Arthritis & Rheumatism (Arthritis Care & Research)
Vol. 55, No. 3, June 15, 2006, pp 355–363
DOI 10.1002/art.22002

SPECIAL ARTICLE

The Pediatric Rheumatology International Trials Organization/American College of Rheumatology Provisional Criteria for the Evaluation of Response to Therapy in Juvenile Systemic Lupus Erythematosus: Prospective Validation of the Definition of Improvement

NICOLINO RUPERTO,[1] ANGELO RAVELLI,[1] SHEILA OLIVEIRA,[2] MARIA ALESSIO,[3] DIMITRINA MIHAYLOVA,[4] SRDJAN PASIC,[5] ELISABETTA CORTIS,[6] MARIA APAZ,[7] RUBEN BURGOS-VARGAS,[8] FLORENCE KANAKOUDI-TSAKALIDOU,[9] XIMENA NORAMBUENA,[10] FABRIZIA CORONA,[11] VALERIA GERLONI,[12] STEFAN HAGELBERG,[13] AMITA AGGARWAL,[14] PAVLA DOLEZALOVA,[15] CLAUDIA MAGALHAES SAAD,[16] SANG-CHEOL BAE,[17] RICHARD VESELY,[18] TADEJ AVCIN,[19] HELEN FOSTER,[20] CAROLINA DUARTE,[21] TROELS HERLIN,[22] GERD HORNEFF,[23] LOREDANA LEPORE,[24] MARION VAN ROSSUM,[25] LUCIA TRAIL,[1] ANGELA PISTORIO,[26] BOEL ANDERSSON-GÄRE,[27] EDWARD H. GIANNINI,[28] AND ALBERTO MARTINI,[29] FOR THE PEDIATRIC RHEUMATOLOGY INTERNATIONAL TRIALS ORGANIZATION (PRINTO) AND THE PEDIATRIC RHEUMATOLOGY COLLABORATIVE STUDY GROUP (PRCSG)

This criteria set has been approved by the American College of Rheumatology (ACR) Board of Directors as Provisional. This signifies that the criteria set has been quantitatively validated using patient data, but it has not undergone validation based on an external data set. All ACR-approved criteria sets are expected to undergo intermittent updates.

Objective. **To use the Pediatric Rheumatology International Trials Organization (PRINTO) core set of outcome measures to develop a validated definition of improvement for the evaluation of response to therapy in juvenile systemic lupus erythematosus (SLE).**
Methods. **Thirty-seven experienced pediatric rheumatologists from 27 countries, each of whom had specific experience in the assessment of juvenile SLE patients, achieved consensus on 128 patient profiles as being clinically improved or not improved. Using the physicians' consensus ratings as the gold standard measure, the chi-square, sensitivity, specificity, false-positive and false-negative rates, area under the receiver operating characteristic curve, and kappa level of agreement for 597 candidate definitions of improvement were calculated. Only definitions with a kappa value greater than 0.7 were retained. The top definitions were selected based on the product of the content validity score multiplied by its kappa statistic.**
Results. **The definition of improvement with the highest final score was at least 50% improvement from baseline in any 2 of the 5 core set measures, with no more than 1 of the remaining worsening by more than 30%.**
Conclusion. **PRINTO proposes a valid and reproducible definition of improvement that reflects well the consensus rating of experienced clinicians and that incorporates clinically meaningful change in core set measures in a composite end point for the evaluation of global response to therapy in patients with juvenile SLE. The definition is now proposed for use in juvenile SLE clinical trials and may help physicians to decide whether a child with SLE responded adequately to therapy.**

KEY WORDS. Juvenile systemic lupus erythematosus; Core set; Response to therapy; Disease activity; Consensus.

INTRODUCTION

Although the treatment of systemic lupus erythematosus (SLE) has improved markedly, its morbidity and treatment-related damage are still considerable (1–3). In contrast, several promising therapeutic modalities have become available, and many others are likely to appear in the future (4). To ensure maximum efficiency in the evaluation of new treatments, powerful and uniform criteria for the assessment of response in clinical trials are needed (5–7). Recently, the American College of Rheumatology proposed response criteria based on measures of overall dis-

Supported by a grant from the European Union (contract no. QLG1-CT-2000-00514), the IRCCS G. Gaslini, Genoa, Italy, and the NIH (grant RO3-AI-44046).
[1]Nicolino Ruperto, MD, MPH, Angelo Ravelli, MD, Lucia Trail, MD: IRCCS G. Gaslini, Pediatria II, Reumatologia, Pediatric Rheumatology International Trials Organization, Genoa, Italy; [2]Sheila Oliveira, MD: Instituto de Puericultura e Pediatria Martagao Gesteira, Rio de Janeiro, Brazil; [3]Maria Alessio, MD: Università Federico II, Naples, Italy; [4]Dimitrina Mihaylova, MD: University Children Hospital,

Table 1. Final domains and suggested variables included in the core set for the evaluation of response to therapy in juvenile SLE (adapted from ref 12)*

Final domains	Suggested variable(s)
Physician's global assessment of the patient's overall disease activity	10-cm VAS
Renal involvement	24-hour proteinuria
Global juvenile SLE disease activity tool	ECLAM (or SLEDAI or SLAM)
Parent's global assessment of the patient's overall well-being	10-cm VAS
Health-related quality of life assessment	CHQ physical summary score

* SLE = systemic lupus erythematosus; VAS = visual analog scale; ECLAM = European Consensus Lupus Activity measurement; SLEDAI = Systemic Lupus Erythematosus Disease Activity Index; SLAM = Systemic Lupus Activity Measure; CHQ = Child Health Questionnaire.

ease activity (8), as well as criteria for the steroid-sparing ability of interventions in SLE (9).

