


American College of Rheumatology Provisional Criteria for Global Flares in Childhood-Onset Systemic Lupus Erythematosus

HERMINE I. BRUNNER,¹ MICHAEL HOLLAND,¹ MICHAEL W. BERESFORD,² STACY P. ARDOIN,³ SIMONE APPENZELLER,⁴ CLOVIS A. SILVA,⁵ FRANCISCO FLORES,¹ BEATRICE GOILAV,⁶ SCOTT E. WENDERFER ,⁷ DEBORAH M. LEVY,⁸ ANGELO RAVELLI,⁹ RAJU KHUNCHANDANI,¹⁰ TADEJ AVCIN,¹¹ MARISA S. KLEIN-GITELMAN,¹² BRIAN M. FELDMAN,⁸ NICOLINO RUPERTO,⁹ JUN YING,¹³ FOR THE PRCSG AND PRINTO INVESTIGATORS*

Objective. To validate the preliminary criteria of global flare for childhood-onset SLE (cSLE).

Methods. Pediatricians experienced in cSLE care (n = 268) rated unique patient profiles; results of standard cSLE laboratory testing and information about the cSLE flare descriptors were presented as follows: global assessment of patient well-being, physician global assessment of disease activity (MD-global), Disease Activity Index score, protein/creatinine ratio (PCR), and erythrocyte sedimentation rate (ESR). Using rater interpretation of the course of cSLE (baseline versus followup as the gold standard), performance (sensitivity, specificity, area under the receiver operating characteristic curve [AUC]) of the preliminary flare criteria was tested. An international consensus conference was held to rank the preliminary flare criteria as per the American College of Rheumatology recommendations and delineate threshold scores for minor, moderate, and major flares.

Results. The accuracy of the 2 highest-ranked candidate criteria that consider absolute changes (Δ) of the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) or British Isles Lupus Assessment Group (BILAG) (numeric scoring: A = 12, B = 8, C = 1, and D/E = 0), MD-global, PCR, and ESR were confirmed (both AUC >0.93). For the SLEDAI-based criteria ($0.5 \times \Delta\text{SLEDAI} + 0.45 \times \Delta\text{PCR} + 0.5 \times \Delta\text{MD-global} + 0.02 \times \Delta\text{ESR}$) flare scores $\geq 6.4/3.0/0.6$ constituted major/moderate/minor flares, respectively. For the BILAG-based algorithm ($0.4 \times \Delta\text{BILAG} + 0.65 \times \Delta\text{PCR} + 0.5 \times \Delta\text{MD-global} + 0.02 \times \Delta\text{ESR}$) flare scores $\geq 7.4/3.7/2.2$ delineated major/moderate/minor flares, respectively. These threshold values (SLEDAI, BILAG) were all >82% sensitive and specific for capturing flare severity.

Conclusion. Provisional criteria for global flares in cSLE are available to identify patients who experienced a flare. These criteria also allow for discrimination of the severity of cSLE exacerbations.

INTRODUCTION

Systemic lupus erythematosus (SLE) is a complex, chronic, multisystem autoimmune inflammatory disease, with up to 20% of patients diagnosed during childhood (cSLE) (1,2). When disease commences early in life rather than during adulthood, it has a less favorable prognosis, particularly due

to multiorgan and kidney involvement (3,4). The course of cSLE is characterized by episodes of disease flares, followed by periods of improvement, generally due to more intensive drug therapy. There is international consensus that a flare of cSLE is “a measurable worsening of disease activity in at least one organ system, involving new or worse signs of disease that may be accompanied by new or worse SLE

Supported by the NIH (grants 5U01-AR-51868, P30-AR-AR47363, and 2UL-1RR-026314) and by LUPUS UK, who supports the UK Juvenile-Onset Systemic Lupus Erythematosus Cohort Study, along with the NIHR Clinical Research Network (CRN), NIHR CRN Children's Specialty Group, and NIHR Alder Hey Clinical Research Facility. Dr. Silva's work was supported by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP 2015/03756-4), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq 303422/2015-7), and by Núcleo de Apoio à Pesquisa “Saúde da Criança e do Adolescente” da USP (NAP-CriAd).

¹Hermine I. Brunner, MD, MSc, Michael Holland, MD, Francisco Flores, MD: University of Cincinnati and Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio; ²Michael W. Beresford, MBChB, PhD: Institute of Translational Medicine

and Alder Hey Children's NHS Foundation Trust, Liverpool, UK; ³Stacy P. Ardoin, MD: Ohio State University, Nationwide Children's Hospital, and Wexner Medical Center, Columbus, Ohio; ⁴Simone Appenzeller, MD, PhD: University of Campinas, Campinas, Brazil; ⁵Clovis A. Silva, MD, PhD: Children's Institute, Hospital das Clínicas HCFMUSP, Universidade de São Paulo, São Paulo, Brazil; ⁶Beatrice Goilav, MD: Children's Hospital at Montefiore, Albert Einstein College of Medicine, Bronx, New York; ⁷Scott E. Wenderfer, MD, PhD: Baylor College of Medicine and Texas Children's Hospital Houston,

*Correction added after online publication 2 May 2018: The author byline has been corrected to include the name for the PRCSG and PRINTO Investigators.

Significance & Innovations

- Results of the preliminary validation of criteria of global flare for childhood-onset systemic lupus erythematosus are provided.
- Based on the flare scores, mild, moderate, and severe flares can be defined.

symptoms; depending on the severity of the flare, more intensive therapy may be required” (5). Further, using consensus formation techniques, agreement has been achieved regarding preliminary criteria of global flares of cSLE based on changes of the erythrocyte sedimentation rate (ESR), the protein/creatinine ratio (PCR), physician global assessment of cSLE activity (MD-global), and the score of the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) (6,7) or the British Isles Lupus Activity Group index (BILAG) (8). Moreover, there is consensus around the need to discriminate flares as per their severity: mild/minor, moderate, and major/severe flares (5). However, there are no generally accepted criteria or algorithms to determine how to measure the severity of cSLE flares, nor have the preliminary cSLE flare criteria been validated in an independent data set. Thus, the objectives of this phase of the project were to validate the preliminary criteria of global flare of cSLE and to apply consensus formation methodology to define flare threshold levels for minor, moderate, and major flares. These criteria were created to define cSLE flares and their severity for use in clinical trials.

