Arthritis Care & Research
Vol. 70, No. 1, January 2018, pp 98–103
DOI 10.1002/acr.23252

ORIGINAL ARTICLE

Study of Flare Assessment in Systemic Lupus Erythematosus Based on Paper Patients

D. ISENBERG,[1] J. STURGESS,[2] E. ALLEN,[2] C. ARANOW,[3] A. ASKANASE,[4] B. SANG-CHEOL,[5] S. BERNATSKY,[6] I. BRUCE,[7] J. BUYON,[8] R. CERVERA,[9] A. CLARKE,[10] MARY ANNE DOOLEY,[11] P. FORTIN,[12] E. GINZLER,[13] D. GLADMAN,[14] J. HANLY,[15] M. INANC,[16] S. JACOBSEN,[17] D. KAMEN,[18] M. KHAMASHTA,[19] S. LIM,[20] S. MANZI,[21] O. NIVED,[22] C. PESCHKEN,[23] M. PETRI,[24] K. KALUNIAN,[25] A. RAHMAN,[1] R. RAMSEY-GOLDMAN,[26] J. ROMERO-DIAZ,[27] G. RUIZ-IRASTORZA,[28] J. SANCHEZ-GUERRERO,[29] K. STEINSSON,[30] G. STURFELT,[22] M. UROWITZ,[14] R. VAN VOLLENHOVEN,[31] D. J. WALLACE,[32] A. ZOMA,[33] J. MERRILL,[34] AND C. GORDON[35]

Objective. **To determine the level of agreement of disease flare severity (distinguishing severe, moderate, and mild flare and persistent disease activity) in a large paper-patient exercise involving 988 individual cases of systemic lupus erythematosus.**
Methods. **A total of 988 individual lupus case histories were assessed by 3 individual physicians. Complete agreement about the degree of flare (or persistent disease activity) was obtained in 451 cases (46%), and these provided the reference standard for the second part of the study. This component used 3 flare activity instruments (the British Isles Lupus Assessment Group [BILAG] 2004, Safety of Estrogens in Lupus Erythematosus National Assessment [SELENA] flare index [SFI] and the revised SELENA flare index [rSFI]). The 451 patient case histories were distributed to 18 pairs of physicians, carefully randomized in a manner designed to ensure a fair case mix and equal distribution of flare according to severity.**
Results. **The 3-physician assessment of flare matched the level of flare using the 3 indices, with 67% for BILAG 2004, 72% for SFI, and 70% for rSFI. The corresponding weighted kappa coefficients for each instrument were 0.82, 0.59, and 0.74, respectively. We undertook a detailed analysis of the discrepant cases and several factors emerged, including a tendency to score moderate flares as severe and persistent activity as flare, especially when the SFI and rSFI instruments were used. Overscoring was also driven by scoring treatment change as flare, even if there were no new or worsening clinical features.**
Conclusion. **Given the complexity of assessing lupus flare, we were encouraged by the overall results reported. However, the problem of capturing lupus flare accurately is not completely solved.**

INTRODUCTION

In the past 30 years, methods of assessing disease activity in patients with lupus have improved considerably. Both global score systems, such as the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI), the Systemic Lupus Activity Measures, and the European Community Lupus Activity Measure, and more specific instruments such as the British Isles Lupus Assessment Group (BILAG) system, which focuses on individual organs or systems, have emerged as viable and effective activity assessment instruments. SLEDAI and BILAG have been revised to SLEDAI-2K (1) and the BILAG 2004 index

Supported by a grant from a combined American College of Rheumatology and European League Against Rheumatism committee.