The performance of clinical trials in juvenile SLE is made difficult by the small number of eligible patients, the heterogeneity of disease manifestations, and the lack of standardized criteria to assess clinical response. Additionally, there is little consensus about the amount of change in end points that signifies clinically important improvement or worsening. Methodologic advances in the definition of clinical response constitute a suitable manner in which to overcome some of these problems. Standardized criteria for juvenile SLE trials would provide a common basis for comparing different treatment options, permit study and statistical comparison of patients with different disease manifestations, and facilitate comparisons of different clinical trials using meta-analysis.

To help standardize the conduct and reporting of juvenile SLE clinical trials and enhance identification of new therapeutic agents, the Pediatric Rheumatology International Trials Organization (PRINTO) (10), in collaboration with the Pediatric Rheumatology Collaborative Study Group and with the support of the European Union and the US National Institutes of Health, undertook a multinational effort that was aimed at developing, and promulgating a core set of outcome measures and a definition of clinical improvement to evaluate response to therapy in patients with juvenile SLE. The first 2 phases of the project, which were previously reported (11,12), led to the development of a prospective, evidence-based, validated core set of 5 domains for the evaluation of response to therapy (Table 1).

In this article, we report the results of the third phase of the project, the aim of which was to develop a validated definition of improvement to aid in the classification of individual patients included in future therapeutic trials as either improved or not improved. We anticipate that a standardized definition should help physicians decide, in the clinical setting, whether a child with SLE has responded adequately to therapy.

PATIENTS AND METHODS

The overall methodology of this phase of the project was based on the methodologic framework and good results obtained with previous studies in rheumatoid arthritis (13), juvenile arthritis (14–16), and inflammatory myopathies (17). The 5 validated domains included in the final core set of variables for the evaluation of response to

Sofia, Bulgaria; [5]Srdjan Pasic, MD, PhD: Mother and Child Health Institute, Belgrade, Serbia and Montenegro; [6]Elisa-Italy; [7]Maria Apaz, MD: Universidad Catolica, Cordoba, Argentina; [8]Ruben Burgos-Vargas, MD: Hospital General de Mexico, Mexico City, Mexico; [9]Florence Kanakoudi-Tsakalidou, MD: Aristotle University Ippokration General Hospital, Thessalonika, Greece; [10]Ximena Norambuena, MD: Hospital Dr. Exequiel Gonzalez Cortes, Santiago, Chile; [11]Fabrizia Corona, MD: Clinica Pediatrica II "De Marchi," Milan, Italy; [12]Valeria Gerloni, MD: Istituto Gaetano Pini, Milan, Italy; [13]Stefan Hagelberg, MD, PhD: Karolinska University Hospital, Stockholm, Sweden; [14]Amita Aggarwal, MD: Sanjay Gandhi Institute of Medical Sciences, Lucknow, India; [15]Pavla Dolezalova, MD, PhD: 1st Faculty of Medicine and General Faculty Hospital, Prague, Czech Republic; [16]Claudia Magalhaes Saad, MD: Faculdade de Medicina de Botucatu, Universidade Estadual Paulista, Botucatu, Brazil; [17]Sang-Cheol Bae, MD, PhD: Hospital for Rheumatic Diseases, Hanyang University Medical Center, Seoul, South Korea; [18]Richard Vesely, MD: University Hospital, Kosice, Slovakia; [19]Tadej Avcin, MD, MSc: University Medical Centre Ljubljana, Ljubljana, Slovenia; [20]Helen Foster, MD: Royal Victoria Infirmary, Newcastle Upon Tyne, United Kingdom; [21]Carolina Duarte, MD: Centro Nacional de Rehabilitacion, Tlalpan Mexico DF, Mexico; [22]Troels Herlin, MD, PhD: Skejby Sygehus University Hospital, Aarhus, Denmark; [23]Gerd Horneff, MD: Universitatsklinik und Poliklinik fur Kinder-und Jugendmedizin, Halle, Germany; [24]Loredana Lepore, MD: Università degli Studi di Trieste, Trieste, Italy; [25]Marion van Rossum, MD: AMC/Emma Children's Hospital, Amsterdam, The Netherlands; [26]Angela Pistorio, MD, PhD: IRCCS G. Gaslini, Servizio di Epidemiologia e Biostatistica, Genoa, Italy; [27]Boel Andersson-Gäre, MD, PhD: Ryhov's County Hospital, Jönköping, Sweden; [28]Edward H. Giannini, MSc, DrPH: Children's Hospital Medical Center, Cincinnati, Ohio; [29]Alberto Martini, MD: IRCCS G. Gaslini, Pediatria II, Reumatologia and Università degli Studi, Genoa, Italy.

Address correspondence to Nicolino Ruperto, MD, MPH, Pediatric Rheumatology International Trials Organization (PRINTO), IRCCS G. Gaslini, Università di Genova Pediatria II–Reumatologia, Largo Gaslini, 5, 16147 Genoa, Italy. E-mail: nicolaruperto@ospedale-gaslini.ge.it.

Submitted for publication September 27, 2005; accepted in revised form February 17, 2006.