PATIENTS AND METHODS

The overall approach to this project was based on the methodologic framework successfully employed in pediatric rheumatology criteria measures in the past (9–11), aligned with recommendations of the American College of Rheumatology (ACR) Criteria Subcommittee and the Quality of Care Committee (12). The initial results of the consensus process

resulting in preliminary cSLE flare criteria have been described elsewhere (5,13). Briefly, previous research demonstrated that the scores of a disease activity measure alone are inadequate for identifying flares (5). International agreement was reached regarding preliminary criteria to measure global flares of cSLE. Pediatric rheumatologists participated in Delphi surveys that yielded consensus around a common definition of cSLE global flares, and the delineation of cSLE flare descriptors. This was followed by exploration of candidate flare criteria (5) and the identification of preferred algorithms of global cSLE flares (14). Notably, data and analyses all suggested that uniform percentage changes of the cSLE flare descriptors are insufficient to capture cSLE flares with high sensitivity. Further, inclusion of the MD-global assessment of cSLE activity in highly accurate cSLE candidate flare algorithms proved necessary (5,15). During the first consensus conference, the top-performing candidate flare algorithms, derived either from multinomial logistic regression modeling or classification tree analysis (CART), were established.

We now present the phase of the project aimed at validating the preferred preliminary flare algorithms (14) via testing in an independent validation data set (Figure 1). These encompassed patient profile (PP) ratings that were requested from 503 pediatric rheumatologists from Australia, Africa, Asia, Europe, and the Americas who were members of at least 1 of the following organizations: the Pediatric Rheumatology Collaborative Study Group, the Childhood Arthritis Rheumatology Research Alliance, the Pediatric Rheumatology European Society Juvenile Lupus Working Group, and the Pan American League of Associations for Rheumatology (step 1).

The interpretation of the flare or “true” disease course of a given PP was determined using 2 approaches, which resulted in 2 distinct data sets for the subsequent validation exercises (step 2). Using the PP ratings, the preliminary criteria for cSLE global flares were tested for their ability to discriminate patients who experienced different levels of flares (minor, moderate, and major) (step 3). Subsequently, during a consensus conference, the validity of the criteria was critically reviewed, taking into consideration information from the medical literature, statistical performance, reliability, feasibility, and face validity as per the ACR guidance document (12) and the Outcome Measures in Rheumatology filter (16) (step 4).

Preliminary cSLE flare algorithms. We considered the top 4 preliminary flare algorithms (identified in the first consensus conference) based on feasibility, truthfulness, and discrimination (17). Two of the 4 preliminary cSLE flare algorithms (SLEDAI-based criteria: $0.5 \times \Delta\text{SLEDAI} + 0.45 \times \Delta\text{PCR} + 0.5 \times \Delta\text{MD-global} + 0.02 \times \Delta\text{ESR}$, and BILAG-based criteria: $0.4 \times \Delta\text{BILAG} + 0.65 \times \Delta\text{PCR} + 0.5 \times \Delta\text{MD-global} + 0.02 \times \Delta\text{ESR}$) were derived by multinomial logistic regression that considered several of the cSLE flare descriptors, and that yield “flare scores” (or log odds of flare), with a higher score representing a higher likelihood of a flare to have occurred. The other 2 algorithms of the top preliminary flare criteria were derived from CART (SLEDAI-CART: where score = 4 if $3 \leq \text{SLEDAI}$, score = 3 if $0.7 \leq \text{PCR}$ and $3 > \text{SLEDAI}$, score = 2 if $2 \leq \text{MD-global}$ and $0.7 > \text{PCR}$ and $3 > \text{SLEDAI}$, and score = 1 if otherwise, and BILAG-CART: where score = 4 if $2 \leq \text{BILAG}$, score = 3 if $0.7 \leq \text{PCR}$ and $2 > \text{BILAG}$, score = 2 if $2 \leq \text{MD}$ and $0.7 > \text{PCR}$ and $2 > \text{BILAG}$,

Texas; ⁸Deborah M. Levy, MD, MS, Brian M. Feldman, MD, MSc, FRCPC: University of Toronto and Hospital for Sick Children, Toronto, Ontario, Canada; ⁹Angelo Ravelli, MD, Nicolino Ruperto, MD, MPH: Clinica Pediatrica e Reumatologia, Istituto Giannina Gaslini, and Università degli Studi di Genova, Genoa, Italy; ¹⁰Raju Khunchandani, MD: Jaslok Hospital, Mumbai, India; ¹¹Tadej Avcin, MD, PhD: University Children's Hospital, University Medical Centre, Ljubljana, Slovenia; ¹²Marisa S. Klein-Gitelman, MD, MPH: Northwestern University Feinberg School of Medicine and Ann and Robert Lurie Children's Hospital of Chicago, Chicago, Illinois; ¹³Jun Ying, PhD: University of Cincinnati, Cincinnati, Ohio.

Drs. Brunner and Holland contributed equally to this work.

Address correspondence to Hermine I. Brunner MD, MSc, Cincinnati Children's Hospital Medical Center, University of Cincinnati, William S. Rowe Division of Rheumatology, E 4010, 3333 Burnet Avenue, Cincinnati, OH 45229-3039. E-mail: hermine.brunner@cchmc.org.

Submitted for publication July 8, 2017; accepted in revised form March 8, 2018.

