

Preliminary Definition of Disease Flare in Juvenile Rheumatoid Arthritis

HERMINE I. BRUNNER, DANIEL J. LOVELL, BARBARA K. FINCK, and EDWARD H. GIANNINI

ABSTRACT. Objective. To develop preliminary criteria for defining disease flare in patients with polyarticular-course juvenile rheumatoid arthritis (JRA).

Methods. Data from a randomized clinical trial of etanercept in JRA (51 patients) and the 6 core response variables (CRV) for JRA were used to derive flare definitions. The criterion standard of flare was treatment with placebo. Candidate flare definitions were assessed by receiver-operator characteristic (ROC) curve properties and other statistics for diagnostic tests.

Results. Of the possible flare definitions tested with acceptable statistical properties, the one that seemed to be the most useful was worsening in any 2/6 CRV by $\geq 40\%$ without improvement in more than 1 of the remaining CRV by $\geq 30\%$. Two other superior flare definitions were (1) worsening in 3/6 CRV by $\geq 30\%$ and (2) any worsening of the Childhood Health Assessment Questionnaire, worsening of erythrocyte sedimentation rate by $\geq 30\%$ and worsening of the active joint count by $\geq 10\%$.

Conclusions. CRV are useful for defining flare in JRA. Worsening in any 2/6 CRV by $\geq 40\%$ without concomitant improvement of more than one of the remaining CRV by $\geq 30\%$ appears to be the most suitable preliminary flare definition. Because the proposed flare criteria were derived from a small number of patients, it is essential to perform more definitive testing of this and several alternative flare definitions in larger patient populations. (J Rheumatol 2002;29:1058–64)

Key Indexing Terms:

JUVENILE RHEUMATOID ARTHRITIS
DISEASE ACTIVITY

FLARE
CHILDREN

Juvenile rheumatoid arthritis (JRA) is a group of inflammatory autoimmune conditions of childhood¹. Three subgroups are recognized according to the number of joints involved during the first 6 months of disease and the presence of systemic features such as rash, serositis, and fevers. The course of JRA is characterized by changes in the degree of inflammation. To describe clinically important decrease in disease activity, preliminary criteria for improvement in JRA² have been developed, which are based on the 6 core response variables (CRV) of JRA. These preliminary criteria for improvement of JRA are now often used in randomized clinical trials (RCT) to measure the effects of different treatment interventions for children with JRA.

Acute worsening of symptoms is a well known compli-

cation of JRA. Exacerbation of disease activity is referred to as disease flare by most researchers and clinicians, although the exact definition of it differs from one author to another, making it difficult to compare the results of studies. There is no universally accepted definition of disease flare that has been tested in a large number of patients. Such a standard definition would be useful to increase the comparability of studies and is required for certain experimental study designs (withdrawal designs) such as that used for the evaluation of the efficacy of etanercept in JRA³.

The objective of the current study was to develop preliminary criteria for defining disease flare in patients with polyarticular-course JRA by using the CRV for JRA and the data of the RCT of etanercept in polyarticular-course JRA.

MATERIALS AND METHODS

Patients. Data from the RCT of etanercept for the treatment of children with polyarticular-course JRA³ were used. Information on all 51 patients was available, all of whom had shown clinical improvement² taking etanercept by day 90 post initiation of open treatment with the drug. Patients with clinically significant improvement taking etanercept were subsequently randomized to receive treatment in a blinded fashion with either etanercept or placebo until a flare of disease occurred. The definition of disease flare chosen for the purpose of this RCT was as follows: worsening of at least 3 of the 6 CRV by at least 30% without concomitant improvement of more than one of the remaining CRV by 30% or more. A patient with flare also had to have a minimum of 2 active joints. In addition, if either physician global assessment of disease severity or the parent rating of the overall well being of the patient was used to define a patient with

From the William S. Rowe Division of Rheumatology, Cincinnati Children's Hospital Medical Center, and the Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH; and Immunex Corporation, Seattle, WA, USA.

Supported by Immunex Corporation.

H.I. Brunner, MD, MSc; D.J. Lovell, MD, MPH; E.H. Giannini, MSc, DrPH, William S. Rowe Division of Rheumatology, Cincinnati Children's Hospital Medical Center University of Cincinnati College of Medicine, Department of Pediatrics; B.K. Finck, MD, formerly at Immunex Corporation.