Table 2. Examples of 2 patients evaluated according to the PRINTO/ACR definition of improvement (at least 50% improvement from baseline in any 2 of the 5 core set measures with no more than 1 of the remaining worsening by more than 30%*

Variable	Month 0	Month 6	Absolute difference	% difference	Outcome
Patient 1					
Physician's global assessment of patient's overall disease activity (0–10-cm scale) ↑	7.2	1.6	−5.6	−78	Improved
Proteinuria, gm/24 hours ↑	9.3	1.4	−7.9	−85	Improved
ECLAM (range 0–10) ↑	9	4	−5	−56	Improved
Parent's global assessment of overall patient's well-being (0–10-cm scale) ↑	6.9	1.5	−5.4	−78	Improved
CHQ physical summary score ↓	29.9	49.3	19.4	65	Improved
Patient 2					
Physician's global assessment of patient's overall disease activity (0–10-cm scale) ↑	7.8	6.6	−1.2	−15	Not improved
Proteinuria gm/24 hours ↑	0.7	4	3.3	471	Not improved
ECLAM (range 0–10) ↑	10	10	0	0	Not improved
Parent's global assessment of overall patient's well-being (0–10-cm scale) ↑	9.2	9.7	0.5	5	Not improved
CHQ physical summary score ↓	16.9	19.0	2.1	13	Not improved

* ↑ indicates that a higher score for that variable denotes worse disease activity; ↓ indicates that a lower score denotes worse disease activity. PRINTO/ACR = Pediatric Rheumatology International Trials Organization/American College of Rheumatology; ECLAM = European Consensus Lupus Activity Measures; CHQ = Child Health Questionnaire.

therapy in juvenile SLE and the related suggested variables to measure each domain are shown in Table 1. The PRINTO juvenile SLE core set includes the following 5 clinical measures: 1) physician's global assessment of the patient's overall disease activity on a 10-cm visual analog scale (VAS); 2) global disease activity using the European Consensus Lupus Activity Measurement (18,19), the Systemic Lupus Erythematosus Disease Activity Index (20,21), or the Systemic Lupus Activity Measure (22); 3) renal involvement as assessed by 24-hour proteinuria; 4) parent's global assessment of the patient's overall well-being on a 10-cm VAS; and 5) health-related quality of life (HRQOL) assessment through the physical summary score of the parent's version of the Child Health Questionnaire (CHQ PhS) that measures the physical well-being of the child (23,24). The domains included in the core set underwent a careful evidence-based evaluation, which has been previously described (12). In particular, all domains were found to be feasible and to have good construct validity, discriminative ability, and internal consistency; furthermore, they were not redundant, proved responsive to clinically important change in disease activity, and were strongly associated with treatment outcome. Based on these key measurement properties, the domains were included in the final core set. It should be noted, however, that the recommended variables are only a minimal core set and that investigators can measure as many other variables as they deem appropriate for the major hypothesis that is being tested.

Following the selection of domains for the evaluation of response to therapy, a second consensus conference, entitled "International Consensus Conference on Defining Improvement in Juvenile SLE and Juvenile Dermatomyositis," was held in Camogli, Italy, on September 27–30, 2003. The meeting was attended by 37 experienced pediatric rheumatologists from 27 different countries to ensure wide international acceptance of the results of the project, and was facilitated by 4 of the authors (NR, EHG, BAG, AP) each of whom has expertise in nominal group process (25). Briefly, nominal group technique is a structured face-to-face meeting, with round-robin guided discussion, designed to facilitate reaching consensus among a group of experts. The overall goal of the meeting was to reach consensus on a validated definition of improvement based on the PRINTO core set of end points, using a combination of statistical criteria and consensus formation techniques. To achieve this objective, 5 steps were pursued, which are described below.

Step 1. Using nominal group technique, rate each of 128 patient profiles as "clinically importantly improved" or "not improved." Data for the 533 patients with juvenile SLE enrolled in the validation phase of the study (12) were used to select a subgroup of 128 patient profiles that were presented to conference attendees for the evaluation of a therapeutic response. The profiles selected were those that were judged by the conference organizers to be near a putative threshold level of improvement (e.g., patients who showed 100% improvement in all outcome variables were not good candidates for inclusion, because everyone would agree that such patients had improved, and that all of the definitions of improvement would categorize these patients as improved; likewise, for patients who showed 0% improvement in all outcome variables, everyone would agree that such patients were not improved/unchanged, and that all of the definitions of improvement would categorize these patients as not improved). Similar to what has been done for other definitions of improvement in patients with rheumatologic disease (13–17), each

profile contained only information related to the 5 validated PRINTO juvenile SLE core set measures (12). In each profile, and for each core set variable, we reported absolute values at baseline and 6 months and the absolute and percent change from baseline (Table 2). Participants were randomized into 3 equally sized "nominal groups" and asked to rate each of 128 patient profiles as clinically importantly improved or not improved, independently of the other participants. The moderator then asked each member how he or she had voted on each patient. If an 80% consensus was not achieved on whether the patient was improved or not improved, then the patient profile was discussed in a round-robin manner, and a second vote was taken. If 80% consensus was still not attained, the patient profile was declared uninterpretable. Profiles judged "uninterpretable" by 1 or more groups were rediscussed in a plenary session with nominal group technique; if consensus >80% was reached, then the patient profile was retained, otherwise it was deleted from further analysis. It was expected that consensus would be reached for at least 80% of the patients discussed.

Step 2. Several statistical evaluations (see below) were performed, using the physicians' consensus judgment as the "gold standard measure" to identify the best definition of improvement. Because no definitions of improvement that used combinations of the core set variables existed in the literature, we tested 597 different definitions of improvement that were deemed clinically reasonable and that were classified as generic and specific.

Each generic definition required improvement by at least x% in at least k of the 5 core measures, with no more than m other measures showing worsening of 30% or more. The combinations of x = 20%, 30%, 40%, or 50%, k = 2, 3, 4, or 5, and m = 0, 1, or 2 were considered. For the variables that worsened, the amount of worsening selected was based on the median changes observed in the entire data set (12). An example of a generic definition is as follows: at least 20% improvement from baseline in any 2 of the 5 core set measures with no more than 1 of the remaining measures worsening by more than 30%.