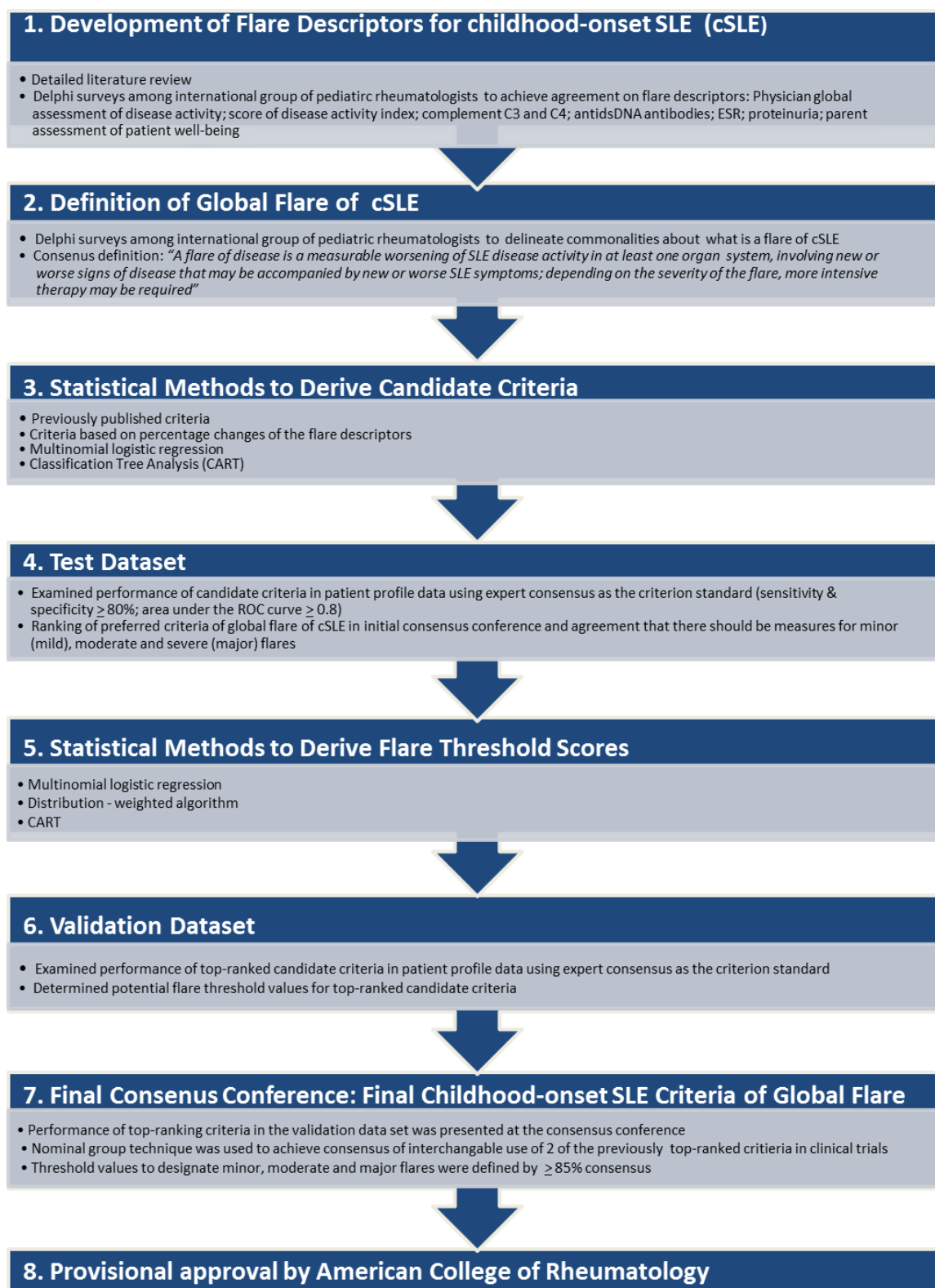


Figure 1. Flow diagram of the entire process used to develop and validate the approved criteria of global flare of cSLE. Steps 1–5 have been summarized in references 5 and 14. The current study begins at step 5 and focuses on steps 6–8. antidsDNA = anti-double-stranded DNA; ESR = erythrocyte sedimentation rate; ROC = receiver operating characteristic curve.

and score = 1 if otherwise). Similar to algorithms derived by multinomial logistic regression, CART-based criteria yield “CART-scores” that can be used to decide on the presence of a flare, including its severity (14).

Step A: PPs and ratings of disease course of a PP. Two of the authors (HIB and MH) conducted a pilot study to test the format of the PP. Built on this pilot study, we generated over 2,996 unique PPs, using prospectively collected data of cSLE patients from the Cincinnati Children’s Hospital Medical Center Lupus Registry (18), the Paediatric Rheumatology International Trials Organisation Lupus Cohort (19), the UK Juvenile-Onset SLE Cohort Study (20), and the Atherosclerosis Prevention in Pediatric Lupus Erythematosus (APPLE) trial (21). Missing observations in the data sets were imputed using multiple imputation methods and expectation-maximization algorithms in computation (22–24).

Each PP provided data about a patient at the time of a baseline visit and a followup visit. For each PP visit, the cSLE flare descriptors were provided (5) as follows: 1) MD-global, measured on a visual analog scale (VAS; 0 = inactive disease, 10 = very active disease); 2) parent assessment of patient overall well-being, measured on a VAS with a range from 0 to 10 (0 = very poor, 10 = very well); 3) proteinuria, measured by timed urine collection or spot PCR; 4) ESR; 5) levels of complement C3 and C4; and 6) item and summary scores of the SLEDAI-2K (7), or the domain and summary scores of the BILAG using the following numeric conversion: A = 12, B = 8, C = 1, and D/E = 0 (8). Information on complete blood counts and differential, serum chemistry, urinalysis, and anti-double-stranded DNA antibodies were also provided. Details on PP formats for the SLEDAI and BILAG are provided in Supplementary Appendices A and B, available on the *Arthritis Care & Research* web site at <http://online.library.wiley.com/doi/10.1002/acr.23557/abstract>.

PP raters were randomly assigned to assess the disease course of a maximum of 51 PPs. Response options offered were major flare, moderate flare, minor flare, unchanged, improved, or “I do not have enough information to make this assessment.” A global flare was considered as “present” whenever the disease course was rated as minor, moderate, or major flare.

Step B: adjudication of disease course of the PP. A randomization scheme was preplanned to ensure that each PP was sent to about 13 raters, with the ratio of American and international raters matching that of the PP raters’ pool (about 1:1). PPs with fewer than 4 ratings were regarded as “invalid” or “unqualified” and excluded from further consideration. Only “qualified” PPs with successful adjudication were considered in step 3.

Given that PP raters may not necessarily agree on the disease course, the “true” overall course of cSLE for a given PP was adjudicated using 2 approaches: 1) 67% rule: at least two-thirds of the raters agreed on a given disease course, and 2) majority-rule: the majority of the raters of a PP agreed on a given disease course. Other rules (50% rule and 75% rule) were also explored, and results were similar to the majority-rule and the 67% rule, respectively; therefore, they are not presented herein.

Step C: assessment of performance. *Statistical analysis in preparation of the testing of preliminary flare criteria.* Considering the intended widespread use of the cSLE flare criteria (14), we tested whether there were systematic differences in the ratings provided by raters from 1) different geographic regions or 2) with varying professional experience as measured by the duration of medical practice. Agreement among raters was assessed using intraclass correlation coefficients (ICCs) and/or kappa (κ) statistics. An ICC or a kappa value can be interpreted as follows: poor agreement: ICC or $\kappa < 0.4$, fair to good agreement: ICC or $\kappa \geq 0.4$ –0.75, and substantial to excellent agreement: ICC or $\kappa > 0.75$ (25).

Performance and accuracy. Each of the 4 flare algorithms (SLEDAI-based criteria, BILAG-based criteria, SLEDAI-CART, and BILAG-CART) was assessed for diagnostic accuracy using receiver operating characteristic (ROC) curve analysis. Specifically, the area under the ROC curve (AUC) was calculated, and the diagnostic accuracy was considered outstanding, excellent, good, fair, and poor if the AUC was in the range of 0.9–1.0, 0.81–0.90, 0.71–0.80, 0.61–0.70, and <0.60, respectively (14,26). In contrast to the flare criteria derived from multinomial regression models (SLEDAI- and BILAG-based criteria), CART-based flare algorithms (SLEDAI-CART and BILAG-CART) result in a single discrete value for sensitivity and specificity, respectively. Considering all possible flare scores, the overall diagnostic accuracy of an algorithm can be estimated.