Address reprint requests to Dr. H.I. Brunner, William S. Rowe Division of Rheumatology, Cincinnati Children's Hospital Medical Center PAV 2-129, 3333 Burnet Avenue, Cincinnati, OH, 45229, USA.

E-mail: brus9z@chmcc.org

Submitted June 27, 2001; revision accepted August 24, 2001.

flare, then either one of these CRV had to have deteriorated by at least 2 units (categorical scale 0–10) to be considered as worsened.

Core set of variables in JRA. The core set of outcome variables consists of 6 CRV⁴. They are the number of joints with active arthritis (AJC), the number of joints with limited range of motion (LROM), the physician global assessment of disease severity, the parent or patient global assessment of overall well being, one laboratory marker of inflammation, and the score of a functional assessment tool. For this study the erythrocyte sedimentation rate (ESR) was used as a laboratory marker of inflammation, and the Childhood Health Assessment Questionnaire (CHAQ)⁵ score was chosen to assess function.

Definition of joints with active arthritis and with limited range of motion. All joints with swelling or with any 2 other signs of inflammation (heat, limited range of motion, tenderness or painful range of motion) were classified as joints with active arthritis (AJC), whereas involved joints with only limited range of motion are referred to as LROM.

Childhood Health Assessment Questionnaire (CHAQ). This tool measures physical function of children of all ages with chronic musculoskeletal disorders⁵. The 53 items of the CHAQ are grouped into 8 domains. A summary score between 0 and 3 can be calculated. Changes of CHAQ scores occur in minimal increments of 0.125, and a score of 0 is given to patients without limitation of physical function as defined by the CHAQ.

Generation of candidate flare definitions. The baseline for assessing change in disease was the day the patient either started placebo or continued etanercept (Day 90 of initial trial). This reference time point (RTP) for subsequent disease flares is sensible, because all patients who underwent randomization ($n = 51$) had shown clinically important improvement taking etanercept, and had failed therapy with other second-line agents. Treatment with placebo was used as the criterion standard for disease flare. Thus all patients receiving placebo were expected to experience a disease flare, and all patients randomized to continuing treatment with etanercept were thought to have no flare.

Candidate definitions for disease flare were generated based on 20 to 50% changes of 2 to 4 CRV. Worsening of less than 20% of the CRV was deemed less likely to be clinically relevant. Moreover, candidate flare definitions that were only based on the physician global assessment of disease severity and/or the parent global assessment of overall well being were tested. To avoid spurious effects, only worsening of at least 2 units on the ordinal scale of physician or parent global assessments was considered clinically relevant (both measured on a categorical scale with range 0–10). In addition, patients with flare had to have at least 2 active joints.

Generalized estimation equation (GEE) models⁶ were used to identify additional possible definitions of flare. GEE modeling takes into account that the data set contained multiple observations per patient (longitudinal data set). Therefore the multiple observations recorded from the same patient are not independent of each other as would be required for standard statistical procedures. GEE also take into account that patients differed in the duration of followup and that they were assessed at differing time intervals. Thus, GEE models can incorporate the fact that more information may be obtained from patients with longer followup time than from those with shorter observation periods. The fit of the GEE models was assessed by Pearson chi-square, and the ratio of deviance/degree of freedom (scaled deviance) was determined as a measure of dispersion. The results of the GEE with regard to the best predictors and their relative impact on explaining disease flare were then used to generate additional candidate flare definitions.

Assessment of candidate flare definitions. The measurement characteristics of the generated flare definitions were assessed by calculating the sensitivity, specificity, positive predictive value, and negative predictive value to detect flare. For the purpose of this analysis, sensitivity of candidate flare definitions was defined by the ratio of the number of patients with flare receiving placebo (numerator) and the number of all patients receiving placebo (denominator; $n = 26$). Similarly, specificity of a candidate flare definition was measured by using the number of patients without flare

taking etanercept (numerator) divided by the number of all patients treated with etanercept (denominator; $n = 25$). The positive predictive value of a candidate flare definition was estimated by the ratio of the number of patients with flare receiving placebo (numerator) and the number of all patients with flare (denominator). Accordingly, the negative predictive value corresponds to the ratio of the number of patients without flare taking etanercept (numerator) divided by the number of all patients without flare (denominator). For all possible flare definitions, the median time to flare, when treated with etanercept or placebo, was calculated by using the log-rank test. Flare definitions that led to early identification of patients receiving placebo were deemed superior to those with a longer time lag. This is clinically sensible because early detection of flare will minimize patient discomfort and may decrease the risk of permanent disease damage and complications associated with increased disease activity in patients with JRA. A receiver operator characteristic (ROC) curve⁷ was constructed. ROC curves are graphs of sensitivity on the y-axis and $(1 - \text{specificity})$ on the x-axis. The best candidate definitions of flare are located in the left upper quadrant of the graph. To more fully characterize the ability of candidate flare definitions to identify flares, we also used the ROC curve approach proposed by Deyo, *et al*⁸. The results of ROC analysis were summarized by the value of the area under the curve (AUC) (range: 0–1), which is calculated by multiplying the value of the sensitivity of a candidate flare definition with the value of the definition's specificity.

