomized in a 1:1 ratio. The total study duration was 24 weeks: 12 weeks active treatment and 12 weeks follow up in which responders could continue or stop their study medication. At week 0, 2, 4, 8, 12 and 24 of the trial, (objective) SCORAD, EASI, POEM, investigator global assessment (IGA) and patients global assessment (PGA) were assessed.

The second group of subjects (n=38) was recruited for a parallel-group double blinded RCT (PROVE) comparing prednisolone with cyclosporine for severe AE conducted at the Carl Gustav Carus Medical Faculty, Technical University of Dresden, Germany and three other centres in Germany (Kiel, Münster, Hamburg).² In the prednisolone group patients were treated for three weeks with prednisolone and then four weeks with placebo. In the cyclosporine group patients were treated for six weeks. Duration of follow-up was 12 weeks in which no medication was given. At week 0, 2, 4, 6, 10, 14 and 18 the objective SCORAD, POEM, IGA and PGA were assessed.

Main inclusion criteria for both trials were age ≥ 18 years, state of general good health, severe eczema (objective SCORAD ≥ 40 or severe according to Rajka and Langeland criteria).³

In the third RCT (TASCO), which was performed at the Carl Gustav Carus Medical Faculty, Technical University of Dresden, Germany, an evidence-based treatment algorithm was compared with individualized symptom-oriented treatment in 63 subjects with AE aged two years or older, who had at least moderate AE (SCORAD > 20).⁴ Duration of the study was one year in which the objective SCORAD and IGA were assessed every four weeks. Treatment regimens included topical therapy (corticosteroids and calcineurine inhibitors), phototherapy and systemic therapy (corticosteroids and cyclosporine).

All patients were diagnosed with AE by an experienced dermatologist and by the use of the U.K. Working Party Diagnostic criteria.⁵ Approval by the ethics committees was obtained for all studies and all patients signed a written informed consent prior to study related procedures.

Outcome measures

The SCORAD combines both 'objective' items as affected area and intensity of the lesions, and 'subjective' items as extent of pruritus and sleep loss. The 'objective' part of the SCORAD is referred to as 'objective SCORAD' and is often used alone. To calculate the SCORAD, affected area is scored by applying *the rule of nine* after drawing the lesions on an evaluation form. The intensity is determined by grading each intensity item (erythema, oedema/papulation, oozing/crusts, excoriation, lichenification and dryness) on a scale of 0 to 3. Pruritus and sleep loss were rated by the patient on a visual analogue scale. The total score is the sum of extent/5 + 7 x intensity/2 + sum of symptoms. Range of the scores is from 0 to 103 for the SCORAD and from 0 to 83 for the objective SCORAD. Over the years, SCORAD has been validated in several studies.⁶

The EASI combines affected area and intensity of the lesions.^{7,8} For affected area, the EASI assigns proportionate body surface areas to the head (10%), trunk (30%), upper extremities (20%), and lower extremities (40%). Intensity of the lesions is determined by 4 items (erythema, infiltration/papulation, excoriation and lichenification) in each particular region (head, trunk, upper and lower extremities). Each item is assessed on a scale from 0 to 3 for each region. The region score is a product of the total intensity score for that region and the proportional area score. The total score is a sum of each region score and can range from 0 to 72. To date, the EASI has been validated in several studies, but adequate evidence concerning sensitivity to change is missing.⁶

Both SCORAD and EASI showed adequate convergent and divergent construct validity according to those validation studies. Thereby, it should be noted that the (objective) SCORAD and EASI measure extent in different ways.⁹

The POEM was inaugurated in 2004 by Charman *et al.* to enable patient-oriented assessment of disease severity.¹⁰ The POEM questionnaire consists of seven questions regarding the amount of days during the last week that patient experienced itch, sleep loss, bleeding, cracking weeping, flaking of the skin and dry skin. The number of days per item can represent a score from 0 to 4. The total score is a sum of all individual scores. Maximum score is 28 points. From the measures investigated

here, the POEM is the only outcome measurement that has been documented to show adequate internal consistency.¹⁰

IGA and PGA were scored using a six-point Likert scale; 0= clear, 1= almost clear, 2= mild disease, 3= moderate disease, 4= severe disease and 5= very severe disease. The IGA has a close correlation with both the EASI and the SCORAD.^{11;12} The POEM correlated well with the PGA.¹⁰ In our study the correlation between IGA and EASI was 0.83, between IGA and (objective) SCORAD 0.85 and between PGA and POEM 0.65.

The POEM questionnaire was translated into Dutch and German. The process involved forward and back-translation per standard protocol. Since all investigators had a thorough command of English language, the other outcome measures were not translated.

Assessments

Outcome assessors of the included RCTs were instructed to use the outcome measures prior to the studies by written details and/or oral presentation. In the MAcAD and PROVE trial, outcome assessors were nurses and trial doctors who were blinded for treatment allocation. Evaluation of patients was conducted under standardized conditions in each study. All patients in the different groups were instructed not to apply any creams or ointments to their eczema directly before assessment. The sum scores were calculated in SPSS 18.0.

Statistical analysis

For measuring the MCID and responsiveness of the EASI and SCORAD, data from the MAcAD trial were used. For the POEM, data from the PROVE and MAcAD trial were used. The MCID and responsiveness of the objective SCORAD was measured by pooled data from the MAcAD and PROVE trial and by data from the TASCOT trial concerning adults and children (<18 years).

PGA and IGA were used as anchors/ global ratings of change in our study. A clinically relevant change was predefined as an 'improvement' or 'decline' of ≥ 1 in PGA and IGA scores since it would be highly likely that a clinical relevant difference occurs in health status if the patients changes one point on the global assessment.