Threshold score candidates for algorithms derived by multinomial logistic analysis. In the absence of strong guidance from the ACR, we used 2 statistical methods to define potential threshold scores. First, in an earlier phase of the project, consensus had been achieved that “flare score threshold” for a given algorithm should reflect the highest conditional AUC among all candidate thresholds on a ROC curve. Hence, these flare score thresholds represent the point on the ROC curves with the highest precision of correctly classifying the severity of a cSLE flare. Second, we also explored a distribution-weighted approach in which the flare score threshold was calculated based upon the average of means of scores in 2 neighboring flare states weighted by the SDs of the scores. The performance of the candidate thresholds from both statistical analyses described above was calculated, as well as average accuracies for the correct identification of minor, moderate, and major flares for the SLEDAI-based and BILAG-based algorithms.

Step D: ranking of candidate flare criteria and thresholds score. To support decision making, consensus conference participants reviewed a syllabus that provided the results of the preceding Delphi surveys, relevant published medical literature, and the results of the statistical analyses prior to the consensus conference (see step 3). Participants in the consensus conference were 13 experienced pediatric rheumatologists and nephrologists from South America, North America, Asia, and Europe with substantial clinical and research experience in cSLE (HIB, MWB, SPA, SA, CAS, FF, BG, SEW, DML, AR, RK, TA, and MKG).

A priori, the consensus level at the consensus conference was set at 75%, i.e., comparable or even somewhat higher than that chosen for similar studies in the past (15–18).

Using nominal group technique guided by an experienced moderator (BMF), the expert panel assessed each of the 4 top candidate flare algorithms (14) and potential flare score thresholds according to 1) feasibility (i.e., practicability: can the items be measured easily?), 2) reliability (i.e., reproducibility: can the items be measured precisely?), 3) redundancy (are there 2 or more items included in the candidate criteria measuring the same aspect of the disease?), 4) face validity (i.e., credibility: are the criteria sensible?), 5) content validity (i.e., comprehensiveness: do the criteria sample all of the domains of the disease?), 6) criterion validity: based on AUC, do the criteria accurately approximate the gold standard? (i.e., the adjudicated disease course as per the 67% rule or majority-rule), 7) sensitivity and specificity (do the criteria effectively identify patients with cSLE flares and distinguish them from patients who do not have a flare of their cSLE?), and 8) discriminant validity: do the criteria detect the smallest clinically important change? (i.e., discriminate patients with 1 of the following disease courses: minor flare, moderate flare, major flare, or no flare). Based on the above considerations, the consensus conference experts were asked to rank the candidate flare criteria from 1 (lowest criterion) to 4 (highest criterion).

The survey source data were batch processed, and open-source online survey software, LimeSurvey, was used for response management and as a presentation layer (see <http://www.limesurvey.org/>). All analyses were done using SAS, version 9.4, software and SYSTAT 12 software. *P* values less than 0.05 were considered statistically significant.

Table 1. Baseline characteristics of validation cohort*

	Majority rule (n = 1,860)	67% rule (n = 818)
Mean age, years	15.0	15.1
Female sex, %	81.7	82.5
Protein-creatinine ratio†	0.39	
≤0.2	63.8	67.5
>0.2	36.2	32.5
>0.5	14.5	13.0
>2.0	3.4	2.7
Organ involvement with active cSLE at baseline		
Neuropsychiatric	2.7	7.0
Musculoskeletal	12.4	8.67
Mucocutaneous	21.7	22.6
Hematologic	15.4	12.7
Renal	24.1	20.5
Cardiopulmonary	1.2	1.0
Constitutional symptoms	2.7	8.1

* Values are the percentage of the number, unless indicated otherwise. cSLE = childhood-onset systemic lupus erythematosus.
† Either from 24-hour urine or random urine sample (mg protein/mg urine creatinine).

Ethics review. The study was approved by the institutional review boards of the participating pediatric rheumatology centers. Informed consent was obtained from all parents and, as appropriate, participants assented prior to the study procedures.

Table 2. Change of descriptors in relationship to cSLE disease course*

Flare descriptor/rule	(1) Improved/ no change	(2) Minor flare	(3) Moderate flare	(4) Major flare	(1) vs. (2), adjusted <i>P</i>	(2) vs. (3), adjusted <i>P</i>	(3) vs. (4), adjusted <i>P</i>
ESR							
Majority	-0.02 ± 1.30	8.81 ± 1.34	22.80 ± 1.38	28.99 ± 1.68	< 0.0001	< 0.0001	0.023
67%	0.54 ± 1.58	7.28 ± 2.07	31.95 ± 2.39	35.34 ± 2.41	0.048	0.000	0.749
MD-global							
Majority	0.66 ± 0.50	3.05 ± 0.52	5.92 ± 0.53	7.95 ± 0.65	0.005	0.001	0.075
67%	0.76 ± 0.60	2.70 ± 0.79	7.74 ± 0.91	9.79 ± 0.92	0.210	< 0.0001	0.392
Protein-creatinine ratio							
Majority	0.02 ± 0.07	0.10 ± 0.07	0.66 ± 0.07	1.44 ± 0.08	0.843	< 0.0001	< 0.0001
67%	0.03 ± 0.07	0.02 ± 0.09	0.64 ± 0.11	1.61 ± 0.11	1.000	< 0.0001	< 0.0001
SLEDAI							
Majority	1.81 ± 0.26	4.58 ± 0.28	8.45 ± 0.29	16.00 ± 0.36	0.000	< 0.0001	< 0.0001
67%	1.56 ± 0.35	4.63 ± 0.48	9.98 ± 0.56	19.88 ± 0.55	0.000	< 0.0001	< 0.0001
BILAG							
Majority	3.12 ± 1.08	7.76 ± 0.93	15.19 ± 0.95	24.19 ± 1.15	0.007	< 0.0001	< 0.0001
67%	1.79 ± 1.34	8.63 ± 1.50	15.64 ± 1.61	28.71 ± 1.75	0.005	0.010	< 0.0001
SLEDAI-based flare algorithm							
Majority	-0.23 ± 0.17	1.66 ± 0.18	4.79 ± 0.19	9.88 ± 0.23	< 0.0001	< 0.0001	< 0.0001
67%	-0.34 ± 0.21	1.67 ± 0.29	5.84 ± 0.34	12.34 ± 0.34	< 0.0001	< 0.0001	< 0.0001
BILAG-based flare algorithm							
Majority	0.40 ± 0.56	3.00 ± 0.48	7.10 ± 0.49	11.96 ± 0.60	0.003	< 0.0001	< 0.0001
67%	-0.11 ± 0.66	3.49 ± 0.76	8.23 ± 0.79	15.05 ± 0.88	0.003	< 0.0001	< 0.0001

* Values presented are changes in means ± SDs, adjusted for multiple comparisons using the Tukey's method, unless indicated otherwise. cSLE = childhood-onset systemic lupus erythematosus; ESR = erythrocyte sedimentation rate; MD-global = physician global assessment of disease activity; SLEDAI = Systemic Lupus Erythematosus Disease Activity Index; BILAG = British Isles Lupus Assessment Group.

