

# Which outcome measures in SLE clinical trials best reflect medical judgment?

Aikaterini Thanou,<sup>1</sup> Eliza Chakravarty,<sup>1</sup> Judith A James,<sup>1,2</sup> Joan T Merrill<sup>1</sup>

**To cite:** Thanou A, Chakravarty E, James JA, *et al.* Which outcome measures in SLE clinical trials best reflect medical judgment? *Lupus Science & Medicine* 2014;**1**:e000005. doi:10.1136/lupus-2013-000005

Received 19 December 2013  
Revised 28 January 2014  
Accepted 31 January 2014

## ABSTRACT

**Objectives:** To compare two measures of systemic lupus erythematosus (SLE) response: the British Isles Lupus Assessment Group (BILAG)-based Composite Lupus Assessment (BICLA) and the Systemic Lupus Responder Index (SRI) against a clinician's assessment of improvement.

**Methods:** Ninety-one lupus patients were identified with two visits at which Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) and BILAG had been scored and with active disease (SLEDAI $\geq$ 6) at the first visit. A physician rated the disease activity at the second visit as clinically significant improvement, no change or worsening. SRI and BICLA were scored both with and without the medication criteria often used in trials to restrict response definitions.

**Results:** 68 patients were considered improved, 17 same and 6 worse at follow-up. SRI versus BICLA, performed without considering medication changes, captured physician-rated improvement with 85% vs 76% sensitivity and 74% vs 78% specificity. With medication limits both instruments had 37% sensitivity and 96% specificity for physician-assessed improvement. Seven patients considered improved by the clinician met the BICLA but not the SRI definition of improvement by failing to achieve a four-point improvement in SLEDAI. 13 clinician-rated responders met SRI but not BICLA by improving in less than all organs.

**Conclusions:** Shortfalls of SRI and BICLA may be due to BICLA only requiring partial improvement but in all organs versus SRI requiring full improvement in some manifestation(s) and not all organs. SRI and BICLA with medication restrictions are less likely to denote response when the physician disagrees and could provide stringent proof of efficacy in appropriately powered clinical trials.

## INTRODUCTION

Despite many promising lupus treatments reaching early-phase human studies over the past several decades, none except belimumab has yet demonstrated efficacy in phase III trials. Attempts to explain the disappointing results of most clinical trials in lupus have pointed to pitfalls in the application

## KEY MESSAGES

- ▶ Shortfalls of SRI and BICLA may be due to BICLA requiring only partial improvement but in all organs versus SRI requiring full improvement in some manifestations and not necessarily in all organs.
- ▶ BICLA may be less sensitive than SRI in determining clinically significant improvement, particularly in SLE patients with multiple organs involved at baseline.
- ▶ Each instrument could likely be an optimal primary endpoint depending on the population under study and the design of a given clinical trial.

and interpretation of clinical endpoints.<sup>1</sup> According to recommendations by the Food and Drug Administration, European League Against Rheumatism and Outcome Measures in Rheumatology, the ideal endpoint should be able to detect both improvement and worsening in different manifestations and discern acute lupus disease activity from chronic damage and changes related to other causes.<sup>2</sup> Detection of both improvement and worsening is not possible using a global measure of disease activity such as the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI), and it is not a trivial undertaking even with an organ-based assessment such as the British Isles Lupus Assessment Group (BILAG) since different events within one organ may not be differentiated even with that complex instrument. This led to proposals of composite endpoints to detect improvement without worsening, two of which have already been widely incorporated in clinical trials.<sup>3–5</sup>

The SLE Responder Index (SRI) was derived in part by posthoc analysis of data from a failed phase II study of belimumab.<sup>2</sup> This was subsequently used as the primary outcome measure in two successful phase III trials (BLISS-52 and BLISS-76), leading to



<sup>1</sup>Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma, USA

<sup>2</sup>University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma, USA

**Correspondence to**  
Dr Aikaterini Thanou;  
aikaterini-thanou@omrf.org

belimumab approval by regulatory agencies.<sup>3, 4</sup> The SRI consists of scores derived from the SLEDAI and the BILAG Index: (1)  $\geq 4$ -point reduction in SLEDAI global score, (2) no new severe disease activity (BILAG A organ score) or more than one new moderate organ score (BILAG B) and (3) no deterioration from baseline in the physician's global assessment ( $\leq 10\%$  of scale). Although a requirement for no change in treatment was not formally included in the SRI definition, initiation of off-protocol medications defined non-response in the BLISS studies. Based on the success of the BLISS programme, SRI is currently being used in a number of clinical trials for SLE. Nonetheless, the treatment effect achieved by adding belimumab to standard of care using the original SRI (four-point drop in SLEDAI) was at best modest, and it remains unclear whether this is the optimal discriminatory endpoint.