Confidence intervals of the AUC were calculated as suggested for ROC curves derived from the same patients⁹. Any flare definition with an AUC > 0.5 is responsive to change. The greater the AUC, the better the overall performance of a certain candidate flare definition. As previously suggested for ROC, candidate flare definitions were also compared based on the likelihood ratio for a positive test (LR)^{10–12}. The LR is defined as the ratio of the probability in a patient treated with placebo of having a flare (numerator) to the corresponding probability of flare in a patient treated with etanercept (denominator). The LR corresponds to $\text{sensitivity} \div (1 - \text{specificity})$. The greater the LR of a candidate flare definition, the greater the chance of a patient with flare to be identified as having a flare. Ninety-five percent confidence intervals of AUC and LR estimates were calculated and compared for important differences as previously suggested⁹.

Selection of the best candidate flare definitions. From the pool of generated candidate flare definitions, the best flare definitions were selected according to (in sequence of importance) (a) the AUC, (b) the LR, (c) the position on the ROC curve, and (d) the PPV and NPV. Besides statistical performance, the clinical usefulness of the flare definitions was also considered. For instance, in order to avoid failures to recognize flares, we deemed less pronounced worsening of CRV to be more clinically relevant than more extreme changes in CRV. Also, clinically relevant flare definitions needed to be easy to use in daily practice. Thus candidate flare definitions considering fewer CRV seemed to be preferable over those requiring the assessment of the complete set of 6 CRV. Moreover, the assessment of both relevant improvement and important worsening of CRV for flare definitions seemed more difficult compared to determining only important deterioration in CRV.

RESULTS

Demographics. Twenty-six patients received placebo, and 25 were treated with etanercept. Patients were followed until disease flare occurred (as defined for the purpose of the RCT of etanercept therapy for JRA³). The median time to flare taking placebo was 30 days (range 6–126). The median time to flare taking etanercept (range 31–131 days) was not reached at the end of the RCT, because more than half of the 25 patients randomized to etanercept therapy did not fulfill criteria of flare (as defined for the purpose of this RCT) at the end of the study period. No

severe adverse drug reactions were observed. None of the patients was lost to followup or excluded based on protocol violations.

Core response variables (CRV) at reference time point (RTP) of flare definitions (day 90 of RCT). The parent global assessment of overall well being ranged between 0 and 7 (mean 1.88, median 1), while the physician global assessment ranged between 0 and 8 at baseline (mean 2.33, median 1). The average AJC at RTP was 11 (median 9, range 0–29), and the number of joints with LROM was between 0 and 53 (mean 18, median 15). The ESR ranged from 1 to 98 mm/h (mean 21 mm/h, median 11 mm/h). The CHAQ scores of the patients ranged between 0 and 3 (mean 0.825, median 1). There were no statistically significant differences with respect to the CRV at the RTP between patients randomized to placebo and those randomized to continue etanercept therapy.

Flare definitions based on 6 CRV. Definitions of disease flare were tested based on 20–50% changes in different numbers of the CRV. For some of the candidate flare definitions, concomitant improvement of no more than one of the remaining CRV by 30% or more was allowed (Table 1), while for other definitions of flare possible improvement of CRV was not considered (Table 2). Generally, candidate flare definitions based on changes of at least 40% in the CRV were very similar to those based on minimal worsening of 50% in the CRV (Tables 1 and 2) with regard to their sensitivity, specificity, PPV, or NPV. However, flare definitions considering worsening of CRV by at least 50%

tended to have a longer lag time (median time to flare) to identify patients with flare.