RESULTS

PP raters and validation data set (steps A and 2). A total of 2,996 ratings were provided to 503 pediatric rheumatologists and used for step 2. The response rate of the pediatric rheumatologists to the PP was 54% (274 of 503; locations: 30% from the US and Canada, 8% from Australia/Asia, 3% Africa/Middle East, 40% South and Central America, and 19% from Europe). The majority (69%) of PP raters had over 10 years of experience in treating cSLE. There were 1,860 PPs (1,860 of 2,996 [62%]) that were rated by at least 4 raters and therefore considered “qualified” for inclusion in step 3. There were no significant differences of distribution of flares between qualified and unqualified PPs ($P = 0.62$ by Fisher’s exact test).

When the majority rule was applied to the “qualified” PPs, there were 1,318 PPs representing global flares (510 minor flares, 483 moderate flares, and 325 major flares) and 542 unchanged/improved (29% of 1,860 PPs). When applying the 67% rule to the 1,860, only 818 PPs remained available for analysis, among them 484 representing a flare (194 minor flares, 146 moderate flares, and 144 major flares) and 334 PPs without cSLE flare. The patient characteristics reflected in these PPs are summarized in Table 1. PP raters from different geographic locations did not differ systematically in the disease course assignment for a given PP (North America versus other countries: ICC 0.658). Similarly, there was fair to good agreement among PP raters with different duration of medical experience (3–5, 6–10, 10–15, and >15 years) for the interpretation of the disease courses (ICC 0.656). Additionally, we explored other selection criteria (50% rule, 75% rule) and found no systematic differences with the 50% rule and 75% rule, resulting in similar

adjudication of the PP compared to the majority-rule and the 67% rule, respectively (data not shown).

Performance of preliminary algorithms of cSLE global flares (step C). The absolute baseline-to-followup changes of the parameters considered in the preliminary flare algorithms by flare severity and rule are provided in Table 2. Irrespective of the data set (67% rule, majority-rule), most of the cSLE flare descriptors included in the preliminary cSLE flare criteria (ESR, PCR, MD-global, SLEDAI, and BILAG) significantly changed between the baseline and followup visit, by flare severity. Notably, the accuracy of the SLEDAI-based algorithm was outstanding (AUC 0.93, 95% confidence interval [95% CI] 0.91–0.95), as was that of the BILAG-based algorithm (AUC 0.93, 95% CI 0.89–0.98). The CART-SLEDAI algorithm had an excellent accuracy for identifying patients with global flare of cSLE (any severity) (AUC 0.89, sensitivity 88.8%, and specificity 87.1%). The same was true for the CART-BILAG criteria (AUC 0.84, sensitivity 93.9%, and specificity 72.9%). Comparisons of accuracies in the development data set in 2010 (18) and this validation data set are summarized in Table 3.

Figure 2A and B depict potential thresholds for defining minor, moderate, and major flares. In this final consensus conference, again consensus (100%) was reached to use the statistically optimal threshold from logistic models to define all threshold scores for the both SLEDAI-based and the BILAG-based algorithms. As shown in Figure 3A and B, using these threshold cutoff scores allows for the discrimination of minor from moderate or severe flares, all with sensitivities and specificities of $\geq 82\%$. Neither of the CART-based algorithms was suited

Table 3. Comparison of the performance of the preliminary flare algorithm in the development and validation data set*

Algorithm		Flare category	AUC	
			2010 data	2017 data
SLEDAI-based flare score†	Score = $0.5 \times \text{SLEDAI} + 0.45 \times \text{PCR} + 0.5 \times \text{MD} + 0.02 \text{ ESR}$	Major flare	0.95	0.93
		At least moderate flare	0.85	0.94
		At least minor flare	0.86	0.93
BILAG-based flare score†	Score = $0.4 \times \text{BILAG} + 0.65 \times \text{PCR} + 0.5 \times \text{MD} + 0.02 \text{ ESR}$	Major flare	0.93	0.91
		At least moderate flare	0.85	0.92
		At least minor flare	0.85	0.93
SLEDAI-based CART rule	Score = 4 if $3 \leq \text{SLEDAI}$ Score = 3 if $0.7 \leq \text{PCR}$ and $3 > \text{SLEDAI}$ Score = 2 if $2 \leq \text{MD}$ and $0.7 > \text{PCR}$ and $3 > \text{SLEDAI}$ Score = 1 if otherwise	Major flare	0.85	0.76
		At least moderate flare	0.80	0.80
		At least minor flare	0.84	0.89
BILAG-based CART rule	Score = 4 if $2 \leq \text{BILAG}$ Score = 3 if $0.7 \leq \text{PCR}$ and $2 > \text{BILAG}$ Score = 2 if $2 \leq \text{MD}$ and $0.7 > \text{PCR}$ and $2 > \text{BILAG}$ Score = 1 if otherwise	Major flare	0.86	0.71
		At least moderate flare	0.80	0.75
		At least minor flare	0.82	0.84

* Values presented represent the area under the receiver operating characteristic curve (AUC) considering patient profile (PP) with consensus as defined by the 67% rule. Numeric values larger than or equal to the flare score signify a flare; higher scores are seen with more severe flare. See reference 14 for details about algorithm development. SLEDAI = Systemic Lupus Erythematosus Disease Activity Index; PCR = urine protein/creatinine ratio from random urine sample; MD = physician global assessment of disease measured on a visual analog scale (range 0–10, where 0 = inactive disease); ESR = erythrocyte sedimentation rate; BILAG = British Isles Lupus Assessment Group; CART = classification tree analysis.

† Algorithm considers for the change (baseline – followup) of each of the flare descriptors included.