The BILAG-Based Composite Lupus Assessment (BICLA) was derived by expert consensus<sup>6</sup> and employed as the primary endpoint in the EMBLEM trial of epratuzumab in lupus, where it appeared to discriminate well between standard of care and epratuzumab added to standard of care, at least in some doses tested.<sup>5</sup> BICLA response is defined as (1) at least one gradation of improvement in baseline BILAG scores in all body systems with moderate or severe disease activity at entry (eg, all A (severe disease) scores falling to B (moderate), C (mild), or D (no activity) and all B scores falling to C or D); (2) no new BILAG A or more than one new BILAG B scores; (3) no worsening of total SLEDAI score from baseline; (4)  $\leq 10\%$  deterioration in physicians global assessment and (5) no initiation of non-protocol treatment. It should be noted that the EMBLEM protocol was more restrictive than the BLISS protocols in the increases allowed in background medications at baseline.

The BICLA and SRI might be compared in two possible ways. The clinical components could be studied in isolation of the medication restrictions in order to simply determine presence or absence of clinical improvement. Adding medication restrictions (as has been done in clinical trials) allows an assessment of 'response' to an intervention given at baseline without clouding that assessment by the impact of other medications that might have been added when patients were not responding to that treatment. Direct comparison of BICLA and SRI has been addressed by only few studies to date,<sup>7, 8</sup> and the impact of the requirement for no medication changes on the simple determination of improvement has not yet been examined for either instrument. In this real-world exercise, the BICLA and SRI definitions were compared with physician's clinical assessment of change in disease state with and without medication restriction rules to distinguish between improvement and response.

## MATERIALS AND METHODS

This study was performed using data from the Oklahoma Lupus Cohort study. All patients underwent

informed consent procedures and Health Insurance Portability and Accountability Act disclosures in compliance with good clinical practice.

Individuals were identified who met the 1997 modified American College of Rheumatology classification criteria for SLE,<sup>9</sup> had two cohort study encounters at which the BILAG, SLEDAI and physician's global assessment (PGA) had been scored and had a SLEDAI  $\geq 6$  at the earlier (baseline) visit. In a minor deviation from either the SELENA-SLEDAI (used in the SRI) or the SLEDAI-2K (used in the BICLA), a Hybrid SLEDAI is used for this cohort (identical to the SELENA-SLEDAI except for the scoring of proteinuria, which uses the SLEDAI-2K definition). During a quality check of the data, where discrepancies were found between SLEDAI and BILAG scores and the original clinic note, these were corrected based on the original source documentation. The following data were determined retrospectively based on the medical record: some PGA values that had not been scored at the time of the clinic encounters, and an overall physician impression of clinical change at the later encounter (follow-up visit) as compared with the first visit, determined by review of clinic notes and laboratory values but without consideration of what treatments had been given in the interim. This was categorised as clinically significant improvement (physician-rated improvement (PRI)), deterioration or no change. Medication changes between baseline and follow-up and at the follow-up visit were recorded, and the SRI-4 (SRI) and BICLA composite responses were calculated, both with and without consideration of the medication restrictions that characterised the major clinical trials. SRI-3 and SRI-5 were computed similarly to SRI-4, except for a minimal three-point or five-point improvement in SLEDAI being required, respectively.<sup>2</sup> When medication criteria were applied, improvement criteria were not considered to be met at any follow-up visit where lupus medication changes were made after the initial changes that were instituted at baseline with the exception of NSAIDs or topical agents. The performance of BICLA and SRI was compared against clinician-rated improvement. Potential causes of discrepancies in each instrument were explored, including the role of each component endpoint.