Responsiveness (synonym: sensitivity to change) is the ability of an outcome measure to detect change over time in the construct to be measured and can be seen as a form of longitudinal

validity.¹³ Firstly, the mean scores of the global assessments (reference test) were examined in relation to the mean scores of the outcome measure of interest (index test) per time-point in each treatment group, thereby globally illustrating responsiveness. Secondly, the receiver operating characteristic (ROC) curves - area under the curve (AUC) was used to estimate responsiveness.^{14;15} An ROC curve is a graph that plots sensitivity against 1- specificity. The AUC illustrates how well score changes in the outcome measures can discriminate between patients that did or did not change, defined as a 1-point change in global assessment score between time-points. We calculated only improvements versus no improvement, since data on decline were too scarce. If area under the ROC curve is at least 0.70, the responsiveness is considered fair, 0.8 indicates good and 0.9 excellent responsiveness.¹⁶

The MCID is the smallest change in an outcome measure that represents a clinically relevant outcome. We estimated MCID using three complementary approaches. Primarily, we analyzed the absolute changes observed within individuals during treatment using the global assessment as anchors for clinical change.¹⁷ Anchors for relevant change were PGA for the POEM and IGA for the (objective) SCORAD and EASI. We calculated the mean change score on outcome measures when an improvement of one point on global assessment was seen. Sensitivity analysis was done by calculating the cut-off point of the ROC-curve at which correct classification (improved versus not improved on global assessment) was optimal i.e. the cut-off with the highest sum of true positive and true negative classifications.

Finally, to assess the expected amount of variability that can be expected in stable patients, we compared the scores within measurement occasions of patients that did not change on the global assessment using the method of Bland & Altman.¹⁸ The limits of agreement were calculated to judge the absolute magnitude of the measurement error i.e. any score deviation larger than the expected zero. Limits of agreement were calculated as $\pm 1.96 \times$ standard deviation of the measurement error. When the measurement error is random, 95% of the deviation scores are expected to fall within 1.96 SD. In the absence of bias, the measurement error should approximate zero and have equal variance over the total range of possible scores.

Missing data were minimal (< 2%). Missing observations were excluded from analysis. Data was analysed by the first author using SPSS 18.0.

Reference List

1. Schram ME, Roekevisch E, Leeftang MM *et al.* A randomized trial of methotrexate versus azathioprine for severe atopic eczema. *J Allergy Clin Immunol* 2011.
2. Schmitt J, Schakel K, Folster-Holst R *et al.* Prednisolone vs. ciclosporin for severe adult eczema. An investigator-initiated double-blind placebo-controlled multicentre trial. *Br J Dermatol* 2010; **162**: 661-8.
3. Rajka G, Langeland T. Grading of the severity of atopic dermatitis. *Acta Derm Venereol Suppl (Stockh)* 1989; **144**: 13-4.
4. Schmitt J, Meurer M, Schwanebeck U *et al.* Treatment following an evidence-based algorithm versus individualised symptom-oriented treatment for atopic eczema. A randomised controlled trial. *Dermatology* 2008; **217**: 299-308.
5. Williams HC, Burney PG, Hay RJ *et al.* The U.K. Working Party's Diagnostic Criteria for Atopic Dermatitis. I. Derivation of a minimum set of discriminators for atopic dermatitis. *Br J Dermatol* 1994; **131**: 383-96.
6. Schmitt J, Langan S, Williams HC. What are the best outcome measurements for atopic eczema? A systematic review. *J Allergy Clin Immunol* 2007; **120**: 1389-98.
7. Hanifin JM, Thurston M, Omoto M *et al.* The eczema area and severity index (EASI): assessment of reliability in atopic dermatitis. EASI Evaluator Group. *Exp Dermatol* 2001; **10**: 11-8.
8. Fredriksson T, Pettersson U. Severe psoriasis--oral therapy with a new retinoid. *Dermatologica* 1978; **157**: 238-44.
9. Charman CR, Venn AJ, Williams HC. Measurement of body surface area involvement in atopic eczema: an impossible task? *Br J Dermatol* 1999; **140**: 109-11.
10. Charman CR, Venn AJ, Williams HC. The patient-oriented eczema measure: development and initial validation of a new tool for measuring atopic eczema severity from the patients' perspective. *Arch Dermatol* 2004; **140**: 1513-9.
11. Breuer K, Braeutigam M, Kapp A *et al.* Influence of pimecrolimus cream 1% on different morphological signs of eczema in infants with atopic dermatitis. *Dermatology* 2004; **209**: 314-20.
12. Barbier N, Paul C, Luger T *et al.* Validation of the Eczema Area and Severity Index for atopic dermatitis in a cohort of 1550 patients from the pimecrolimus cream 1% randomized controlled clinical trials programme. *Br J Dermatol* 2004; **150**: 96-102.
13. Revicki D, Hays RD, Cella D *et al.* Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008; **61**: 102-9.

14. Beaton DE, Bombardier C, Katz JN *et al.* A taxonomy for responsiveness. *J Clin Epidemiol* 2001; **54**: 1204-17.
15. Beaton DE, Katz JN, Fossel AH *et al.* Measuring the whole or the parts? Validity, reliability, and responsiveness of the Disabilities of the Arm, Shoulder and Hand outcome measure in different regions of the upper extremity. *J Hand Ther* 2001; **14**: 128-46.
16. Hosmer DW. *Applied logistic regression*. New York: John Wiley and sons, 2000.
17. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989; **10**: 407-15.
18. de Winter AF, Heemskerk MA, Terwee CB *et al.* Inter-observer reproducibility of measurements of range of motion in patients with shoulder pain using a digital inclinometer. *BMC Musculoskelet Disord* 2004; **5**: 18.