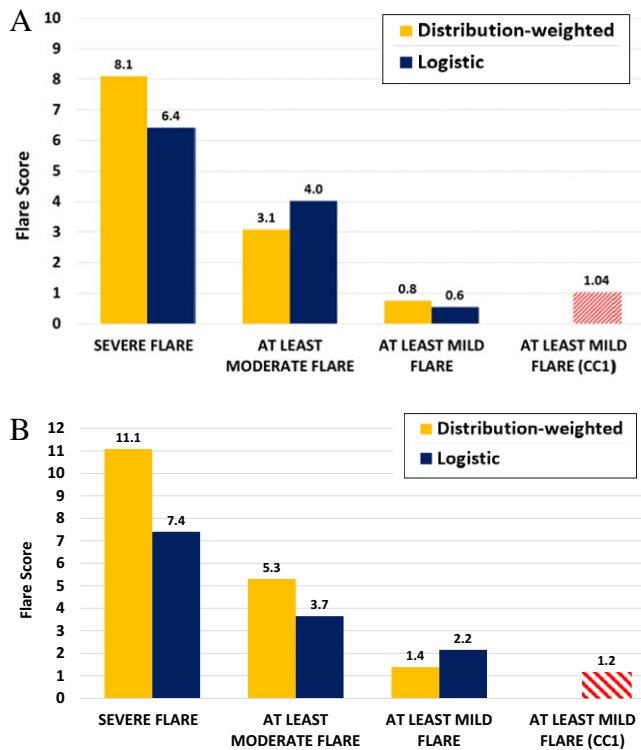


Figure 2. Potential flare thresholds to define childhood-onset systemic lupus erythematosus flare severity. **A**, Systemic Lupus Erythematosus Disease Activity Index–based algorithm, **B**, British Isles Lupus Assessment Group–based algorithm. Flare threshold values based on multinomial logistic regression models (blue bars) and distribution-weighted strategies (yellow bars) for each flare category (minor, moderate, major flare) were presented to the experts participating in the final consensus conference. There was 100% agreement to use threshold values derived from multinomial logistic regression, i.e., thresholds with the best statistical performance in receiver operating characteristic (ROC) curve analysis. Each threshold had the largest summation of sensitivity and specificity on the ROC curve. Red bars indicate the scores using each algorithm to assess the 2010 data (14). CC1 = Threshold value for flare as agreed upon in the initial consensus conference (see Step 4, Figure 1).

to discriminate between mild and moderate cSLE flares (Figure 3C and D).

Ranking of the preliminary cSLE flare algorithms (step D). Consensus conference participants achieved consensus that the BILAG-based (92%) and SLEDAI-based (100%) flare algorithms both have construct validity for measuring global flares of cSLE. There was consensus (100%) to recommend that both measures be collected in future cSLE clinical trials and that either one may be chosen as the primary end point. Consistent with their performance in the validation data set, no consensus was reached whether one of these 2 algorithms was preferable to the other. Consensus was achieved that CART-based algorithms are not suited for use in clinical trials, given that these algorithms cannot be used to discriminate minor from moderate cSLE flares. The results of this study were reviewed by the ACR Criteria Subcommittee and the ACR Quality of Care Committee.

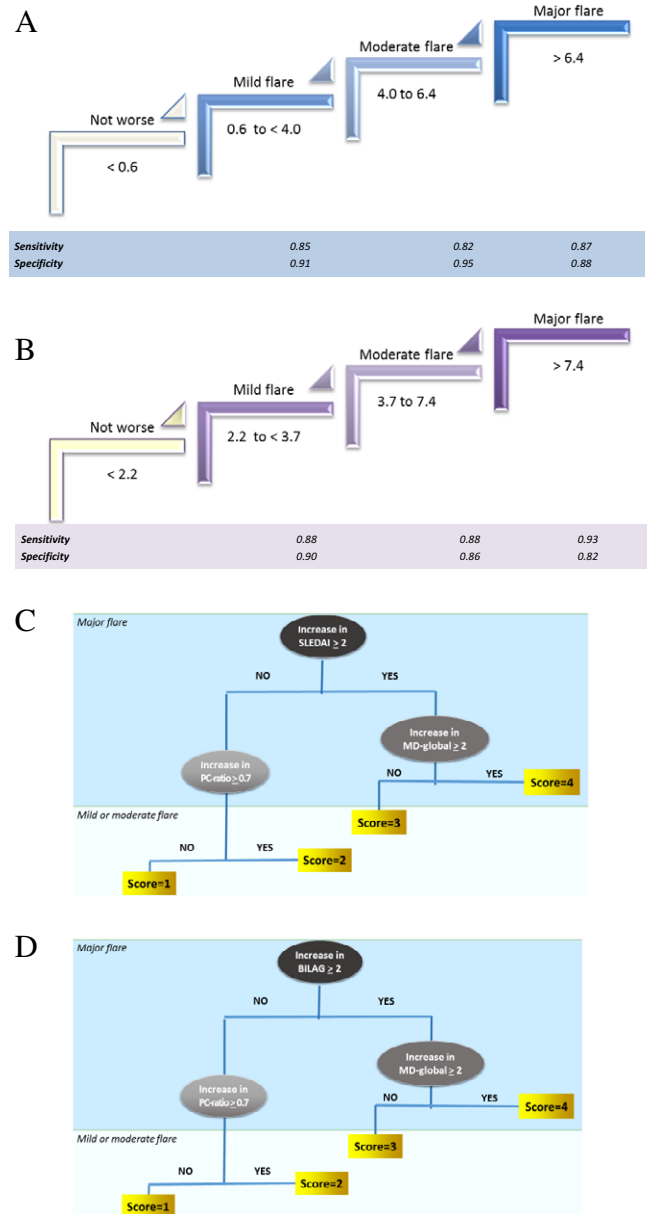


Figure 3. Flare score interpretation. Flare scores represent the cutoff score on the receiver operating characteristic curves that provide the best discrimination between adjacent disease states (no flare, minor or mild flare, moderate flare, major or severe flare) with childhood-onset systemic lupus erythematosus (cSLE). Sensitivities and specificities are shown for the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI)–based algorithm (**A**) and the British Isles Lupus Assessment Group (BILAG)–based algorithm (**B**). The SLEDAI–classification tree analysis (CART) algorithm (**C**) (where score = 4 if $3 \leq$ SLEDAI, score = 3 if $0.7 \leq$ protein/creatinine ratio [PC-ratio] and $3 >$ SLEDAI, score = 2 if $2 \leq$ physician global assessment [MD-global] and $0.7 >$ PC-ratio and $3 >$ SLEDAI, and score = 1 if otherwise) and in the BILAG-CART algorithm (**D**) (where score = 4 if $2 \leq$ BILAG, score = 3 if $0.7 \leq$ PC-ratio and $2 >$ BILAG, score = 2 if $2 \leq$ MD-global and $0.7 >$ PC-ratio and $2 >$ BILAG, and score = 1 if otherwise) are only able to distinguish major flares from other cSLE disease courses. Thus, the other 2 of the top preliminary flare criteria (SLEDAI-CART, BILAG-CART) were unable to discriminate minor from moderate cSLE flare.