## Statistical analysis

Descriptive statistics (mean SD) were used to describe measures of disease activity (PGA, SLEDAI, global BILAG and number of BILAG domains involved) at baseline and follow-up. Comparison of disease activity measures between baseline and follow-up was performed by paired t test. Disease activity measures between PRI responders that did or did not meet the BICLA or SRI endpoints were compared by Mann-Whitney rank-sum test. Spearman's rank test was used to correlate PRI with SRI and BICLA responses. The number of BILAG domains with persistent and/or B scores at follow-up among PRI responders stratified by the number of

**Table 1** Disease activity at baseline and follow-up among patients clinically improving (n=68) or not (n=23)

Improving by PRI (n=68)	Baseline Mean (SD)	Follow-up Mean (SD)	p Value (baseline to follow-up)*
PGA	1.7 (0.4)	0.8 (0.5)	<0.001
SLEDAI	10.4 (4.4)	3.6 (3.7)	<0.001
Global BILAG	15.7 (6.9)	4.8 (5.4)	<0.001
Number of BILAG domains (A, B, C)	2.8 (0.9)	1.8 (1.0)	<0.001
Number of BILAG domains (A, B)	1.7 (0.8)	0.4 (0.7)	<0.001
<b>Not improving by PRI (n=23)</b>			
PGA	1.7 (0.3)	1.7 (0.5)	0.541
SLEDAI	9.0 (2.7)	8.7 (4.3)	0.747
Global BILAG	13.3 (4.8)	12.5 (8.1)	0.615
BILAG domains (A, B, C)	2.3 (0.6)	2.2 (1.1)	0.732
BILAG domains (A, B)	1.5 (0.7)	1.4 (0.9)	0.680

\*Paired t test.

BILAG, British Isles Lupus Assessment Group; PGA, physician's global assessment; PRI, physician-rated improvement; SLEDAI, Systemic Lupus Erythematosus Disease Activity Index.

BILAG A and/or B scores at baseline was compared by Fisher's exact test. SigmaPlot V.12.5 (Systat Software, Inc) was used for all statistical analyses.

## RESULTS

### Demographic characteristics

In total, 91 patients eligible for the analysis were identified, including 86 women and 5 men with SLE. The mean age at the time of the baseline visit was 41 years (SD 11.8), ranging from 21 to 68. Also, 46 subjects were Caucasian, 22 African-American, 14 Native American, 5 Caucasian/Hispanic and 4 Asian. The patients in this analysis were all assessed between June 2009 and April 2012 with a mean interval of 6.24 months between baseline and follow-up visits (range 1–25 months).

### The PRI was associated with improvements in disease activity scores between baseline and follow-up

In total, 68 patients improved by PRI, 17 remained the same and 6 deteriorated. Disease activity in patients who did or did not improve by PRI was compared using the PGA, SLEDAI, global BILAG scores and the number of BILAG organ domains involved (table 1). A statistically significant improvement between baseline and follow-up in all disease activity instruments was found in patients improving by PRI, but no differences in those indices were observed in group not deemed by the physician to be

improved. This supports previous literature that outcome measures discriminate clinically relevant differences.<sup>10 11</sup>

### Sensitivity of SRI compared with BICLA in detecting PRI

Of 68 patients determined to be improving by the physician (PRI), 58 met the clinical criteria for SRI and 52 met the BICLA endpoint, with sensitivity of agreement with PRI 85% and 76%, respectively (table 2). This evaluation was for improvement not response and did not include the restriction that change of medications over-rules 'response' to the baseline treatment. Of the 23 patients not improving by PRI, 17 did not meet the SRI clinical criteria and 18 did not improve by the BICLA; therefore, specificity of SRI and BICLA for PRI was 74% and 78%. Sensitivities and specificities were similarly computed for SRI-3 and SRI-5. Spearman's rank correlations to PRI were SRI-3 0.605, SRI-4 0.563, SRI-5 0.541 and BICLA 0.492 (all p values<0.000001).