Each specific definition required improvement by at least x% in at least k of the 5 specific core measures, with no more than m other measures showing worsening of 30% or more. The combinations of x = 20%, 30%, 40%, or 50%, k = 1 specific key variable (e.g., physician's global assessment of the patient's overall disease activity) alone or in combination with 1 or 2 other specific key variables, and m = 0, 1, or 2 were considered. An example of a specific definition is as follows: physician's global assessment of the patient's overall disease activity and 24-hour proteinuria improved by at least 30%, 2 of any remaining 3 measures improved by at least 20%, and none worsened by more than 30%.

We evaluated the ability of the 597 candidate definitions of improvement to classify individual patients as improved or not improved, and then assessed the agreement between the definitions and consensus of the physicians. We used only patient profiles for which physician consensus was achieved. For each definition, we calculated the chi-square test (1 degree of freedom) and the corresponding *P* value, sensitivity (ability of the definition to identify a patient as improved who had been classified as improved by the physicians), specificity (ability of the definition to identify a patient as not improved who had been classified as not improved by the physicians), the false-positive rate (percent falsely identified as improved by criteria/all patients identified as improved × 100), the false-negative rate (percent falsely identified as not improved by the criteria/all patients identified as not improved × 100), and area under the receiver operating characteristic curve (ROC) (26). Moreover, the kappa statistic (27) was used to measure the strength of agreement between the definitions and consensus of the physicians, using the following cut-offs as proposed by Landis and Koch (28): 0.01–0.2 = slight, 0.21–0.4 = fair, 0.41–0.6 = moderate, 0.61–0.8 = substantial, and 0.81–1 = almost perfect agreement. Only definitions with kappa statistics >0.7 (substantial agreement), sensitivity and specificity >80%, and false-positive and false-negative rates of <20% were retained, while the remaining were eliminated from further considerations. On the next day, the results of the statistical analyses were presented to the conference attendees.

Step 3. Using nominal group technique, decide upon which of the remaining definitions of improvement is easiest to use and most credible (highest content validity). The attendees were again randomly split into 3 groups and, using nominal group technique, were asked to determine which of the definitions of improvement that performed best were easiest to use and most credible (content validity), ranking the 5 best from 1 (lowest) to 5 (highest content validity).

Step 4. The content validity score was multiplied by the kappa values to obtain the "best" definition. For each definition, the 3 content validity rankings obtained by the 3 nominal groups were summed, and the resulting sum was multiplied by the corresponding value of the kappa statistic, to obtain the final score that incorporated both statistical evaluations and experts' judgments.

Step 5. Concordance between response assessments made in the core set validation phase and those made at the consensus conference was assessed. Concordance between evaluation of the response to therapy (improved versus stable/not improved) made independently by the attending physicians and the parents at the time of the prospective data collection (12) and that made by the consensus conference attendees (improved versus not improved) were assessed again by means of kappa statistics (27) using the cut-offs proposed by Landis and Koch (28).

Association between changes in each of the 5 core measures and the overall outcome. The association between the change in each core set measure and the evaluation of response to therapy was analyzed by multiple logistic regression, which used as explanatory variables the baseline–to–6-month change in each core set variable and as dependent outcome the physicians' consensus evaluation

Table 3. Comparison of baseline demographic features and the baseline and 6-month values of the PRINTO juvenile SLE core set variables between the patients evaluated at the consensus conference (n = 128) and the rest of the sample collected (n = 405) for the validation of the final core set for the evaluation of response to therapy*

	Month 0			Month 6		
	Validation patients	Consensus patients	*P*	Validation patients	Consensus patients	*P*
Demographic variables						
Age at onset, years	12.1 ± 2.9	11.8 ± 3.2	0.4†			
Age at first observation, years	12.6 ± 2.9	12.5 ± 2.8	0.7†			
Age at diagnosis, years	12.7 ± 2.9	12.2 ± 3.1	0.2†			
Age at study visit, years	13.6 ± 2.8	13.9 ± 2.6	0.3†			
Disease duration, years	1.5 ± 2.3	2.1 ± 2.5	0.045†			
Sex, no. (%) female	351 (82)	108 (84)	0.4‡			
Core set variables						
Physician's global assessment of patient's overall disease activity (0–10-cm scale) ↑	5.8 ± 2.7	5.8 ± 2.5	0.9†	1.3 ± 1.8	3.3 ± 2.4	< 0.0001§
ECLAM (range 0–10) ↑	6.1 ± 2.5	5.9 ± 2.4	0.5†	1.9 ± 1.9	3.3 ± 2.3	< 0.0001§
Proteinuria, gm/24 hours ↑	1.0 ± 1.8	1.5 ± 2.4	0.2§	0.4 ± 0.9	0.9 ± 1.9	< 0.0001§
Parent's global assessment of patient's overall well-being (0–10 cm scale) ↑	4.6 ± 3.0	3.6 ± 3.2	0.002†	1.3 ± 1.9	2.6 ± 2.6	< 0.0001§
CHQ physical health summary score (range 40–60) ↓	38.3 ± 12.4	39.7 ± 11.9	0.3†	49.3 ± 8.0	45.6 ± 10.4	< 0.0001§

* Except where indicated otherwise, values are the mean ± SD. ↑ indicates that a higher score for that variable denotes worse disease activity; ↓ indicates that a lower score denotes worse disease activity. PRINTO = Pediatric Rheumatology International Trials Organization; SLE = systemic lupus erythematosus; ECLAM = European Consensus Lupus Activity Measures; CHQ = Child Health Questionnaire.
† By *t*-test for independent samples.
‡ By Pearson's chi-square test.
§ By Mann-Whitney U test for independent samples.

of patient improvement. Variables were dichotomized according to the best cut-offs provided by the ROC analysis (26). Determining the best cut-offs for each core set variable will help physicians decide if a patient is improved based on the absolute change in that particular measure.