DISCUSSION

The need to develop internationally agreed upon criteria for disease flares has become more urgent since the introduction of randomized withdrawal trials in pediatric rheumatology, in which time to flare or the proportion of patients who experience a flare are used as primary efficacy measures (27). We confirm the outstanding accuracy of the previously developed preliminary criteria of global flares of cSLE, based on large international data sets used for validation. Consensus has been achieved on how to interpret flare scores. The preferred cSLE global flare algorithms for use in clinical trials were derived from multinomial logistic regression models. These algorithms consider the differential and complementary contribution of select cSLE flare descriptors in identifying disease flares in this disease with highly variable multiorgan involvement. Despite consensus that CART-based algorithms are potentially of value when used in clinical care settings, there was agreement that they should not be used in clinical trials. As for SLE in adulthood, measures of the overall course are especially relevant because not all cSLE features improve or worsen in parallel. Current drugs used in cSLE therapy are not equally effective in reducing disease activity in the various organ systems. Thus, it is reasonable to assume that the same holds true for new or emerging drugs for cSLE. In clinical trials aimed at reducing cSLE-mediated inflammation in certain organ systems, it appears mandatory to ensure that global disease, i.e., disease manifestations in other than the target organ systems, is not worsening. The results of this study support that the SLEDAI-based and the BILAG-based flare scores are both highly suited to provide such information.

Based on the current evidence about these algorithms, they are similarly sensitive, specific, and accurate. Hence, consensus conference experts considered both algorithms equally valuable and suitable for use in clinical trials. Different from what is currently used to gauge response to therapy in juvenile idiopathic arthritis (11), flare algorithms derived from regression models allow for consideration of the differential importance of changes in individual cSLE flare descriptors when recognizing cSLE flares. The SLEDAI-based and BILAG-based flare scores are reminiscent of the Disease Activity Score (DAS) used in rheumatoid arthritis (28). However, the DAS considers the natural logarithm of the ESR and square roots of the number of swollen or tender joints, while the preliminary cSLE flare criteria require at most simple arithmetic maneuvers to calculate a cSLE flare score, supporting their ease of use (14).

All flare score algorithms consider changes in proteinuria, despite the inclusion of proteinuria assessment in the SLEDAI and BILAG scores. This allows for detection of renal SLE flares that occur in patients with existing proteinuria and also allows for the consideration of increases in proteinuria that would otherwise not be captured given the item definition used in the SLEDAI and BILAG, respectively. As reported previously, exclusion of changes in proteinuria from the flare algorithms resulted in inferior accuracy in predicting cSLE flares (14).

In line with our earlier studies (5,8), both cSLE flare criteria from CART and multinomial logistic regression analysis

showed excellent or even outstanding accuracy. Statistically, they were superior to algorithms that considered equally weighted percentage changes from a statistical point of view in the past.

Given the simplicity of CART-based criteria, they appear particularly suited for clinical settings, but a potential shortcoming of CART-based criteria includes so-called "over-fitting of the mathematical model," which can make them prone to less favorable statistical performance in subsequent validation studies (14). Mild cSLE flares often do not prompt clinicians to change therapy, whereas moderate cSLE flares generally require more intensive antiinflammatory therapy. Although CART-based flare algorithms were highly accurate for discriminating any kind of global flare when tested in this validation data set, they were unable to distinguish minor from moderate cSLE flares. This limitation prompted the agreement among the consensus conference experts to not recommend CART-based algorithms for use as outcome measures in clinical trials.

We chose 2 approaches to adjudicate the disease course (67% rule, majority rule) presented in the various PPs, which might have introduced bias. However, both approaches yielded comparable results.

The ACR has outlined a series of validation steps necessary before new criteria are to be widely used for clinical care or research (12). Among others, one step is to use data from clinical trials for developing response criteria. However, clinical trial data from interventions that impact cSLE activity are unavailable at present. In our study, the presence of a flare was based on the PP raters' perception of the course of cSLE instead. Given their prospective character and the expertise of the PP raters, we consider the quality of our data to be high, and the number of PPs per flare severity category yielded robust provisional cSLE flare criteria.

We would like to point out that PP raters from different parts of the world and different degrees of experience all showed excellent concordance (interrater agreement) in their assessment of the cSLE course. This supports the robustness of this validation study. A limitation might be that only 54% of those physicians approached to provide PP ratings provided feedback. Nonetheless, responses from 274 pediatric rheumatologists were obtained, which is a much larger number than for many similar validation exercises (9–11).

In addition to criteria for global flare and improvement, criteria for changes of cSLE in specific organ systems are likely needed. Depending on the proposed effect of a cSLE drug candidate, the Cutaneous Lupus Activity and Severity Index (29), pediatric lupus nephritis response measures (30), and standardized joint assessments for children (11) have already been validated to adequately capture the proposed therapeutic effects. To further provide support for the accuracy of the provisional criteria of global flare of cSLE, data from clinical trials will be needed.

Taken together, a methodologically stringent validation process has been employed to calculate a flare score that can be used to interpret the course of cSLE over time with respect to the degree of worsening that might have occurred. Based on the data available, these algorithms cannot be used to quantify potential improvement over time.

ACKNOWLEDGMENTS

We thank Kasha Wiley (overall study coordination), Susan Priest (consensus conference logistics), Pinar Avar (consensus conference support and data management), and Carly Muller, Malea Rolfsen, Allen Watts, Gaurav Gulati, and Jamie Meyers-Eaton (patient profile testing) from Cincinnati Children's Hospital Medical Center (CCHMC), as well as CCHMC Biomedical Informatics (web-based data management application development). We also thank Drs. Laura Schanberg and Christy Sandberg and the Childhood Arthritis and Rheumatology Research Alliance for provision of the data from the Atherosclerosis Prevention in Pediatric Lupus Erythematosus clinical trial, as well as the UK Juvenile-Onset SLE (JSLE) Study Group, for provision of the data from the UK JSLE Cohort Study. We are indebted to the members of the External Scientific Advisory Committee of this study for their advice in the study implementation, conduction, and its statistical analysis: Drs. Tuhina Neogi, Ian Bruce, David Isenberg, Nicola Ruperto, and James Witter. For a list of physicians who made important contributions to this work by providing their expertise when rating the patient profiles, see Supplementary Appendix C, available on the *Arthritis Care & Research* web site at <http://onlinelibrary.wiley.com/doi/10.1002/acr.23557/abstract>.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be submitted for publication. Dr. Brunner had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Brunner, Holland, Beresford, Ardoin, Appenzeller, Silva, Flores, Goilav, Wenderfer, Levy, Ravelli, Khunchandani, Avcin, Klein-Gitelman, Feldman, Ruperto, Ying.
Acquisition of data. Brunner, Holland, Beresford, Ardoin, Appenzeller, Silva, Flores, Goilav, Wenderfer, Levy, Ravelli, Khunchandani, Avcin, Klein-Gitelman, Feldman, Ruperto, Ying.
Analysis and interpretation of data. Brunner, Holland, Beresford, Ardoin, Appenzeller, Silva, Flores, Goilav, Wenderfer, Levy, Ravelli, Khunchandani, Avcin, Klein-Gitelman, Feldman, Ruperto, Ying.

