Systemic Lupus Erythematosus Disease Activity Index 2000 Responder Index-50: A Reliable Index for Measuring Improvement in Disease Activity

ZAHI TOUMA, MURRAY B. UROWITZ, PAUL R. FORTIN, CAROLINA LANDOLT, SERGIO M. TOLOZA, CLAIRE RIDDELL, VINOD CHANDRAN, LIHI EDER, AQEEL GHANEM, OLGA ZIOUZINA, SHAHRZAD TAGHAVI-ZADEH, DOMINIQUE IBAÑEZ, and DAFNA D. GLADMAN

ABSTRACT. ***Objective.*** To test the interrater and intrarater reliability of the Systemic Lupus Erythematosus Disease Activity Index 2000 (SLEDAI-2K) Responder Index (SRI-50), an index designed to measure ≥ 50% improvement in disease activity between visits in patients with systemic lupus erythematosus.

Methods. This was a multicenter, cross-sectional study with raters from Canada, the United Kingdom, and Argentina. Patient profile scenarios were derived from real adult patients. Ten rheumatologists from university and community hospitals and postdoctoral rheumatology fellows participated. An SRI-50 data retrieval form was used. Each rheumatologist scored SLEDAI-2K at the baseline visit and SRI-50 on followup visit, for the same patients, on 2 occasions 2 weeks apart. Physician global assessment (PGA) was determined on a numerical scale at baseline visit and a Likert scale on followup visit. Interrater and intrarater reliability was assessed using intraclass correlation coefficient (ICC) and kappa statistics whenever applicable.

Results. Forty patient profiles were created. The ICC performed on 80 patient profiles for interrater ranged from 1.00 for SLEDAI-2K and SRI-50 to 0.96 for PGA. The intrarater ICC for SLEDAI-2K, SRI-50, and PGA scores ranged from 1.00 to 0.86. Substantial agreement was determined for the interrater Likert scale, with a kappa statistic of 0.57.

Conclusion. The SRI-50 is reliable to assess ≥ 50% improvement in lupus disease activity. Use of the SRI-50 data retrieval form is essential to ensure optimal performance of the SRI-50. SRI-50 can be used by both rheumatologists and trainees and performs equally well in trained as well as untrained rheumatologists. (J Rheumatol First Release Feb 15 2011; doi:10.3899/jrheum.101080)

Key Indexing Terms:
SYSTEMIC LUPUS ERYTHEMATOSUS DISEASE ACTIVITY AND RESPONDER INDEX
SRI-50 SLEDAI-2K RELIABILITY

Systemic lupus erythematosus (SLE) is a complex disease with highly variable patterns of organ involvement and prognosis[1]. During the course of their disease, patients with lupus experience events that are related to acute disease activity or to chronic damage, which makes the disease difficult to monitor[1]. Lupus disease activity is an important

From the Toronto Western Hospital, University Health Network, Toronto, Ontario, Canada; Hospital San Juan Bautista, Catamarca, Argentina; and Antrim Area Hospital, Northern Ireland, UK.

Dr. Touma is a recipient of the Lupus Ontario Geoff Carr Fellowship and the University of Toronto Arthritis Centre of Excellence Fellowship. The Lupus Clinic is supported by The Lupus Flare Foundation, Arthritis and Autoimmune Centre Foundation, Toronto General-Toronto Western Hospital Foundation, and the Smythe Foundation.

Z. Touma, MD, University of Toronto Lupus Clinic, Centre for Prognosis Studies in the Rheumatic Diseases, Toronto Western Hospital; M.B. Urowitz, MD, FRCPC, Professor of Medicine, University of Toronto, Senior Scientist, Toronto Western Research Institute, Director, University of Toronto Lupus Clinic, Centre for Prognosis Studies in the Rheumatic Diseases, Toronto Western Hospital; P.R. Fortin, MD, FRCPC, MPH, Professor of Medicine, University of Toronto, Senior Scientist, Toronto Western Research Institute, Director, University of Toronto Lupus Clinic, Centre for Prognosis Studies in the Rheumatic Diseases, Toronto Western Hospital; C. Landolt, MD, FRCP, Assistant Professor of Medicine, University of Toronto, Senior Scientist, Toronto Western Research Institute, Director, University of Toronto Lupus Clinic, Centre for Prognosis Studies in the Rheumatic Diseases, Toronto Western Hospital; S.M. Toloza, MD, Hospital San Juan Bautista; C. Riddell, MD, Antrim Area Hospital; V. Chandran, MBBS, MD, DM, Assistant Professor of Medicine, University of Toronto, Senior Scientist, Toronto Western Research Institute, Director, University of Toronto Lupus Clinic, Centre for Prognosis Studies in the Rheumatic Diseases, Toronto Western Hospital; L. Eder, MD, University of Toronto Lupus Clinic, Centre for Prognosis Studies in the Rheumatic Diseases, Toronto Western Hospital; A. Ghanem, MD, University of Toronto Lupus Clinic, Centre for Prognosis Studies in the Rheumatic Diseases, Toronto Western Hospital; O. Ziouzina, MD, FRCPC, University of Toronto Lupus Clinic, Centre for Prognosis Studies in the Rheumatic Diseases, Toronto Western Hospital; S. Taghavi-Zadeh, Research Analyst, University of Toronto Lupus Clinic, Centre for Prognosis Studies in the Rheumatic Diseases, Toronto Western Hospital; D. Ibañez, MSc, University of Toronto Lupus Clinic, Centre for Prognosis Studies in the Rheumatic Diseases, Toronto Western Hospital; D.D. Gladman, MD, FRCPC, Professor of Medicine, University of Toronto, Senior Scientist, Toronto Western Research Institute, Co-Director, University of Toronto Lupus Clinic, Centre for Prognosis Studies in the Rheumatic Diseases, Toronto Western Hospital.