Of the 10 out of 68 patients improving by PRI but not by SRI, 7 did meet the BICLA improving criteria. These seven SRI-negative/BICLA-positive discordant cases failed the SRI because of less than four-point improvement on SLEDAI, which scores mild, moderate or severe arthritis four points and mild, moderate or severe rash two points (six individuals had a two-point improvement and one had a three-point improvement; therefore, all did have complete resolution of at least one feature, but the scores given to these particular features were not

**Table 2** Sensitivity and specificity of SRI and BICLA compared with PRI

PRI R (n=68) without medication criteria			PRI non-R (n=23) without medication criteria		
	n	Sensitivity vs PRI		n	Specificity vs PRI
SRI-3 R	60	88	SRI-3 non-R	17	74
SRI-4 R	58	85	SRI-4 non-R	17	74
SRI-5 R	48	71	SRI-5 non-R	21	91
BICLA R	52	76	BICLA non-R	18	78

BILAG, British Isles Lupus Assessment Group; BICLA, BILAG-based Composite Lupus Assessment; non-R, non-responders; PRI, physician-rated improvement; R, responders; SRI, Systemic Lupus Responder Index.

high enough to define improvement by SRI). Among 16 patients who improved by PRI but not BICLA, 13 were defined as improving by SRI. These 13 SRI-positive/BICLA-negative discordant cases failed BICLA because of one or more organ systems without even partial improvement, despite complete resolution of at least one other feature counting four points on the SLEDAI. Persistent B (moderate) scores in the musculoskeletal and mucocutaneous domains were present in six patients each, and there was one persistent B in each of the following domains: gastrointestinal (lupoid hepatitis), cardiorespiratory (pleurisy) and renal (proteinuria). One patient had a persistent A (severe) score in the cardiorespiratory domain (interstitial lung disease).

Six of the twenty-three patients who did not improve by PRI met the SRI criteria, three of whom did not improve by BICLA. These three SRI/BICLA discordant cases achieved four-point reduction in SLEDAI due to resolving mild manifestations (rash, alopecia, mucosal ulcers) despite other ongoing (but not worsening) more severe organ involvement. Five PRI failures met the BICLA, two of which did not meet the SRI. One of these two patients developed new BILAG B arthritis, and one new B score is allowed in the BICLA definition of improvement. Another had only moderate improvement in arthritis (musculoskeletal BILAG A score (severe arthritis) decreased to a high B (significant moderate arthritis)), but the patients were globally judged as overall clinically unchanged.

#### Those considered improved by the physician but failed the BICLA had more active organs at baseline

Since patients improving by PRI often failed BICLA (false negative) because improvement did not occur in all domains, we hypothesised this may be more common if there is a greater number of organs active at baseline. Therefore, the total number of BILAG domains involved in BICLA improvement versus non-improvement was assessed in those patients meeting PRI (table 3).

BICLA responders tended to have fewer domains involved at baseline compared with non-responders, an observation that was more pronounced when only organs with baseline A (severe disease) and B (moderate disease) were counted. This supports the hypothesis that the more BILAG organs are significantly involved at baseline, the more likely it is to have persistent disease in one or more organs at follow-up despite significant improvement in some aspects of the disease and an overall clinical impression of improvement. Among patients improving by PRI, 33.3% of those with two or more BILAG A or B scores at baseline had persistent A or B scores at follow-up (BICLA failure) compared with 10.3% in those with one or less BILAG A or B scores at baseline (Fisher's exact test,  $p=0.0924$ ). This suggests that the BICLA might be less sensitive at detecting improvement when more organs are involved.

PRI responders meeting the BICLA criteria also had less disease activity at follow-up compared with those not meeting the BICLA, evident by the PGA, SLEDAI and global BILAG scores, as well as the number of BILAG organs involved ( $p<0.02$  for all comparisons), indicating the possibility that the BICLA may provide some additional meaningful discrimination beyond what is captured by PRI or be considered a higher bar for determining improvement.

Disease activity indices at baseline and follow-up were not as informative when comparing those improved by PRI who did or did not meet the SRI criteria (table 4). With the exception of PGA, there was no difference in baseline disease activity measures in those improving by SRI versus failures.