Data were entered in an Access XP database and analyzed with Excel XP software (Microsoft, Redmond, WA), *XL*STAT 6.1.9 software (Addinsoft, Brooklyn, NY), Statistica 6.0 software (StatSoft, Tulsa, OK), and Stata version 7.0 software (Stata, College Station, TX).

RESULTS

Table 3 shows the demographic features and the baseline and 6-month values for the core set variables in the subgroup of 128 patients used to create the patient profiles for the consensus conference and in the remaining 405 patients collected for the validation of the core set for the evaluation of response to therapy (12). The demographic features of the 2 cohorts were comparable, although the consensus patients had slightly longer disease duration. At baseline, the 2 cohorts were comparable for all core set variables except the parent's global assessment of the patient's overall well-being, the score for which was lower among consensus patients. The finding that at 6 months values for all core set variables were worse in consensus patients was expected, because this subgroup comprised patients who were near a putative threshold level of improvement (see Patients and Methods), whereas the other subgroup included the remaining patients who achieved the most pronounced levels of improvement.

Results of scoring the patient profiles. Consensus ≥80% was achieved for 109 (85%) of the 128 patients, with 70 (64%) of the 109 patients being judged as achieving clinically important improvement, and 39 (36%) of the 109 patients being judged as not improved. In no case did 1 nominal group rate a patient as improved and the other 2 groups rate the same patient as not improved.

Identification of 10 definitions of improvement as the best performers. Ten of the 597 definitions of improvement reached a kappa value of ≥0.7 (substantial agreement); the corresponding chi-square values, *P* values, sensitivity, specificity, percent false-positive and false-negative rates, area under the curve, and kappa statistics are shown in Table 4.

Content validity of the 10 definitions of improvement and final resolution. After presentation of the above data, the attendees, using nominal group technique, selected the 5 best definitions for content validity and ranked them on a 1–5 scale, with 5 being the highest. The sum of the combined ranks from the 2 groups is presented in Table 4 (minimum–maximum 2–150). Then, the sum of the ranking was multiplied by its kappa statistic to obtain the final score (minimum–maximum 2–118), and the definitions of

Table 4. Final results for the 10 best definitions of improvement (DI)*

Definition of improvement	χ^2†	Sensitivity, %	Specificity, %	False-negative rate, %	False-positive rate, %	AUC	κ	Rank	Final score
DI 11. 2 of any 5 improved by at least 50%, no more than 1 worse by more than 30%	67	90	90	17	6	90	0.78	150	118
DI 8. 2 of any 5 improved by at least 40%, no more than 1 worse by more than 30%	63	93	82	14	10	87	0.76	99	75
DI 17. 3 of any 5 improved by at least 30%, no more than 1 worse by more than 30%	63	81	97	25	2	89	0.74	87	64
DI 12. 2 of any 5 improved by at least 50%, no more than 2 worse by more than 30%	67	94	85	11	8	89	0.80	74	59
DI 14. 3 of any 5 improved by at least 20%, no more than 1 worse by more than 30%	64	84	95	23	3	90	0.75	29	22
DI 9. 2 of any 5 improved by at least 40%, no more than 2 worse by more than 30%	63	93	82	14	10	87	0.76	23	17
DI 5. 2 of any 5 improved by at least 30%, no more than 1 worse by more than 30%	59	94	77	12	12	86	0.73	13	10
DI 6. 2 of any 5 improved by at least 30%, no more than 2 worse by more than 30%	60	99	69	4	15	84	0.72	6	4
DI 18. 3 of any 5 improved by at least 30%, no more than 1 worse by more than 30%	59	81	95	26	3	88	0.72	4	3
DI 15. 3 of any 5 improved by at least 20%, no more than 2 worse by more than 30%	63	86	92	22	5	89	0.75	2	2

* Definitions are ordered according to the final score. AUC = area under the curve; Rank = consensus attendees selected which definitions of improvement performed best, were easiest to use, and most credible (content validity); Final score = the sum of the content validity rankings was multiplied by the corresponding kappa statistic.
† $P < 0.0001$

improvement with the highest final score were identified. The definition of improvement that scored highest was the following: at least 50% improvement in any 2 of the 5 core set variables, with no more than 1 of the remaining variables deteriorating by more than 30%.

As seen in Table 4, the definitions that ranked second and third highest are similar to the highest-ranking definition but required a lower degree of improvement. The similarity of the top-ranking definitions indicates convergent validity of the measures. Because the statistical performance of the 10 best definitions all had kappa statistics >0.7, the selection of the final definition of improvement was driven mainly by the ranking (content validity) of the top 5 definitions.

Concordance in the evaluation of response. The level of agreement in the evaluation of response to therapy between the physicians who assessed the patients in the validation phase and those who assessed the patients during the consensus conference was in the moderate range ($\kappa = 0.4$, 95% confidence interval [95% CI] 0.2–0.6). The level of agreement in the evaluation of response to therapy between the parents who assessed the patients in the validation phase and the physicians who assessed the patients during the consensus conference was in the fair range ($\kappa = 0.3$, 95% CI 0.1–0.5). To explain the observed level of agreement, it should be specified that the physicians attending the consensus conference made their evaluation based only on changes in the 5 core set variables, while physicians who assessed the patients in the validation phase were the attending physicians who judged the complete clinical status of their patients. The same holds true for the parents who assessed their children. Furthermore, it should be remembered that, for statistical purposes, we chose to create the profiles of those patients in whom the clinical change from baseline to 6 months was less pronounced (and thus potentially more controversial) in the direction of either improvement or worsening.