[1]D. Isenberg, MD, A. Rahman, PhD: University College London, London, UK; [2]J. Sturgess, PhD, E. Allen, PhD: The Hospital For Tropical Diseases, London, UK; [3]C. Aranow, MD: Feinstein Institute for Medical Research, Manhasset, New York; [4]A. Askanase, MD: Columbia University, New York, New York; [5]B. Sang-Cheol, MD: Hanyang University Hospital for Rheumatic Diseases, Seoul, South Korea; [6]S. Bernatsky, MD: McGill University, Quebec, Ontario, Canada; [7]I. Bruce, PhD: The University of Manchester, Central Manchester University Hospitals NHS Foundation Trust, and Manchester Academic Health Science Centre, Manchester, UK; [8]J. Buyon, MD: New York School of Medicine, New York; [9]R. Cervera, MD: Universitat de Barcelona, Barcelona, Spain; [10]A. Clarke, MD: Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada; [11]Mary Anne Dooley, MD: University of North Carolina, Chapel Hill; [12]P. Fortin, MD: Université Laval, Quebec City, Québec, Canada; [13]E. Ginzler, MD: Downstate Medical Center Rheumatology, Brooklyn, New York; [14]D. Gladman, MD, M. Urowitz, MD: Krembil Research Institute, Toronto Western Hospital, and University of Toronto, Toronto, Ontario, Canada; [15]J. Hanly, MD: Nova Scotia Rehabiliation Center, Halifax, Nova Scotia, Canada; [16]M. Inanc, MD: Istanbul University, Istanbul, Turkey; [17]S. Jacobsen, MD: Rigshospitalet, Copenhagen, Denmark; [18]D. Kamen, MD: Medical University of South Carolina,

Significance & Innovations

- This study addresses the ongoing dilemma of how best to capture flare in patients with systemic lupus erythematosus. In our previous attempt to measure flare using 16 live patients, we could only assess a modest number of lupus manifestations.
- In the current study we have used nearly 1,000 paper-based case histories to determine the capacity of 3 flare activity instruments (British Isles Lupus Assessment Group 2004 index, Safety of Estrogens in Lupus Erythematosus National Assessment [SELENA] flare index, and the revised SELENA flare index to capture flare.
- We show that all 3 instruments are able to do this in many cases, but there is an ongoing need to do even better.

(2), with large-scale studies undertaken to demonstrate their validity, reliability, and sensitivity to change (3).

Although both the SLEDAI and BILAG activity assessments are widely used in large-scale international trials of new biologic drugs (4–6), there are few data assessing their usefulness in determining clinical flares in patients with lupus (7). The Lupus Foundation of America held 2 investigator meetings in 2007 and 2008 that sought to agree on the definition of flare in lupus patients (8). They concluded that flare is "a measurable increase in disease activity in 1 or more organ systems involving new or worse clinical signs and symptoms and/or laboratory measurements. It must be considered clinically significant by the assessor and usually there will be at least consideration of initiation or increase in treatment."

A particular challenge in patients with lupus has been distinguishing mild, moderate, and severe flares and distinguishing them from ongoing, persistent disease. This problem is in part related to difficulties in agreeing what constitutes such flares in different organs and systems, but also because when a patient is flaring there may be a difference in the degree of flare in different organs or systems.

A live patient study of 16 flaring patients with lupus (9) took place in London in May 2009. In that study, 3 assessment instruments were used to assess flare. One was based on the BILAG 2004 index, a second was the classic Safety of Estrogens in Lupus Erythematosus National Assessment (SELENA) flare index (SFI), and the third was the revised SELENA flare index (rSFI), an organ-based system that is based on, but not directly linked to, the SLEDAI.