Flare definitions derived from less than 6 CRV. Flare definitions that included only the global assessments of disease severity by the physicians and/or the global assessment of well being by the parents (Table 3) did not perform better than those based on the complete set of outcome variables (6 CRV) (Tables 1 and 2). The GEE model using any worsening of the CHAQ, deterioration of the ESR by at least 30%, and concomitant increase in the AJC by at least 10% performed best with a scaled deviance of 1.04 (desired value: 1.0) and a Pearson chi square (df = 36) of 0.95 (desired value: 1.0). In this GEE model, ESR, AJC, and CHAQ scores were all highly significant predictors of disease flare, with chi-square values of at least 5.7 (df = 1), and the associated p values were as follows: $p < 0.01$ (ESR), $p < 0.004$ (CHAQ), and $p < 0.002$ (AJC). GEE models using other CRV or other percentage changes of ESR, CHAQ, and AJC had inferior performance (Table 3). All candidate flare definitions derived from GEE modeling had, on average, higher PPV and specificity compared to other candidate flare definitions, but lower NPV and sensitivity.

Flare definitions based on absolute changes of 6 core response variables. Exploratory analysis was done to examine absolute differences of CRV, rather than differences based on percentage changes of CRV, for their usefulness to define disease flare in JRA. However, candidate flare definitions based on absolute changes of CRV were inferior to those considering percentage changes of 6 CRV (data not shown).

Table 1. Possible definitions of flare for JRA based on worsening of CRV that allow no more than one of the remaining CRV to improve by 30% or more.

Definitions of Flare (Worsening of CRV)	Number of Patients with Flare		Sensitivity ^a	Specificity ^b	PPV ^c	NPV ^d	Chi-square p value	Median Time to Flare (days) ^e	
	On Etanercept	On Placebo						On Etanercept	On Placebo
Change in at least 2 CRV									
20%	11	23	0.88	0.56	0.68	0.82	< 0.001	122	28
30%	8	22	0.85	0.68	0.73	0.81	< 0.001	> 131	29
40%	5	22	0.85	0.80	0.81	0.83	< 0.001	> 131	29
50%	5	22	0.85	0.80	0.81	0.83	< 0.001	> 131	30
Change in at least 3 CRV									
20%	7	20	0.77	0.72	0.74	0.75	< 0.001	> 131	29
30%	6	20	0.77	0.76	0.77	0.76	< 0.001	> 131	30
40%	5	18	0.69	0.80	0.78	0.71	< 0.001	> 131	31
50%	5	18	0.69	0.80	0.78	0.71	< 0.001	> 131	32
Change in at least 4 CRV									
20%	5	18	0.69	0.80	0.78	0.71	0.001	> 131	32
30%	5	16	0.62	0.80	0.76	0.67	0.003	> 131	34
40%	5	16	0.62	0.80	0.76	0.67	0.003	> 131	34
50%	5	15	0.58	0.80	0.75	0.65	0.006	> 131	77

^a Sensitivity: number of patients with flare receiving placebo/number of patients receiving placebo (n = 26); ^b Specificity: Number of patients without flare on etanercept/number of patients treated on etanercept (n = 25); ^c Positive predictive value: number of patients with flare receiving placebo/number of all patients with flare; ^d Negative predictive value: number of patients without flare on etanercept/number of all patients without flare; ^e For all flare definitions, the time to flare was statistically longer ($p < 0.0004$ using log-rank test) for patients receiving placebo compared to patients treated with etanercept. Patients with polyarticular-course JRA were treated with placebo (n = 26) or with etanercept (n = 25). Treatment with placebo was used as criterion standard for flare.

Table 2. Possible definitions of flare allowing any improvement in the remaining CRV for JRA.

Definitions of Flare (Worsening of CRV)	Number of Patients with Flare		Sensitivity ^a	Specificity ^b	PPV ^c	NPV ^d	Chi-square p value	Median Time to Flare (days) ^e	
	On Etanercept	On Placebo						On Etanercept	On Placebo
Change in at least 2 CRV									
20%	14	24	0.92	0.44	0.63	0.85	< 0.003	>131	28
30%	11	23	0.88	0.56	0.68	0.82	< 0.001	>131	29
40%	7	23	0.88	0.72	0.77	0.86	< 0.001	>131	29
50%	7	23	0.88	0.72	0.77	0.86	< 0.001	>131	30
Change in at least 3 CRV									
20%	7	21	0.81	0.72	0.75	0.78	< 0.001	>131	29
30%	6	21	0.81	0.76	0.78	0.79	< 0.001	>131	30
40%	5	19	0.73	0.80	0.79	0.74	< 0.001	>131	31
50%	5	19	0.73	0.80	0.79	0.74	< 0.001	>131	32
Change in at least 4 CRV									
20%	5	18	0.69	0.80	0.78	0.71	< 0.001	>131	32
30%	5	16	0.62	0.80	0.76	0.67	< 0.003	>131	34
40%	5	16	0.62	0.80	0.76	0.67	< 0.003	>131	34
50%	5	16	0.62	0.80	0.76	0.67	< 0.006	>131	77