To obtain further insights into this issue, we evaluated the level of concordance between the classification derived from the "top definition of improvement" and either attending physicians' or parents' evaluations in all 533 patients (12). Compared with the previous analysis, the level of agreement with attending physicians increased from 0.4 to 0.5, and the level of agreement with parents increased from 0.3 to 0.5. These kappa statistics fall in the moderate range (28). Bearing in mind that the evaluation of a patient in the clinical setting implies different issues than those involved in a clinical trial, this level of concordance is more than acceptable, particularly when taking into account the fact that no standards for comparison with our findings are available. Indeed a similar analysis has never been attempted for other rheumatic diseases (13,14,17) for which a standard definition of clinical improvement was based only on the results of consensus conferences and not on ad hoc prospectively collected data as were used for this project.

Further refinement of the definition of improvement. After selecting the definition of improvement, the consensus organizers were asked to perform a further analysis on the top definitions by adding the requirement that the 24-hour proteinuria cannot worsen. The inclusion of this "contingency" led to definitions with lower statistical per-

Table 5. Results of logistic regression to predict improvement according to the evaluation of the participants at the consensus conference*

Variable	OR	95% CI	*P*, likelihood ratio test
Physician's global assessment of patient's overall disease activity (0–10-cm scale) ↑	25.4	5.5–116.2	< 0.0001
CHQ physical summary score	12.2	2.1–72.2	0.0021
ECLAM (range 0–10) ↑	8.7	1.9–38.9	0.0028
Proteinuria, gm/24 hours ↑	4.1	0.9–19.3	0.0623
Parent's global assessment of patient's overall well-being (0–10-cm scale) ↑	1.7	0.3–9.2	0.547

* Sample entered in the model was equal to 109 patients and the area under ROC curve of the model = 0.96. ↑ indicates that a higher score for that variable denotes worse disease activity; ↓ indicates that a lower score denotes worse disease activity. Prediction was based on absolute change of the variables included in the final core set. Variables have been dichotomized according to the best cut-offs obtained from the receiver operating characteristic curve analysis. Best cut-offs for the variables included in the model were as follows: for physician's global assessment of patient's overall disease activity, ≤ −1.7 (sensitivity 91.4% specificity 79.5%); for physical well-being of the CHQ, >3.6 (sensitivity 72.9% specificity 82.1%); for ECLAM ≤ −2 (sensitivity 82.9% specificity 71.8%); for 24-hour proteinuria, ≤ −0.1 (sensitivity 58.6% specificity 79.5%); for parent's global assessment of patient's overall well-being, ≤ −1.0 (sensitivity 58.6% specificity 79.5%). Area under the curve = 0.96. OR = odds ratio; 95% CI = 95% confidence interval; CHQ = Child Health Questionnaire; ECLAM = European Consensus Lupus Activity Measures.

formance; for example, for definition 11 the sensitivity decreased from 90% to 84%, the specificity increased from 90% to 92%, and the kappa value decreased from 0.78 to 0.73.

Association between changes in each of the 5 core measures and the overall outcome. The association between the change in each core set measure and response to therapy was analyzed in a multivariate analysis, which used as explanatory variables in baseline–to–6-month change in each of the 5 core set variables and as dependent outcome the physician's consensus evaluation of the patient's improvement. In the final model (Table 5), the physician's global assessment of the patient's overall disease activity appeared to be the strongest predictor of response to therapy (odds ratio [OR] 25.4), followed by the CHQ-physical health well-being, and the European Consensus Lupus Activity Measurement (OR 12.2 and 8.7, respectively), whereas 24-hour proteinuria and the parent's global assessment of the patient's overall well-being, despite having ORs in the right direction (OR 4.1 and 1.7, respectively), did not reach the level of statistical significance.

Practical application of the validated PRINTO definition of improvement. The domains and suggested variables included in the final core set for the evaluation of response to therapy in juvenile SLE are shown in Table 1. The suggested variables to measure each domain are those used for validation of the core set and of the definition of improvement but researchers can use other variables that might be more appropriate based on their study design or new validation data that will appear in the literature in the future.

Two examples with data from real patients used at the consensus conference are shown in Table 2.

DISCUSSION

Using a combination of data-driven and consensus-formation processes, pediatric rheumatologists with specific experience in the assessment of juvenile SLE developed a validated definition of improvement that PRINTO proposes for inclusion in future juvenile SLE clinical trials. Based on the top definition, improvement in individual patients with juvenile SLE can be defined as follows: improvement in any 2 of 5 core set variables by at least 50% versus baseline, with no more than 1 of the remaining variables worsening by more than 30%.

It is interesting to note that during the nominal group discussion, the consensus conference attendees pointed out that, given the similarity in kappa agreement for the top definitions, when evaluating a severe disease such as juvenile SLE, they prefer to aim for the highest improvement suggested by statistical analysis (50% improvement). Indeed, the 50% improvement recommended for juvenile SLE is higher than the 30% improvement requested for juvenile arthritis (14) and the 20% improvement for juvenile myositis (29) and the inflammatory myopathies (17).

The PRINTO definition includes objective measures, such as a global measure of SLE activity and measurement of 24-hour proteinuria, and a physician's subjective assessment of the level of disease activity, but it also considers parent-reported outcomes, such as assessment of overall well-being and HRQOL. The definition selected by the consensus panel performed well in the available data set, with high sensitivity and specificity, and low false-positive and false-negative rates. Furthermore, the definition revealed a good ability to discriminate between patients who improved and those who did not. The consensus process indicated that this definition had the best content validity as well.

Besides the consensus of a large number of experienced pediatric rheumatologists from many countries that provided wide international acceptance to the project, and its good statistical properties, the strengths of this definition are its evidence-based selection process and the validation of its core set components (12), which were performed in a very large sample of patients assessed in a prospective manner.