In that live patient study (9), a panel of rheumatologists, one of whom was the patient's own clinician, determined the severity of flare in each patient, and then different individual rheumatologists assessed each patient for flare with all 3 instruments. Intraclass correlation coefficients (with 95% confidence interval [95% CI]) were calculated to indicate a measure of internal reliability. The results were 0.54 (95% CI 0.32–0.78) for the BILAG 2004 flare assessment compared with 0.21 (95% CI 0.08–0.48) for the revised SELENA flare index and 0.18 (95% CI 0.06–0.45) for the physicians global assessment of flare. Severe flare was associated with good agreement between the 3 instruments, but the mild-to-moderate flares were less consistent. Although assessing real patients offers the tangible advantage of testing potential flare instruments in a more realistic way, obviously the numbers of patients (and their clinical problems) that can be studied at any given time is restricted. To expand our experience of flare assessment in lupus, we have now undertaken a major paper-patient–based exercise, which has allowed us to review a much broader array of lupus symptomatology and to use a larger number of assessors to determine flare status, using individual instruments on paper case reports based on real cases. The objective of this study was to determine the level of agreement of flare severity (severe/moderate/mild/none, i.e., persistent disease activity) identified in paper-patient cases using 3 flare instruments and physician-defined flares determined by a panel of 3 physicians.

MATERIALS AND METHODS

Generation of clinical case scenarios. Participating physicians were given a standardized report form to complete, which included 4 sections: previous lupus assessment and treatment, details of assessment at the visit being evaluated for flare, results of relevant investigations (blood, urine, imaging, biopsies, etc.) influencing this assessment, and treatment changes at this visit. Thirty physicians submitted a total of 988 anonymous individual paper case reports, based on the medical records of their patients. Each patient met the revised classification criteria of the American College of Rheumatology (10) and/or the 2012 Systemic Lupus International Collaborating Clinics criteria (11). The

Charleston; [19]M. Khamashta, FRCP: King's College London, London, UK; [20]S. Lim, MD: Emory University, Atlanta, Georgia; [21]S. Manzi, MD: Allegheny Health Network, Pittsburgh, Pennsylvania; [22]O. Nived, MD, G. Sturfelt, MD: Lund University, Lund, Sweden; [23]C. Peschken, MD: University of Manitoba, Winnipeg, Manitoba, Canada; [24]M. Petri, MD: Johns Hopkins University, Baltimore, Maryland; [25]K. Kalunian, MD: University of California at San Diego; [26]R. Ramsey-Goldman, MD: Northwestern University, Feinberg School of Medicine, Chicago, Illinois; [27]J. Romero-Diaz, MD: Instituto Nacional de Ciencias Médicas y Nutrición, Mexico City, Mexico; [28]G. Ruiz-Irastorza, MD: Hospital Universitario Cruces and University of the Basque Country, Barakaldo, Spain; [29]J. Sanchez-Guerrero, MD: Mount Sinai Hospital and University Health Network and University of Toronto, Toronto, Ontario, Canada; [30]K. Steinsson, MD: Landspitali University Hospital, Reykjavik, Iceland; [31]R. van Vollenhoven, MD: Karolinska University Hospital, Solna, Sweden; [32]D. J. Wallace, MD: University of California at Los Angeles; [33]A. Zoma, PhD: Hairmyres Hospital, East Kilbride, Scotland, UK; [34]J. Merrill, MD: Oklahoma Medical Research Foundation, Oklahoma City; [35]C. Gordon, PhD: College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK.

Address correspondence to D. Isenberg, MD, Centre for Rheumatology, University College London, London, UK. E-mail: d.isenberg@ucl.ac.uk.

Submitted for publication August 3, 2016; accepted in revised form April 4, 2017.

submitting physician was also asked to provide their assessment of the flare category for the paper-patient case (severe, moderate, or mild flare or persistent/ongoing disease) and to submit roughly equal numbers of each category. The actual distribution of paper-patient cases received was 25% severe, 28% moderate, 22% mild, and 25% persistent disease activity without flare.

Each patient case report was then assessed for flare category by 2 additional reviewing physicians who did not know the patient. The cases were allocated at random, and randomization was designed to ensure an equal distribution of submitted category types and to ensure that no physician reviewed their own submitted cases or the same case twice. The flare category assigned by each of the 3 physicians (1 submitting and 2 reviewing) was then compared.