#### Response differs from improvement: the addition of medication criteria dramatically decreased the number of patients meeting the BICLA and SRI response definitions

Clinical trial endpoints using the SRI and BICLA constructs to determine response to a treatment initiated at baseline have used restrictions in rescue medications in

**Table 3** Comparison of disease activity indices between PRI responders that met ( $n=52$ ) or not ( $n=16$ ) the BICLA endpoint

Baseline	PRI R/BICLA R ( $n=52$ ) Mean (SD)	PRI R/BICLA non-R ( $n=16$ ) Mean (SD)	p Value*
PGA	1.7 (0.4)	1.8 (0.3)	0.114
SLEDAI	10.0 (4.2)	11.6 (5.1)	0.203
Global BILAG	14.8 (7.0)	18.7 (6.1)	0.027
Number of BILAG domains (A, B, C)	2.6 (0.9)	3.1 (0.8)	0.065
Number of BILAG domains (A, B)	1.6 (0.8)	2.1 (0.7)	0.024
<b>Follow-up</b>			
PGA	0.6 (0.4)	1.1 (0.4)	<0.001
SLEDAI	3.0 (3.5)	5.6 (3.5)	0.003
Global BILAG	2.7 (3.3)	11.6 (5.1)	<0.001
Number of BILAG domains (A, B, C)	1.6 (0.9)	2.3 (1.0)	0.019
Number of BILAG domains (A, B)	0.1 (0.4)	1.2 (0.6)	<0.001

\*Mann-Whitney rank-sum test.

BILAG, British Isles Lupus Assessment Group; BILCA, BILAG-based Composite Lupus Assessment; non-R, non-responders; PGA, physician's global assessment; PRI, physician-rated improvement; R, responders; SLEDAI, Systemic Lupus Erythematosus Disease Activity Index.



**Table 4** Comparison of disease activity indices between PRI responders that met (n=58) or not (n=10) the SRI endpoint

Baseline	PRI R/SRI R (n=58) Mean (SD)	PRI R/SRI non-R (n=10) Mean (SD)	p Value*
PGA	1.7 (0.4)	2.0 (0.5)	0.042
SLEDAI	10.6 (4.5)	8.9 (4.2)	0.121
Global BILAG	15.4 (6.7)	17.6 (8.4)	0.497
Number of BILAG domains (A, B, C)	2.8 (0.9)	2.6 (1.0)	0.827
Number of BILAG domains (A, B)	1.7 (0.8)	1.8 (0.9)	0.883
<b>Follow-up</b>			
PGA	0.7 (0.5)	0.8 (0.5)	0.388
SLEDAI	3.0 (3.2)	7.0 (4.8)	0.002
Global BILAG	4.3 (5.1)	7.3 (6.3)	0.084
Number of BILAG domains (A, B, C)	1.7 (1.0)	2.0 (0.9)	0.433
Number of BILAG domains (A, B)	0.4 (0.6)	0.7 (0.8)	0.165

\*Mann–Whitney rank-sum test.

BILAG, British Isles Lupus Assessment Group; non-R, non-responders; PGA, physician's global assessment; PRI, physician-rated improvement; R, responders; SLEDAI, Systemic Lupus Erythematosus Disease Activity Index; SRI, Systemic Lupus Responder Index.

the definition of response (underscoring a distinction between the concepts of improvement and response). A requirement for no off-protocol medication changes was included in the BICLA response definition, used in the EMBEM trial.<sup>5</sup> In the BLISS trials of belimumab in SLE where the SRI was used,<sup>3 4</sup> medication increases were restricted during the latter months, and protocol deviations defined non-response accomplishing a similar end, although medication restrictions were less restrictive in those protocols than in the EMBLEM study. To examine the impact of medication restrictions on how BICLA and SRI compare to physician's determination of improvement, we performed the SRI and BICLA determination in this real-world sample of patients (no protocolised restriction on treatment). All pharmacological changes in treatment after baseline interventions and prior to and/or at follow-up visit, except for NSAIDs or topical agents, excluded the designation of improvement by BICLA or SRI in this analysis. As expected, medication restrictions dramatically decreased the number of patients meeting BICLA and SRI response definitions. Surprisingly, when medications were factored in as 'response' criteria, this increased specificity for PRI (decreased the likelihood that BICLA and SRI will detect improvement when the physician did not agree), even though the physician was not considering the treatment changes in the PRI assessment (table 5).