The validated definition of improvement was based on a composite of outcome measures that were set up to detect a broad range of clinical change. Until now, single-organ measures have been used in most SLE clinical trials (e.g., in SLE nephritis). Advantages and disadvantages are associated with each approach (4,8). Although the use of measures related to single-organ involvement certainly provides more meaningful information to the trial, this focus limits information on the clinical status of patients and thus the value of the results. Assessing multiple organ systems alone is impractical because it would lead to assessment of only a small number of patients, due to disease heterogeneity. Alternatively, the use of measures of SLE activity as a whole would "dilute" measures related to a particular organ, because of contributions from other systems. However, it has been suggested that the use of comprehensive and nonredundant pooled outcome measures offers the advantage of increased clinical validity and improved sensitivity (30). Furthermore, because juvenile SLE has a broad phenotype, there is concern about the ability of any single measure to capture the treatment effect reliably.

In contrast to the JIA response criteria (14), the core set of measures in juvenile SLE probably will not cover all changes brought on by trials of potential therapeutic approaches ranging from topical treatment to the more potent immunosuppressive regimens. For these reasons, we believed that the primary function of response criteria is to provide information related to the patient as a whole. Therefore, we included a disease activity tool, which is likely to incorporate any change in major organ manifestations, and physician- and patient-centered outcome measures as suggested by different groups of investigators (5–9,31). These criteria may constitute a secondary end point in a trial focused on a patient's primary problem or organ involvement that would be the primary end point. To our knowledge, no evidence-based information exists on the relative performance of organ-specific versus broader measures of response, because none of the previous SLE trials compared these 2 sets of end points.

Our study should be viewed in light of certain limitations, which include the facts that it was not conducted in the context of a real clinical trial, and that the PRINTO-validated definition of improvement showed a 17% false-negative rate; this aspect should be further evaluated in future studies. The main strength of the study resides in the prospective collection of a large amount of data, which has never been attempted for other rheumatic diseases (13,14,17); this approach ensured an evidence-based validation analysis of the juvenile SLE core set (12) and provided data for the consensus conference evaluations.

In summary, PRINTO investigators developed a validated definition of improvement that will help standardize the conduct of juvenile SLE clinical trials and assist clinicians in the classification in daily practice of patients as being either responder or nonresponder. In the absence of available therapeutic trial data in juvenile SLE, this definition deserves validation in future controlled studies to examine its discriminant validity in detecting a therapeutic response greater than that of placebo or the active comparator, and to assess whether further refinements of the currently available instruments are required.

ACKNOWLEDGMENTS

We are indebted to Drs. Anna Tortorelli, Monica Tufillo, and Elisabetta Maggi for their help in data handling, their organization skills, and overall management of the project. We are also thankful to Dr. Luca Villa and Mr. Michele Pesce for their help in database development.

We would like to acknowledge the organizers, attendees, and external observers of the Camogli, Italy International Consensus Conference on defining improvement in juvenile SLE and JDM for their work during the meeting: organizers Alberto Martini, MD, Nicolino Ruperto, MD, MPH, Angelo Ravelli, MD, Angela Pistorio, MD, PhD (Italy); Edward H Giannini, MSc, DrPH, Daniel J Lovell, MD, MPH (United States); and Boel Andersson-Gäre, MD, PhD (Sweden); attendees Carmen de Cunto, MD, Ruben Cuttica, MD (Argentina); Rik Joos, MD (Belgium); Claudia Magalhaes Saad, MD, Sheila Oliveira, MD (Brazil); Dimitrina Mihaylova, MD (Bulgaria); Brian Feldman, MD (Canada); Miroslav Harjacek, MD (Croatia); Pavla Dolezalova, MD (Czech Republic); Susan Nielsen, MD (Denmark); Pekka Lahdenne, MD (Finland); Anne Marie Prieur, MD (France); Hans Iko Huppertz, MD (Germany); Florence Kanakoudi Tsakalidou, MD (Greece); Yosef Uziel, MD (Israel); Ingrida Rumba, MD (Latvia); Ruben Burgos Vargas, MD (Mexico); Nico Wulffraat, MD (The Netherlands); Berit Flato, MD (Norway); Malgorzata Wierzbowska, MD (Poland); Jose Antonio Melo-Gomes, MD (Portugal); Gordana Susic, MD (Serbia and Montenegro); Richard Vesely, MD (Slovakia); Tadej Avcin, MD (Slovenia); Michael Hofer, MD (Switzerland); Huri Ozdogan, MD (Turkey); Clarissa Pilkington, MD, Madeleine Rooney, MD (United Kingdom); Lisa Rider, MD, Phil Hashkes, MD, Anne Reed, MD, Robert Rennebohm, MD, Lauren Pachman, MD, and Carol Wallace, MD (United States); and external observers Marcia Bandeira, MD (Brazil); Jenny Pratsidou, MD (Greece); Stella Maris Garay, MD (Argentina).

REFERENCES

1. Flanc RS, Roberts MA, Strippoli GF, Chadban SJ, Kerr PG, Atkins RC. Treatment of diffuse proliferative lupus nephritis: a meta-analysis of randomized controlled trials. Am J Kidney Dis 2004;43:197–208.
2. Contreras G, Pardo V, Leclercq B, Lenz O, Tozman E, O'Nan P, et al. Sequential therapies for proliferative lupus nephritis. N Engl J Med 2004;350:971–80.
3. Lockshin MD. Therapy for systemic lupus erythematosus. N Engl J Med 1991;324:189–91.
4. Schiffenbauer J, Hahn B, Weisman MH, Simon LS. Biomarkers, surrogate markers, and design of clinical trials of new therapies for systemic lupus erythematosus. Arthritis Rheum 2004;50:2415–22.