All 3 physicians agreed on the flare category in 451 cases (46%). We refer to this flare assessment agreement as 3-physician consensus (TPC), and these are the cases that were carried forward to be assessed by the flare instruments in the study. The TPC result for each case formed the reference standard for our later analyses. Flare category distribution in these cases was 32% severe, 21% moderate, 24% mild, and 23% persistent disease without flare.

TPC was not achieved in 535 cases (54%). In 376 cases (38%) there was partial agreement, i.e., any 2 physicians agreed with each other (but not with the third physician), in 41 cases (4%) there was no agreement between any physician, i.e., all 3 physicians recorded different levels of flare or persistent activity for these cases. One or 2 of the reviewing physicians rejected 118 cases (12%) as unable to code on the information provided. Two of the authors (DI and CG) reviewed these cases to assess whether there were any particular reasons why it was difficult to achieve TPC.

The 451 TPC patient case histories were carefully reviewed by one author (DI) and assigned into 1 of 8 clinical groups: musculoskeletal and/or skin disease only, joint and/or skin and renal disease, mainly serositis, mainly renal, mainly gastrointestinal, mainly central nervous system, joints and/or skin plus serositis, and other, which included predominantly hematologic and/or constitutional or other combinations. Two cases were excluded from further assessment, as 1 was found to be a duplicate, and in the other the case report form had been completed incorrectly.

Assessment of TPC paper case reports using standard instruments. The 451 TPC patient case histories were distributed to 18 pairs of physicians. The cases were randomized in a manner designed to ensure each pair received a fair case mix, based on the clinical groups set out above and an equal distribution of flare by category type.

The 18 pairs of rheumatologists were each asked to agree on the level of flare in 20–26 individual paper cases. Each pair was assigned to use 1 of the 3 flare assessment instruments: the BILAG 2004 index (12) (6 pairs), the SFI (13) (6 pairs), and the rSFI (9) (6 pairs). Full details of the flare instruments are shown in Supplementary Appendices A–E (available on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.23252/abstract). Analyses were done both including and excluding flare defined only by treatment change.

The pairs of rheumatologists were asked to undertake the assessments of flare using the instrument assigned to them separately and then to confer to achieve full agreement on the score and level of flare using that instrument or to record persistent activity without flare.

Statistical methods used. For each flare instrument, the level of flare agreed by the assessing pair of physicians was compared to the TPC flare assessment (the reference standard) for each case. The level of agreement between the flare instrument used and the TPC flare assessment was determined in 2 ways, first by simply calculating the percentage of cases where there was complete agreement, and second by calculating a weighted kappa coefficient with quadratic weights (equivalent to an intraclass correlation coefficient). The weighted kappa coefficient gives weights to the frequencies in each cell of the table according to their distance from the diagonal, thus allowing different levels of agreement to contribute.

RESULTS

Of the 451 cases, 7 were rejected by the pairs of physicians for providing too little information to score using the assigned flare instrument (6 cases for the BILAG 2004 index and 1 for the SFI flare instrument). Each type of flare that was assessed by the BILAG 2004 is shown in Table 1, by the SFI in Table 2, and by the rSFI in Table 3, and compared with the TPC level of flare. Different clinical cases were assessed by each instrument, so the indices cannot be directly compared with each other.

For all 3 flare instruments, agreement on the level of flare as assessed by each instrument and by the TPC assessment occurred in a similar proportion of cases. The level of flare

Table 1. Number of cases by BILAG 2004 flare assessment and percent agreement with TPC flare assessment*

Assessment	No	Mild	Moderate	Severe	Total
No flare	27 (75)†	11 (31)	2 (7)	0	40 (27)
Mild flare	4 (11)	23 (64)†	6 (21)	2 (4)	35 (24)
Moderate flare	1 (3)	0	9 (31)†	4 (8)	14 (9)
Severe flare	2 (6)	0	11 (38)	41 (85)†	54 (36)
Insufficient information	2 (6)	2 (6)	1 (3)	1 (2)	6 (4)
Total	36	36	29	48	149

* Values are the number (%). BILAG = British Isles Lupus Assessment Group; TPC = 3-physician consensus.
† Statistically significant.