^a Sensitivity: number of patients with flare receiving placebo/number of patients receiving placebo (n = 26);

^b Specificity: number of patients without flare on etanercept/number of patients treated on etanercept (n = 25);

^c Positive predictive value: number of patients with flare receiving placebo/number of all patients with flare;

^d Negative predictive value: number of patients without flare on etanercept/number of all patients without flare;

^e For all flare definitions, the time to flare was statistically shorter (p < 0.0004 using log-rank test) for patients receiving placebo compared to patients treated with etanercept. Patients with polyarticular-course JRA were treated either with placebo (n = 26) or with etanercept (n = 25).

Table 3. Possible definitions of disease flare using a subset of the 6 CRV of JRA.

Definitions of Flare (Worsening of CRV)	Number of Patients with Flare		Sensitivity ^a	Specificity ^b	PPV ^c	NPV ^d	Chi-square p value	Median Time to Flare (days) ^e	
	On Etanercept	On Placebo						On Etanercept	On Placebo
Worsening of physician global assessment by ≥ 2 units	5	17	0.65	0.80	0.77	0.69	0.001	> 131	32
Worsening of parent global assessment by ≥ 2 units	7	19	0.73	0.72	0.73	0.72	0.001	> 131	32
Worsening of physician's and parent's global assessment by ≥ 2 units	9	19	0.73	0.64	0.67	0.70	0.008	131	30
Any worsening of CHAQ and ≥ 30% increase of ESR and ≥ 10% increase of AJC ^f	3	17	0.65	0.88	0.85	0.71	0.001	> 131	32
Any worsening of CHAQ and ≥ 40% increase of ESR and ≥ 20% increase of AJC ^f	2	15	0.58	0.92	0.88	0.67	0.001	> 131	31
Any worsening of CHAQ and ≥ 30% increase of ESR and ≥ 30% increase of AJC ^f	2	15	0.58	0.92	0.88	0.67	0.001	> 131	32

^a Sensitivity: number of patients with flare receiving placebo/number of patients receiving placebo (n = 26);

^b Specificity: number of patients without flare on etanercept/number of patients treated on etanercept (n = 25);

^c Positive predictive value: number of patients with flare receiving placebo/number of all patients with flare;

^d Negative predictive value: number of patients without flare on etanercept/number of all patients without flare;

^e For all flare definitions, the time to flare was statistically shorter (p < 0.0004 using log-rank test) for patients receiving placebo compared to patients treated with etanercept. Patients with polyarticular JRA were treated either with placebo (n = 26) or with etanercept (n = 25). Treatment with placebo was used as criterion standard for flare. The worsening of the physician's global assessment of disease severity and the parent's global assessment of overall well-being was measured on categorical scales (range: 0–10).

^f Definition derived from generalized estimation equations (GEE).

Selection of the best candidate flare definitions. Irrespective of the flare definition chosen, patients treated with placebo flared significantly earlier than patients taking etanercept (all p values by log rank test < 0.0004). The AUC of the generated candidate flare definitions ranged between 0.406 and 0.677 (25th percentile at 0.494; 50th percentile at 0.550; 75th percentile at 0.585) and the LR ranged between 1.65 and 7.25 (25th percentile at 2.98, 50th percentile at 3.16, 75th percentile at 3.56). All candidate flare definitions whose AUC or whose LR was higher than the 50th percentile of the generated candidate flare definitions are shown in Table 4. They are all responsive measures of disease flares (AUC > 0.5).

Definition 1. Based on the largest AUC, a high LR (Table 4), and the best position on the ROC graph (Figure 1), definition 1 appeared to be the best candidate flare definition. All other candidate flare definitions based on the change of 2 CRV are clinically less relevant, because they either depend on more pronounced changes (50%) of the CRV or had a lower LR compared to the other disease flare definitions

summarized in Table 2.