REFERENCES

- Silva CA, Avcin T, Brunner HI. Taxonomy for systemic lupus erythematosus with onset before adulthood. *Arthritis Care Res (Hoboken)* 2012;64:1787–93.
- Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* 1997;40:1725.
- Brunner HI, Gladman DD, Ibañez D, Urowitz MD, Silverman ED. Difference in disease features between childhood-onset and adult-onset systemic lupus erythematosus. *Arthritis Rheum* 2008;58:556–62.
- Hiraki LT, Benseler SM, Tyrrell PN, Hebert D, Harvey E, Silverman ED. Clinical and laboratory characteristics and long-term outcome of pediatric systemic lupus erythematosus: a longitudinal study. *J Pediatr* 2008;152:550–6.
- Brunner HI, Klein-Gitelman MS, Higgins GC, Lapidus SK, Levy DM, Eberhard A, et al. Toward the development of criteria for global flares in juvenile systemic lupus erythematosus. *Arthritis Care Res (Hoboken)* 2010;62:811–20.
- 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.
- Gladman DD, Ibañez D, Urowitz MB. Systemic lupus erythematosus disease activity index 2000. *J Rheumatol* 2002;29:288–91.
- Isenberg DA, Rahman A, Allen E, Farewell V, Akil M, Bruce IN, et al. BILAG 2004: development and initial validation of an updated version of the British Isles Lupus Assessment Group's disease activity index for patients with systemic lupus erythematosus. *Rheumatology (Oxford)* 2005;44:902–6.
- Wallace CA, Ravelli A, Huang B, Giannini EH. Preliminary validation of clinical remission criteria using the OMERACT filter for select categories of juvenile idiopathic arthritis. *J Rheumatol* 2006;33:789–95.
- Ruperto N, Ravelli A, Oliveira S, Alessio M, Mihaylova D, Pasic S, et al. 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. *Arthritis Rheum* 2006;55:355–63.
- 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.
- Singh JA, Solomon DH, Dougados M, Felson D, Hawker G, Katz P, et al. Development of classification and response criteria for rheumatic diseases. *Arthritis Rheum* 2006;55:348–52.
- Brunner H, Baker A, Cedeno A, Huggins JL, Sagcal-Gironella AC, Ying J, et al. The pediatric automated neuropsychological assessment metrics has reproducibility and criterion validity in childhood-onset lupus [abstract]. *Arthritis Rheum* 2011;63:S780.
- Brunner HI, Mina R, Pilkington C, Beresford MW, Reiff A, Levy DM, et al. Preliminary criteria for global flares in childhood-onset systemic lupus erythematosus. *Arthritis Care Res (Hoboken)* 2011;63:1213–23.
- Brunner HI, Higgins GC, Klein-Gitelman MS, Lapidus SK, Olson JC, Onel K, et al. Minimal clinically important differences of disease activity indices in childhood-onset systemic lupus erythematosus. *Arthritis Care Res (Hoboken)* 2010;62:950–9.
- Boers M, Kirwan JR, Gossec L, Conaghan PG, D'Agostino MA, Bingham CO III, et al. How to choose core outcome measurement sets for clinical trials: OMERACT 11 approves filter 2.0. *J Rheumatol* 2014;41:1025–30.
- Lassere MN. A users guide to measurement in medicine. *Osteoarthritis Cartilage* 2006;14 Suppl A:10–3.
- Mina R, Harris JG, Klein-Gitelman MS, Appenzeller S, Centeville M, Eskra D, et al. Initial benchmarking of the quality of medical care in childhood-onset systemic lupus erythematosus. *Arthritis Care Res (Hoboken)* 2016;68:179–86.
- Ruperto N, Ravelli A, Pistorio A, Ferriani V, Calvo I, Ganser G, et al. The provisional Paediatric Rheumatology International Trials Organisation/American College of Rheumatology/European League Against Rheumatism Disease activity core set for the evaluation of response to therapy in juvenile dermatomyositis: a prospective validation study. *Arthritis Rheum* 2008;59:4–13.
- Watson L, Leone V, Pilkington C, Tullus K, Rangaraj S, McDonagh JE, et al. Disease activity, severity, and damage in the UK Juvenile-Onset Systemic Lupus Erythematosus Cohort. *Arthritis Rheum* 2012;64:2356–65.
- Schanberg L, Sandborg C, Barnhart H, Ardoin S, Yow E, Evans G, et al. Use of atorvastatin in systemic lupus erythematosus in children and adolescents. *Arthritis Rheum* 2012;64:285–96.
- Little RJ. Pattern-mixture models for multivariate incomplete data. *J Am Stat Assoc* 1993;88:125–34.
- Schafer JL. Multiple imputation: a primer. *Stat Meth Med Res* 1999;8:3–15.
- Rubin DB. Multiple imputation for nonresponse in surveys. Hoboken (NJ): Wiley-Interscience; 2004.

25. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420–8.
26. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.
27. Caldwell JR, Furst DE, Smith AL, Clark JA, Bonebrake RA, Gruhn WB, et al. Flare during drug withdrawal as a method to support efficacy in rheumatoid arthritis: amiprilose hydrochloride as an example in a double blind, randomized study. *J Rheumatol* 1998;25:30–5.
28. Balsa A, Carmona L, González-Alvaro I, Belmonte MA, Tena X, Sanmartí R. Value of Disease Activity Score 28 (DAS28) and DAS28-3 compared to American College of Rheumatology-defined remission in rheumatoid arthritis. *J Rheumatol* 2004;31:40–6.
29. Albrecht J, Taylor L, Berlin JA, Dulay S, Ang G, Fakharzadeh S, et al. The CLASI (Cutaneous Lupus Erythematosus Disease Area and Severity Index): an outcome instrument for cutaneous lupus erythematosus. *J Invest Dermatol* 2005;125:889–94.
30. Mina R, von Scheven E, Ardoin SP, Eberhard BA, Punaro M, Ilowite N, et al. Consensus treatment plans for induction therapy of newly diagnosed proliferative lupus nephritis in juvenile systemic lupus erythematosus. *Arthritis Care Res (Hoboken)* 2012;64:375–83.