## DISCUSSION

The performance of two composite indices of lupus disease improvement and/or response was compared using as reference a clinician's global rating of improvement. Although the PGA, when performed in a paper patient exercise, has been demonstrated to have poor agreement between different physicians,<sup>12</sup> a retrospective overall determination of whether or not a patient has a clinically meaningful improvement might be a valuable tool when determined in a consistent manner by one qualified assessor. This was supported in the current exercise, where changes in disease activity (including SLEDAI and BILAG performed prospectively at the time of the visits) were consistent with the global clinician ratings. Retrospectively assessing clinical change between visits is possibly limiting the validity of our observation and warrants conformation prospectively. Although the visits used to determine improvement or response were separated by various timepoints in the current study, this did not affect the global clinical assessment of change between visits and is consistent with the manner in which SRI and BICLA are used in clinical trials as landmark assessments performed at different timepoints compared with a baseline visit.

The assessment of response to an intervention is not the same as a measurement of clinical improvement (which may occur without response to a given

**Table 5** Sensitivity and specificity of SRI and BICLA with medication criteria

PRI R (n=68) with medication criteria			PRI non-R (n=23) with medication criteria		
	n	Sensitivity vs PRI		n	Sensitivity vs PRI
SRI-3 R	27	40	SRI-3 non-R	22	96
SRI-4 R	25	37	SRI-4 non-R	22	96
SRI-5 R	19	28	SRI-5 non-R	22	96
BICLA R	24	35.3	BICLA non-R	22	96

BILCA, British Isles Lupus Assessment Group-based Composite Lupus Assessment; non-R, non-responders; PRI, physician-rated improvement; R, responders; SRI, Systemic Lupus Responder Index.

intervention if rescue treatments have been given). A comparison of BICLA and SRI without using the medication criteria for 'response' to the baseline treatment was first performed to determine their utility in simply defining clinical improvement. In this assessment, BICLA was less sensitive than SRI at capturing physician-determined improvement. When only those patients ranked as improved by the physician were evaluated, BICLA but not SRI seemed to define patients with greater improvement in disease activity, suggesting the possibility that it could be more discriminatory in some settings. Since more organs involved at baseline decreased the sensitivity of the BICLA, it can be hypothesised that this instrument might be particularly useful in patients with less widespread organ involvement, including those patients considered suitable for clinical trials in which less background medication is allowed. In fact, in the EMBLEM trial of epratuzumab in lupus, which limited background medication changes more restrictively than the belimumab trials, the BICLA response (including the medication criteria) at 12 weeks effectively discriminated the 2400 mg/month cumulative epratuzumab dosage groups from placebo with placebo responses lower than in the BLISS trials.<sup>5</sup>

A comparison of two outcome measures used in different trials does not account for potential differences in patient populations, background treatments or efficacy of the test articles. However, it is worth observing that the apparent improved discriminatory capacity in the EMBLEM trial (using the BICLA) did not appear to be due to increased efficacy rates in the treatment group but to lower rates in the placebo (standard of care) group,<sup>7</sup> which does not suggest increased sensitivity in detection of improvement, but might either reflect a more generally ill population, less background treatments or, indeed, the possibility of increased specificity of the improvement measurement. In posthoc analysis of this same trial dataset using the SRI,<sup>7</sup> SRI rates were higher than BICLA rates in all arms, including the placebo group, losing discriminatory capacity with a loss of significant differences between treatment and placebo. Disagreement between BICLA and SRI in the EMBLEM trial may, however, have been driven by the baseline distribution of items with high SLEDAI weights that tended to improve at follow-up, with the greatest difference in groups with activity scored as eight-point SLEDAI items (vasculitis, lupus headache) at baseline. Scoring SLEDAI descriptors for mild/moderate cutaneous vasculitis and headache risks SLEDAI scores that are high in relation to the degree of illness of the patients. SRI discriminatory performance might improve in a protocol, restricting the scoring of these features. In the current analysis, no lupus headaches were scored (consistent with the evaluation here that they are very rare), and although four individuals had cutaneous vasculitis at baseline, this resolved in only one at follow-up, thereby not accounting for discrepancies in SRI and BICLA performance found here.