Table 2. SFI flare assessment and percent agreement with TPC flare assessment, and assessment excluding flare defined by treatment change*

Assessment	No	Mild/moderate (combined)	Severe	Total
SELENA				
No flare	17 (49)†	5 (7)	0	22 (15)
Mild/moderate flare	8 (23)	49 (67)†	0	57 (38)
Severe flare	9 (26)	19 (26)	44 (100)†	72 (47)
Insufficient information	1 (3)	0	0	1
Total	35	73	44	152
SELENA without flare defined by treatment change				
No flare	22 (63)†	10 (14)	0	32 (21)
Mild/moderate flare	8 (23)	53 (73)†	0	61 (40)
Severe flare	4 (11)	10 (14)	44 (100)†	58 (38)
Insufficient information	1 (3)	0	0	1
Total	35	73	44	152

* Values are the number (%). SFI = Safety of Estrogens in Lupus Erythematosus National Assessment/Systemic Lupus Erythematosus Disease Activity Index; TPC = 3-physician consensus.
† Statistically significant.

matched the TPC assessment of flare in the individual cases in a similar proportion of times using the original versions of these instruments: 67% for BILAG 2004 index, 72% for SFI, and 70% for rSFI. The corresponding weighted kappa coefficients for each instrument were BILAG 2004 index $\kappa = 0.82$, SFI $\kappa = 0.59$, and rSFI $\kappa = 0.74$.

An analysis of the discrepant cases where there was a difference in the assessment of flare between the TPC and the flare instrument was undertaken. There was a consistent pattern across all 3 instruments of a tendency to score moderate flares as severe, as well as scoring some of the no flare/persistent activity cases as flare when the SFI and rSFI instruments were used. Close examination of the data in Tables 2 and 3 suggested that this overscoring was driven by scoring a treatment change as flare, even if there were no new or worsening clinical features. Tables 2 and 3 also show the adjusted results. Only 1 of the 17 cases assessed by the SFI flare instrument as a flare, but as no flare by TPC, had increased lupus activity without treatment change, and in 2 cases the treatment change led to a higher level of flare than the clinical features alone would have scored. With the rSFI instrument, only 1 of 5 cases recorded as severe SELENA flare, but no flare by TPC, had clinical features of flare; the others were due to treatment change alone. In 1 case there was mild flare by TPC but severe SELENA flare due to treatment change. Excluding flare defined primarily by treatment change when assessing flares (Tables 2 and 3) using the SFI and rSFI led to an increase in the proportion of cases with agreement between the TPC and the instruments to 79% and 78%, respectively, and improved the weighted kappa scores to $\kappa = 0.82$ and $\kappa = 0.73$, respectively.

For the cases assessed by the BILAG 2004 index in Table 1, there were 2 cases with severe BILAG flare that

Table 3. Revised SELENA flare assessment and percent agreement with TPC flare assessment, and assessment excluding flare defined by treatment change*

Assessment	No	Mild	Moderate	Severe	Total
Revised SELENA flare index					
No flare	18 (58)†	0	0	0	18 (12)
Mild flare	4 (13)	16 (44)†	0	0	20 (13)
Moderate flare	4 (13)	19 (53)	19 (63)†	1 (2)	43 (29)
Severe flare	5 (16)	1 (3)	11 (37)	52 (98)†	69 (46)
Insufficient information	0	0	0	0	0
Total	31	36	30	53	150
Revised SELENA flare index without flare defined by treatment change					
No flare	23 (74)†	0	0	2 (4)	25 (17)
Mild flare	4 (13)	25 (69)†	5 (16.7)	1 (2)	35 (23)
Moderate flare	3 (10)	11 (31)	23 (77)†	3 (6)	40 (27)
Severe flare	1 (3)	0 (0)	2 (7)	47 (89)†	50 (33)
Insufficient information	0	0	0	0	0
Total	31	36	30	53	150