Address correspondence to Dr. Gladman, Centre for Prognosis Studies in the Rheumatic Diseases, Toronto Western Hospital, Room 1E-410B, 399 Bathurst Street, Toronto, Ontario M5T 2S8, Canada. E-mail: dafna.gladman@utoronto.ca

Accepted for publication December 29, 2010.

domain that must be assessed in clinical trials and outcome studies. Other domains, namely damage resulting from lupus activity or its therapy, health-related quality of life, adverse events, and economic costs including health utilities, are utilized to adequately describe the effects of the disease[1,2]. It is essential that measures used to monitor such outcomes have evidence of validity and reliability[3]. The Systemic Lupus Erythematosus Disease Activity Index 2000 (SLEDAI-2K) Responder Index (SRI-50) is a valid index able to demonstrate incomplete but clinically significant ≥ 50% improvement in disease activity in lupus patients[4].

The SRI-50 comprises the same 24 descriptors, covering 9 organ systems, and reflects disease activity over the previous 30 days as does SLEDAI-2K[4,5,6,7,8]. The SRI-50 data retrieval form standardizes the documentation of the descriptors and performed extremely well in all descriptors, which is especially relevant for multicenter studies that form the backbone of any therapeutic evaluation for SLE[4]. The practical applicability of the SRI-50, including ease of administration, low costs of data collection, method of scoring and ease of score interpretation, and construct validity, has been demonstrated[4].

Clinicians seeking a tool to measure disease activity should look for evidence of reliability, e.g., stability of a tool when no change has occurred in disease activity, test-retest or intrarater reliability, and within-rater reliability or interrater reliability[3,9,10].

Our study assessed the interrater and intrarater reliability of the SRI-50 in patient profile scenarios derived from real adult lupus patients with the participation of rheumatologists from different centers in different countries.

MATERIALS AND METHODS

Patient selection. This study was performed on patient profile scenarios derived from a longitudinal cohort of lupus patients receiving followup care at a single center. All patients in the cohort are followed longitudinally and met the American College of Rheumatology (ACR) classification criteria for SLE[11,12]. Patients attend the lupus clinic at 2–6 month intervals regardless of the state of activity of their lupus. Patients are assessed using a standard protocol that includes complete history, physical examination, and laboratory evaluation. Collection and storage of data at the lupus clinic are conducted in accord with the Declaration of Helsinki and is approved by the Research Ethics Board of the University Health Network, Toronto, Canada. Signed informed consent is obtained from patients at the time of enrollment into the cohort at the lupus clinic.

The sampling strategy adopted in this study to evaluate the reliability of SRI-50 assured that each of the 24 descriptors of SLEDAI-2K was represented in at least 1 patient profile[6]. After selecting the patients that would be included in the study, 40 patient profiles were created based on the information available for the selected visit. Each patient profile was composed of an initial visit and a followup visit. The patient profile was based on the patient's subjective complaints and the objective findings of the clinical, laboratory, and radiological assessments. This was based on the data available from the lupus clinic database, from the medical chart, and from the electronic medical record. On followup visits, there were patients who either had improvement in all active systems as compared to baseline visit, or had improvement in one system and/or worsening in another. This gave the raters the possibility to determine if there had been improvement in the descriptors.

Assessment of disease activity. SLEDAI-2K 30 days. Disease activity was measured by the SLEDAI-2K, a valid measure of disease activity in SLE[6,7], at the first visit. SLEDAI-2K was modeled on clinicians' global judgment to standardize and measure disease activity. SLEDAI-2K is based on the presence of 24 descriptors in 9 organ systems over the patient's past 10 days. SLEDAI-2K 30 days was validated against SLEDAI-2K 10 days to describe disease activity over the previous 30 days[7,8]. The total score of SLEDAI-2K falls between 0 and 105, with higher scores representing increased disease activity[6].