Some PRI responders failed the SRI, providing potential insight into limitations of the SRI. Patients rated as improving by the physician usually fail to meet the SRI improvement criteria due to less than four-point improvement in SLEDAI, which requires not only improvement but complete resolution of at least one manifestation. As expected, this limitation was less evident for SRI-3 at the expense of lower specificity for the detection of PRI, whereas the opposite was the case for the SRI-5. Our results are consistent with the posthoc sensitivity analysis of data from the BLISS-76 trial, where modifications of the SRI using higher thresholds for SELENA-SLEDAI improvement ( $\geq 5$ -point to  $\geq 10$ -point reductions) increased differentiation of belimumab from placebo at 52 and 76 weeks but occurred less frequently.<sup>4</sup>

## CONCLUSIONS

In an assessment using a physician's global rating of clinically significant improvement, the BICLA may be less sensitive than the SRI, particularly in determining improvement in patients with SLE with multiple organs involved at baseline. On the other hand, the BICLA may be more discriminatory than the SRI in selecting those patients with a greater change in disease activity. When SRI and BICLA were discrepant, this was usually due to the BICLA requiring only partial improvement but in all organs versus the SRI requiring full improvement and not necessarily in all organs. Each instrument could likely be an optimal primary endpoint depending on the population under study and the design of a given clinical trial.

**Contributors** All authors fulfil the criteria of authorship and no one else who fulfils these criteria has been excluded from the list of authors.

**Competing interests** None.

**Ethics approval** Oklahoma Medical Research Foundation IRB.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** No additional data are available.

**Open Access** This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/3.0/>

## REFERENCES

1. Bruce IN, Gordon C, Merrill JT, *et al.* Clinical trials in lupus: what have we learned so far? *Rheumatology (Oxford)* 2010;49:1025–7.
2. Furie RA, Petri MA, Wallace DJ, *et al.* Novel evidence-based systemic lupus erythematosus responder index. *Arthritis Rheum* 2009;61:1143–51.
3. Navarra SV, Guzman RM, Gallacher AE, *et al.* Efficacy and safety of belimumab in patients with active systemic lupus erythematosus: a randomised, placebo-controlled, phase 3 trial. *Lancet* 2011;377:721–31.
4. Furie R, Petri M, Zamani O, *et al.* A phase III, randomized, placebo-controlled study of belimumab, a monoclonal antibody that inhibits B lymphocyte stimulator, in patients with systemic lupus erythematosus. *Arthritis Rheum* 2011;63:3918–30.

5. Wallace DJ, Kalunian K, Petri MA, *et al.* Efficacy and safety of epratuzumab in patients with moderate/severe active systemic lupus erythematosus: results from EMBLEM, a phase IIb, randomised, double-blind, placebo-controlled, multicentre study. *Ann Rheum Dis* 2014;73:183–90.
6. Wallace D, Strand V, Furie R, *et al.* Evaluation of treatment success in Systemic Lupus Erythematosus Clinical Trials: development of the British Isles Lupus Assessment Group-Based Composite Lupus Assessment Endpoint [abstract]. *Arthritis Rheum* 2011;63(Suppl 10):S885.
7. Petri M, Pike MC, Kelley L, *et al.* Systemic Lupus Erythematosus Responder Index Assessment of Responders in EMBLEM, a Phase IIb Study in patients with moderate to Severe Systemic Lupus Erythematosus. *Arthritis Rheum* 2011;63(Suppl 10):S539.
8. Sridharan ST, Zhou T, Immermann F, *et al.* Low placebo responses and clinical components of the Biomarkers of Lupus Disease (BOLD) study may provide useful insights for Systemic Lupus Erythematosus Clinical Trial Design. *Arthritis Rheum Suppl* 2011;63 (Suppl 10):S543.
9. Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* 1997;40:1725.
10. Griffiths B, Mosca M, Gordon C. Assessment of patients with systemic lupus erythematosus and the use of lupus disease activity indices. *Best Pract Res Clin Rheumatol* 2005;19:685–708.
11. Yee CS, Farewell V, Isenberg DA, *et al.* British Isles Lupus Assessment Group 2004 index is valid for assessment of disease activity in systemic lupus erythematosus. *Arthritis Rheum* 2007;56:4113–19.
12. Yee CS, Farewell VT, Isenberg DA, *et al.* The use of Systemic Lupus Erythematosus Disease Activity Index-2000 to define active disease and minimal clinically meaningful change based on data from a large cohort of systemic lupus erythematosus patients. *Rheumatology (Oxford)* 2011;50:982–8.