* Values are the number (%). SELENA = Safety of Estrogens in Lupus Erythematosus National Assessment; TPC = 3-physician consensus.
† Statistically significant.

were appropriately scored, and it was not clear why the TPC assessment was of no flare. For the 2 cases with severe flare by TPC assessed as mild by the BILAG assessors, review suggested that they had not been scored correctly and that the criteria for severe flare (BILAG A scores) were present in both cases. There were also 2 of 4 cases recorded by TPC as severe flare with a moderate flare reported by the BILAG assessors, but the criteria for severe flare were present. In 3 of 4 of these cases scored at a lower level by the BILAG assessors, neurologic items were underscored as B instead of A for reasons that are not clear. The patients with severe flares recorded by BILAG, but moderate flares by TPC, were correctly scored for BILAG in 9 of 11 cases and were mostly mucocutaneous or musculoskeletal flares (8 of 11 cases, with 2 cardiorespiratory flares and 1 renal flare).

DISCUSSION

This paper-patient exercise has allowed the assessment of a much wider range of clinical problems compared to the previous live-patient exercise (9). By using over 40 rheumatologists with a major interest in SLE, we were able to collect a large number of cases for flare assessment and involved a much larger number of assessors. The disposition of case review assignments was arranged to ensure that a similar range of cases was reviewed by all participating physicians. In the first part of the study, the 988 case histories were reviewed independently by 3 rheumatologists, and full agreement as to whether the patient was experiencing a mild, moderate, or severe flare or experiencing persistent/ongoing disease was achieved in 451 cases (46%). Review of discrepant cases suggested that there were some errors in the scoring of the BILAG 2004 index, without which the levels of agreement would have been higher, emphasizing the importance of training in the use of the glossary and the scoring manual. Although there was some evidence that single-system severe A-level flares by the BILAG 2004 index in mucocutaneous and musculoskeletal systems were sometimes perceived to be moderate-level flares by the TPC, the case descriptions varied in the amount of detail provided, and attaching much significance to the small number of cases this applied to is difficult, given that the presence of a significant flare was clear to all assessors.

The case histories with full agreement were then used, in effect, as our gold standard cases. The 3 lupus flare instruments, i.e., BILAG 2004, SFI, and rSFI, were used to assess flare in different cases in the second part of the study and performed equally well, particularly if treatment change was excluded as a component of the flare score. In practice, the scores from the pairs of rheumatologists who completed the BILAG 2004, SFI, or the rSFI demonstrated a high level of agreement overall with the predetermined TPC level of flare, and for the specific types of flare, especially for severe flare (levels of agreement of 85% or higher). Distinguishing mild and moderate flares from persistent activity/no flare and severe flare as assessed by the TPC method was more consistently achieved when the SFI and rSFI instruments were used without the rules that required all treatment change to determine flare. This issue has previously been reported in a smaller retrospective real patient cohort study (14). Apparently, treatment change by itself will quite frequently fail to indicate worsening disease, as adverse events, or treatment intolerance, may complicate matters. As a treat-to-target philosophy gains traction in the management of SLE, this philosophy increases the likelihood that treatment change will be widely used to define flare, when it may simply reflect fine tuning of managing persistent disease. Nevertheless an intent-to-treat principle remains useful to broadly reflect clinically important levels of disease activity.

It should be noted that 7 cases were not scored but were included in the analysis despite insufficient information, predominately those for BILAG flare assessment (6 of 7 cases), as this assessment is more demanding in terms of the information required for accurate and comprehensive completion of case report forms. Where there was initial discord in flare scoring for the other cases reported, the pairs reported that by discussion (usually via telephone conference) agreement could be reached relatively easily in the vast majority of cases, even though the quality of the case scenarios was rather variable. The pairs of assessors only used 1 instrument, and each of the 3 of types of flare instruments was not used on the same cases by the same people.