Definition 2 also seemed to be a very good candidate definition given its large AUC and concomitantly high LR. With the inclusion of more than 2 CRV in the definition of flare, the importance of assessing concomitant improvement of the remaining CRV seems less important. This is supported by the finding that the candidate flare definition, which is based on 30% worsening of the 3 CRV and also considers the absence of improvement of more than one of the remaining CRV (i.e., the definition of flare used in the RCT of etanercept in JRA), does not perform better than definition 2. *Definition 3* has a high AUC and a higher LR than definition 1 or 2. Definition 3 appeared to be a good candidate flare definition, because it requires the assessment of 3 instead of all 6 CRV, to identify patients with disease. Among GEE models, definition 3 was the best flare definition. The use of GEE modeling allows one to assess changes in the CRV and also considers that patients were not followed the same length of time. Definition 3 does not appear to have as good a position on the ROC curve as some

Table 4. The superior candidate definitions of disease flare derived from CRV.

Candidate Flare Definitions	Definition	AUC (95% CI) ^a	LR (95% CI) ^b	Sensitivity ^c	Specificity ^d	PPV ^e	NPV ^f
Worsening of 2 CRV by at least 40% without concomitant improvement of more than 1 variable by ≥ 30%	1*	0.677 (0.571–0.783)	4.23 (1.93–6.54)	0.85	0.80	0.81	0.83
Worsening of 2 CRV by at least 50% without concomitant improvement of more than 1 variable by ≥ 30%		0.677 (0.571–0.783)	4.23 (1.93–6.54)	0.85	0.80	0.81	0.83
Worsening of 2 CRV by ≥ 40%		0.637 (0.530–0.803)	3.16 (0.98–5.33)	0.88	0.72	0.77	0.86
Worsening of 2 CRV by ≥ 50%		0.637 (0.530–0.803)	3.16 (0.98–5.33)	0.88	0.72	0.77	0.86
Worsening of 3 CRV by ≥ 30%	2*	0.614 (0.505–0.723)	3.37 (1.14–5.59)	0.81	0.76	0.78	0.79
Worsening of 3 CRV by ≥ 40%		0.585 (0.466–0.703)	3.65 (1.35–5.96)	0.73	0.80	0.79	0.74
Worsening of 3 CRV by ≥ 50%		0.585 (0.466–0.703)	3.65 (1.35–5.96)	0.73	0.80	0.79	0.74
Worsening of 3 CRV by ≥ 30% without concomitant improvement of more than 1 variable by ≥ 30%		0.585 (0.466–0.703)	3.21 (0.97–5.44)	0.77	0.76	0.77	0.76
Any worsening of CHAQ, ESR by ≥ 30% and AJC by ≥ 10%	3*	0.572 (0.457–0.686)	5.42 (2.78–8.05)	0.65	0.88	0.85	0.71
Any worsening of CHAQ, ESR by ≥ 40% and AJC by ≥ 20%		0.534 (0.422–0.646)	7.25 (4.13–10.37)	0.58	0.92	0.88	0.67
Any worsening of CHAQ, ESR by ≥ 30% and AJC by ≥ 30%		0.534 (0.422–0.646)	7.25 (4.13–10.37)	0.58	0.92	0.88	0.67

^a Area under the curve value of candidate flare definition = sensitivity × specificity.

^b Likelihood ratio for positive test = sensitivity/(1–specificity).

^c Sensitivity: number of patients without flare receiving placebo/number of patients receiving placebo (n = 26);

^d Specificity: number of patients without flare on etanercept/number of patients treated on etanercept (n = 25);

^e Positive predictive value (PPV) : number of patients with flare receiving placebo/number of all patients with flare;

^f Negative predictive value: number of patients without flare on etanercept/number of all patients without flare;

*The AUC of Definition 1 is significantly larger than the AUC of Definition 2 (z = 4.02; p < 0.0001) or of Definition 3 (z = 2.87; p < 0.005). The AUC of Definitions 2 and 3 are not significantly different (z = 1.06; p = NS). There seems to be no statistically significant difference between the LR of Definition 1 compared to that of Definition 3 (z = 1.42; p > 0.15). The LR of Definitions 1 and 3 are both significantly larger than the LR of Definition 2 (z = 2.32–2.62, p < 0.021).