5. Strand V, Gladman D, Isenberg D, Petri M, Smolen J, Tugwell P. Outcome measures to be used in clinical trials in systemic lupus erythematosus. J Rheumatol 1999;26:490–7.
6. Smolen JS, Strand V, Cardiel M, Edworthy S, Furst D, Gladman D, et al. Randomized clinical trials and longitudinal observational studies in systemic lupus erythematosus: consensus on a preliminary core set of outcome domains. J Rheumatol 1999;26:504–7.
7. Liang MH, Corzillius M, Bae SC, Fortin P, Esdaile JM, Abrahamowicz M. A conceptual framework for clinical trials in SLE and other multisystem diseases. Lupus 1999;8:570–80.
8. American College of Rheumatology Ad Hoc Committee on Systemic Lupus Erythematosus Response Criteria. The American College of Rheumatology response criteria for systemic lupus erythematosus clinical trials: measures of overall disease activity. Arthritis Rheum 2004;50:3418–26.
9. Ad Hoc Working Group on Steroid-Sparing Criteria in Lupus. Criteria for steroid-sparing ability of interventions in systemic lupus erythematosus: report of a consensus meeting. Arthritis Rheum 2004;50:3427–31.
10. Ruperto N, Martini A. International research networks in pediatric rheumatology: the PRINTO perspective. Curr Opin Rheumatol 2004;16:566–70.
11. Ruperto N, Ravelli A, Murray KJ, Lovell DJ, Andersson-Gare B, Feldman BM, et al, and the Paediatric Rheumatology International Trials Organization (PRINTO), Pediatric Rheumatology Collaborative Study Group (PRCSG). Preliminary core sets of measures for disease activity and damage assessment in juvenile systemic lupus erythematosus and juvenile dermatomyositis. Rheumatology (Oxford) 2003;42:1452–9.
12. Ruperto N, Ravelli A, Cuttica R, Espada G, Ozen S, Porras O, et al, and the Pediatric Rheumatology International Trials Organization (PRINTO), Pediatric Rheumatology Collaborative Study Group (PRCSG). The Pediatric Rheumatology International Trials Organization criteria for the evaluation of response to therapy in juvenile systemic lupus erythematosus: prospective validation of the disease activity core set. Arthritis Rheum 2005;52:2854–64.
13. Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, et al. American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727–35.
14. Giannini EH, Ruperto N, Ravelli A, Lovell DJ, Felson DT, Martini A. Preliminary definition of improvement in juvenile arthritis. Arthritis Rheum 1997;40:1202–9.
15. Ruperto N, Ravelli A, Falcini F, Lepore L, de Sanctis R, Zulian F, et al, and the Italian Pediatric Rheumatology Study Group. Performance of the preliminary definition of improvement in juvenile chronic arthritis patients treated with methotrexate. Ann Rheum Dis 1998;57:38–41.
16. Albornoz MA. ACR formally adopts improvement criteria for juvenile arthritis (ACR Pediatric 30). ACR News 2002;21:3.
17. Rider LG, Giannini EH, Brunner HI, Ruperto N, James-Newton L, Reed AM, et al. International consensus on preliminary definitions of improvement in adult and juvenile myositis. Arthritis Rheum 2004;50:2281–90.
18. Vitali C, Bencivelli W, Isenberg DA, Smolen JS, Snaith ML, Sciuto M, et al, and the European Consensus Study Group for Disease Activity in SLE. Disease activity in systemic lupus erythematosus: report of the Consensus Study Group of the European Workshop for Rheumatology Research. II. Identification of the variables indicative of disease activity and their use in the development of an activity score. Clin Exp Rheumatol 1992;10:541–7.
19. Vitali C, Bencivelli W, Mosca M, Carrai P, Sereni M, Bombardieri S. Development of a clinical chart to compute different disease activity indices for systemic lupus erythematosus. J Rheumatol 1999;26:498–501.
20. Bombardier C, Gladman DD, Urowitz MB, Caron D, Chang CH, and the Committee on Prognosis Studies in SLE. Derivation of the SLEDAI: a disease activity index for lupus patients. Arthritis Rheum 1992;35:630–40.
21. Brunner HI, Feldman BM, Bombardier C, Silverman ED. Sensitivity of the Systemic Lupus Erythematosus Disease Activity Index, British Isles Lupus Assessment Group Index, and Systemic Lupus Activity Measure in the evaluation of clinical change in childhood-onset systemic lupus erythematosus. Arthritis Rheum 1999;42:1354–60.
22. Liang MH, Socher SA, Larson MG, Schur PH. Reliability and validity of six systems for the clinical assessment of disease activity in systemic lupus erythematosus. Arthritis Rheum 1989;32:1107–18.
23. Landgraf JM, Abetz L, Ware JE. The CHQ user's manual. 1st ed. Boston: The Health Institute, New England Medical Center; 1996.
24. Martini A, Ruperto N, for the Pediatric Rheumatology International Trials Organization (PRINTO). Quality of life in juvenile idiopathic arthritis patients compared to healthy children. Clin Exp Rheumatol 2001;19 Suppl 23:S1–172.
25. Delbecq AL, van de Ven AH, Gustafson DH. Group techniques for program planning: a guide to nominal group and Delphi processes. 1st ed. Glenview (IL): Scott, Foresman and Company; 1975.
26. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283–98.
27. Cohen J. Statistical power analysis for the behavioral sciences. New York: Academic Press; 1977.
28. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
29. Ruperto N, Woo P, Cuttica R, Cortis E, Alessio M, Porras O, et al. The validated PRINTO core set and definition of improvement for juvenile myositis [abstract]. Arthritis Rheum 2004;50 Suppl 9:S534.
30. Schneider M. Response and remission criteria for clinical trials in lupus: what can we learn from other diseases? Lupus 1999;8:627–31.
31. Brunner HI, Giannini EH. Health-related quality of life in children with rheumatic diseases. Curr Opin Rheumatol 2003;15:602–12.