In 535 of the 988 cases (54%) originally submitted for flare assessment, consensus on the level of flare without using a flare instrument could not be agreed by 3 physicians. Although in some cases there was insufficient information for some of the physicians to rate the case, for example because the severity was not clear enough or the time of onset of the deterioration was not adequately defined (to distinguish flare from persistent activity or damage), there were other cases where consensus agreement could not be achieved despite reasonable scenarios. These were mostly those cases with multiple systems involved, presumably because different physicians ranked the components differently, particularly if some systems changed severity to different extents or changed inconsistently, with some symptoms worse and others stable or improved. This variability is reminiscent of a problem noted previously in a study evaluating responsiveness using BILAG and SLEDAI compared with a physician visual analog scale (15) and emphasizes the advantage in obtaining consistency of scoring in flare instruments, with defined glossaries rather than relying solely on physician opinion. However, limitations of disease definitions that rely on predefined thresholds of severity cannot be excluded.

Although the capacity of paper-patient exercises greatly increases the range of possible combinations of lupus clinical features for assessment in studies of this kind, nothing captures the dilemma of lupus assessment as much as a real patient in front of a clinician. Nevertheless the practicalities of getting large numbers of patients who are actually flaring into a clinical assessment study of the type that we reported previously (9) are considerable. In retrospect, we would probably have obtained greater agreement among the raters/assessors and a larger panel of cases for flare assessment with much tighter requirements/standardization for the writing of the case histories. An advantage of the case histories is that they were devised from real clinic patients. We are aware that there may have been biases in

that some of the assessors may have been more or less experienced in the assessment instruments used.

It is not just in SLE that challenges are evident in attempting to capture flare adequately and distinguish it from ongoing disease. The Outcome Measures in Rheumatology Rheumatoid Arthritis Flare Group has been involved in similar studies for several years (16). Their work is ongoing and the fact that they are dealing with a single organ/system highlights the added complexity when dealing with lupus.

Although we cannot say that the problem of capturing flare accurately in patients with lupus is ended, the relatively high levels of agreement obtained with the flare instruments and weighted kappa coefficients, ranging from 0.59 (SFI instrument with treatment) to 0.82 (for rSFI without treatment and BILAG 2004), was encouraging, especially given some problems with inadequate histories and the great diversity of rheumatologists from many countries involved in the study. All of the instruments used in this study have construct and content validity based on their use with these paper-patient scenarios. The choice of instrument to be used in future flare studies will depend on the types of patients to be assessed, the need to distinguish different types of flares, and the training and experience of the likely investigators.

ACKNOWLEDGMENTS

The authors thank the many colleagues who helped us to perform the study, in particular Dr. M. Aringer (Dresden), Dr. S. Chambers (Cayman Islands), Dr. Costedoat-Chalumeau (Paris), Dr. S. Croca (Lisbon), Dr. I. Giles (London), Dr. B. Griffiths (Newcastle), Dr. P. Lanyon (Nottingham), Dr. E. Moreland (Melbourne), Dr. M. Mosca (Pisa), Dr. J. M. Pego-Reigosa (Vigo), Dr. M. Schneider (Dusseldorf), and Dr. C. Vasconcelos (Porto).

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. Isenberg 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. Isenberg, Sturgess, Allen, Romero-Diaz, Merrill, Gordon.

Acquisition of data. Isenberg, Aranow, Askanase, Sang-Cheol, Bernatsky, Bruce, Buyon, Cervera, Clarke, Dooley, Fortin, Ginzler, Gladman, Hanly, Inanc, Jacobsen, Kamen, Khamashta, Lim, Manzi, Nived, Peschken, Petri, Kalunian, Rahman, Ramsey-Goldman, Romero-Diaz, Ruiz-Irastorza, Sanchez-Guerrero, Steinsson, Sturfelt, Urowitz, van Vollenhoven, Wallace, Zoma, Merrill, Gordon.

Analysis and interpretation of data. Isenberg, Sturgess, Allen, Merrill, Gordon.