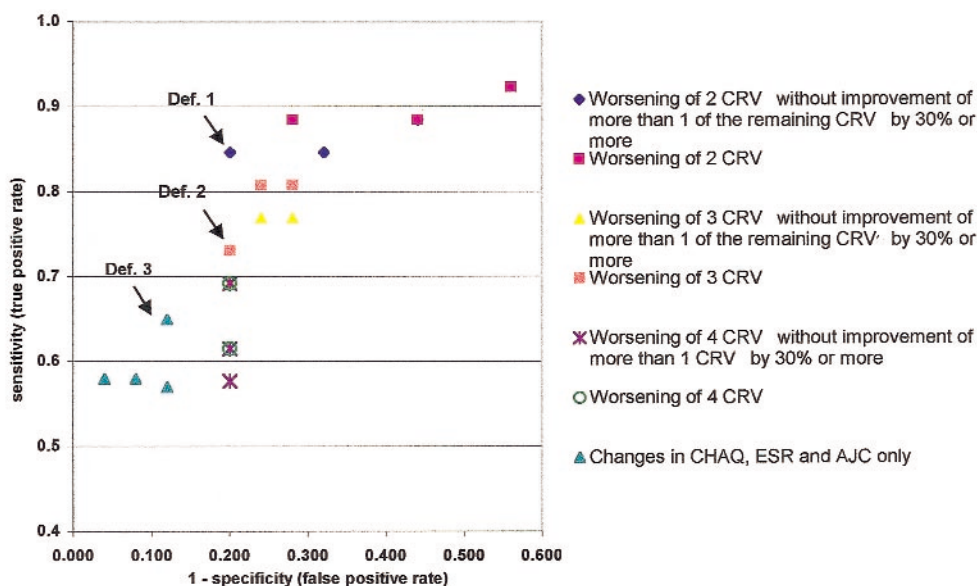


Figure 1. Receiver operator characteristic curve of the generated candidate definitions of disease flare. Various possible definitions of flare were derived by considering 20–50% changes of CRV of JRA or changes in the global assessments performed by parents or physicians. The 3 definitions appeared to be the most suitable candidate definitions of flare considering their sensitivity, specificity, and clinical relevance: Definition 1: at least 40% worsening of 2 CRV without concomitant improvement of more than one of the remaining CRV by 30% or more; Definition 2: at least 30% worsening of 3 CRV irrespective of concomitant improvement of the remaining CRV; Definition 3: any worsening of the CHAQ, increase of the ESR by at least 30%, and worsening of the AJC by at least 10%.

of the other candidate flare definitions. Also the lower limit of the 95% confidence interval of the AUC for this definition is smaller than 0.5, which is a sign in ROC analysis that the definition may not distinguish patients that flared from those that did not flare. However, this definition represents a more comprehensive assessment of the data, because the differential followup times of placebo versus etanercept treated patients are considered. This is not captured by the ROC curve, the ROC analysis (AUC, LR), or any other statistic for diagnostic tests used for this analysis.

Formal statistical comparison of the definitions 1–3. Various statistics have been developed to compare AUC and LR for diagnostic tests. Given the limited sample size of the study population such formal comparisons are of limited validity^{10,11}. The AUC of definition 1 is significantly larger than that of definition 2 ($z = 4.07$; $p < 0.0001$) or definition 3 ($z = 2.87$; $p < 0.005$), whereas the AUC of definition 2 and 3 are not statistically different ($z = 1.06$; $p = \text{NS}$) (Table 2). With regard to the LR, definitions 1 ($z = 2.62$; $p < 0.009$) and 3 ($z = 2.32$; $p < 0.021$) seem superior to definition 2. There is no important difference between the LR of definition 1 and 3 ($z = 1.42$; $p = \text{NS}$).

DISCUSSION

There is no universally accepted definition of disease flare in JRA, although flares of disease prompt clinicians to intensify therapies. This study examined possible definitions of flare and identified 3 flare definitions that seemed clinically and statistically most suitable. Disease flare

defined as at least 40% deterioration of 2 of the 6 CRV without concomitant improvement of more than one of the remaining CRV by 30% or more was overall best suitable to identify patients with flare against the chosen criterion standard (gold standard) of treatment with placebo. Two other definitions of flare with promising measurement characteristics were identified: (a) at least 30% worsening of 3 of the 6 CRV irrespective of concomitant improvement of remaining CRV and (b) any worsening of the CHAQ score, increase of the ESR by at least 30%, and increase of the AJC by at least 10%. The fact that good candidate flare definitions could be derived by using the CRV confirms the usefulness of the core set of outcome variables⁴ in JRA.

The flare definition used for the RCT of etanercept in JRA³ is very similar to Definition 2 identified in this study. Compared to Definition 2, the flare definition chosen for the RCT provides a more conservative rating of the efficacy of etanercept in JRA and was among the superior flare definitions identified in this study. The criterion standard for all the generated candidate flare definitions of this study was treatment with placebo. All included patients were therapy resistant prior to having a favorable response to etanercept during the open label phase of the RCT³. Therefore it was very likely that a deterioration of their disease would occur once active treatment to which the patient had shown clinical response was withdrawn, making treatment with placebo (instead of continued treatment with effective drug) a good surrogate for flare of disease.

Depending on the severity of a disease flare of a patient

with JRA, clinicians will increase the dose of the medications the patient is already taking, or additional therapeutic agents may be introduced, if a more severe increase in disease activity occurs. Recently, a consensus conference developed preliminary flare definitions for systemic lupus erythematosus (SLE)¹³. Definitions for minor and major flares were developed primarily based on the degree of changes in disease activity and the need to alter therapeutic regimens. Similarly, the development of criteria for minor versus major disease flare may be warranted for JRA to allow for more focused and standardized adjustments of therapy.

Other approaches could have been taken to generate candidate flare definitions. One strategy could have been to generate flare definitions by reversing the previously developed 9 final candidate definitions of improvement². An exploratory analysis was done to test those 5 of the 9 reversed final improvement definitions that had not already been examined. However, none of these additional 5 definitions was superior to those that had already been tested for the purpose of this study (all: AUC 0.480–0.585).

Limitations of this study are that followup on randomized treatment was discontinued once a patient met RCT flare criteria and that the study population was relatively small and preselected. This may have affected the sensitivity and specificity of the flare definitions and all statistics based on these values. However, PPV and NPV are statistics for diagnostic tests that are less influenced by the selection of the patient populations, and all 3 proposed preliminary definitions of disease flare had high PPV (all > 0.77) and NPV (all > 0.71).

Another limitation to the study is the small sample size of the study. It has been suggested in the past that at least 100 observations should be used for ROC analysis^{14,15}. Thus the current study is likely under powered. Given the limited sample size, we were also unable to test the generated flare definitions. Therefore a prospective study is warranted to test the usefulness of the proposed definitions of disease flare in clinical practice using a larger study population.

The development of a uniform preliminary definition of disease flare will help to better describe changes in disease activity in clinical practice. Thus, even prior to more definite prospective testing in a larger patient population and in

order to improve the comparability of studies in JRA, we propose the following as the preliminary flare definition: at least 40% worsening of 2 of the 6 CRV without concomitant improvement of more than one of the remaining CRV by 30% or more. The final definition of disease flare in JRA should be reliable and valid and its usefulness in clinical practice and research will require confirmation based on consensus and experience at an international level.

REFERENCES

1. Abrahmsen S, Alarcon G, Arend W, et al. Primer on the Rheumatic Diseases. In: Klippel JH, editor. Atlanta: Arthritis Foundation;1997:455.
2. 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.
3. Lovell DJ, Giannini EH, Reiff A, et al. Etanercept in children with polyarticular juvenile rheumatoid arthritis. Pediatric Rheumatology Collaborative Study Group. *N Engl J Med* 2000;342:763-9.
4. Giannini EH, Lovell DJ, Felson DT, Goldsmith CH. Preliminary core set of outcome variables for use in JRA clinical trials [abstract]. *Arthritis Rheum* 1994;37 Suppl:S428.
5. Singh G, Athreya BH, Fries JF, Goldsmith DP. Measurement of health status in children with juvenile rheumatoid arthritis. *Arthritis Rheum* 1994;37:1761-9.
6. Dobson A. An introduction to generalized linear models. London: Chapman and Hall;1990.
7. Nettleman MD. Receiver operator characteristic (ROC) curves. *Infect Control Hosp Epidemiol* 1988;9:374-7.
8. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Control Clin Trials* 1991;12:142S-158S.
9. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristics curves derived from the same cases. *Radiology* 1983;148:839-43.
10. Lee WC. Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. *Stat Med* 1999;18:455-71.
11. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic tests. *J Clin Epidemiol* 1991;44:763-77.
12. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diagn Imaging* 1989;29:307-35.
13. Petri M, Buyon J, Kim M. Classification and definition of major flares in SLE clinical trials. *Lupus* 1999;8:685-91.
14. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978;8:283-98.
15. Obuchowski NA. Sample size calculations in studies of test accuracy. *Stat Methods Med Res* 1998;7:71-92.