REFERENCES

1. Gladman DD, Ibanez D, Urowitz MB. Systemic lupus erythematosus activity index 2000. J Rheumatol 2000;29:288–91.
2. Yee CS, Farewell V, Isenberg DA, Griffiths B, Teh LS, Bruce IN, et al. The BILAG 2004 index is sensitive to change for assessment of SLE disease activity. Rheumatology (Oxford) 2009;48:691–5.
3. Nuttall A, Isenberg DA. Assessment of disease activity damage and quality of life in systemic lupus erythematosus. Best Pract Res Clin Rheumatol 2013;27:300–16.
4. Merrill JT, Neuwett CM, Wallace DJ, Shanahan JC, Latinis KM, Oates JC, et al. Efficacy and safety of rituximab in moderately-to-severely active systemic lupus erythematosus: the randomized, double-blind, phase II/III systemic lupus erythematosus evaluation of rituximab trial. Arthritis Rheum 2010;62:222–33.
5. Furie R, Petri M, Zamani O, Cervera R, Wallace DJ, Tegzova D, 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.
6. Isenberg D, Gordon C, Licu D, Copt S, Rossi CP, Wofsy D. Efficacy and safety of atacicept for prevention of flares in patients with moderate-to-severe systemic lupus erythematosus (SLE): 52-week data (APRIL-SLE randomised trial). Ann Rheum Dis 2015;74:2006–15.
7. Yee CS, Farewell V, Isenberg DA, Griffiths B, Teh LS, Bruce In, et al. The use of the 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.
8. Ruperto N, Hanrahan LM, Alarcon GS, Belmont HM, Brey RL, Brunetta P, et al. International consensus for a definition of disease flare in lupus. Lupus 2011;20:453–62.
9. Isenberg DA, Allen E, Farewell V, D'Cruz D, Alarcon GS, Aranow C, et al. An assessment of disease flare in patients with systemic lupus erythematosus: a comparison of BILAG-2004 and the flare version of SELENA. Am Rheum Dis 2011; 70:54–9.
10. Hochberg MC, for the Diagnostic and Therapeutic Criteria Committee of the American College of Rheumatology. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus [letter]. Arthritis Rheum 1997;40:1725.
11. Petri M, Orbai AM, Alarcon GS, Gordon C, Merrill J, Fortin PR, et al. Derivation and validation of the Systemic Lupus International Collaborating Clinics classification criteria for systemic lupus erythematosus. Arthritis Rheum 2012;64:2677–86.
12. Yee CS, Cresswell L, Farewell V, Rahman A, Teh LS, Griffiths B, et al. Numerical scoring for the BILAG-2004 index. Rheumatology (Oxford) 2010;49:1665–9.
13. Buyon JP, Petri MA, Kim MY, Kalunian KC, Grossman J, Hahn BH, et al. The effect of combined estrogen and progesterone hormone replacement therapy on disease activity in systemic lupus erythematosus: a randomized trial. Ann Intern Med 2005;142 Pt. 1:953–62.
14. Thanou A, Chakravarty E, James JA, Merrill JT. How should lupus flares be measured? Deconstruction of the safety of estrogen in lupus erythematosus national assessment Systemic Lupus Erythematosus Disease Activity Index flare index. Rheumatology (Oxford) 2014;53:2175–81.
15. Wollaston SJ, Farewell JT, Isenberg DA, Gordon C, Merrill JT, Petri MA, et al. Defining response in systemic lupus erythematosus: a study by the Systemic Lupus Erythematosus International Collaborating Clinic Group. J Rheumatol 2004; 3:2390–4.
16. Bartlett SJ, Bykerk VP, Cooksey R, Choy EH, Alten R, Christensen R, et al. Feasibility and domain validation of rheumatoid arthritis (RA) flare core domain set: report of the OMERACT 2014 RA Flare Group Plenary. J Rheumatol 2015; 42:2185–9.