SRI-50. The SRI-50 is a responder index based on the SLEDAI-2K 30 days that describes partial improvement ≥ 50% in disease activity between visits in lupus patients[4]. SRI-50 score is evaluated at the followup visit and corresponds to the sum of each of the 24 descriptor scores on the SRI-50 data retrieval form. The method of scoring is simple, cumulative, and intuitive and similar to the SLEDAI-2K. One of 3 situations can result when a descriptor is present at the initial visit: (1) the descriptor has achieved complete remission at followup, in which case the score would be "0"; (2) the descriptor has not achieved a minimum of 50% improvement at followup, in which case the score would be identical to its corresponding SLEDAI-2K value; or (3) the descriptor has improved by ≥ 50% (according to the SRI-50 definition) but has not achieved complete remission, in which case the score is evaluated as one-half the score that would be assigned for SLEDAI-2K. If a descriptor was not present at the initial visit, the value for the SRI-50 at the followup visit will be the same as that for SLEDAI-2K. This process is repeated for each of the 24 descriptors. Finally, the SRI-50 score at followup is evaluated as the sum of the scores of the 24 individual descriptors[4].

Physician global assessment. Physician global assessment (PGA) was determined initially at baseline assessment on a 100-mm visual analog scale (VAS; 0 = no disease activity, 100 = very active disease). Physicians documented the PGA based on the baseline assessment of the patient.

Likert scale. During the followup visit a physician response assessment was determined on a 7-point Likert scale (LS), where 7 = much improved, 6 = moderately improved, 5 = slightly improved, 4 = unchanged, 3 = slightly worse, 2 = moderately worse, and 1 = much worse. We defined a 50% improvement as LS ≥ 6. Raters were instructed to circle the appropriate number on the LS to indicate how active the patient's lupus disease activity was on followup visit. The use of numerical scales in the assessment of global disease activity of lupus and rheumatoid arthritis has been adopted in several studies[13,14].

"Standard" SLEDAI-2K and SRI-50 scores. "Standard" SLEDAI-2K and SRI-50 scores were established by the creator of the scenarios (ZT), who described each of the clinical and laboratory variables, and who did not participate in the study as an assessor. The evaluation of raters' scores of SLEDAI-2K and SRI-50 was compared to the "Standard" SLEDAI-2K and SRI-50 results.

Raters, site selection, and procedure at each site. Ten rheumatologists who represented university and community hospitals from 3 centers in different countries, Canada, United Kingdom, and Argentina, participated in this study. All had worked at or had trained at the University of Toronto Lupus Clinic and were comfortable with the use of the original SLEDAI-2K. Four rheumatologists were from university hospitals and 2 from community hospitals, and 4 were postdoctoral rheumatology fellows. The level of training among rheumatologists in the use of the SRI-50 in the reliability study differed. This approach allowed us to evaluate the performance of the SRI-50 among trainees and rheumatologists. Patient profiles were sent to each rater in 2 separate packages, each containing 20 cases. The same 40 equivalent patient profiles were sent again in 2 packages to the same 10 rheumatologists after 2 weeks from the first occasion to complete the SRI-50 data retrieval form, along with LS. This approach was adopted to reduce the possibility of true clinician recall[9]. These patient profiles were returned to the coordinating center after completion, for evaluation and comparison to the "Standard" scores, by one external assessor (ZT).

Statistical analysis. Descriptive statistics were used to describe the characteristics of the patients. We evaluated the number of mis-scorings in each round, and in both rounds for all raters for the SLEDAI-2K 30 days and the SRI-50.

We determined the interrater intraclass correlation coefficient (ICC) for SLEDAI-2K, SRI-50, and PGA. The intrarater ICC were evaluated for each rater separately for SLEDAI-2K, SRI-50, and PGA. Specifically, in all the above analyses we determined both ICC (2,1) and ICC (2,k). The first number "2" designates the model and is used when all subjects are rated by the same raters, who are assumed to be a random subset of all possible raters[15]. The second number signifies the form, using either a single measurement "1" ICC (2,1) or the mean of several measurements "k" ICC (2,k) as the unit of analysis in the model. The mean scores have the effect of increasing reliability estimates, as means are considered better estimates of true scores, theoretically reducing error variance[15,16,17]. As suggested by Streiner and Norman[9] we considered ICC ≥ 0.85 to reflect good reliability. We determined the average intrarater ICC for SLEDAI-2K, SRI-50, and PGA[9,18].

We transformed the data available on 80 patient profiles for SLEDAI-2K and SRI-50 as categorical data, "yes" for right score and "no" for wrong score, compared to the "Standard" solutions. We evaluated the number and percentage of right answers for both SLEDAI-2K and SRI-50 scores as compared to the "Standard" SLEDAI-2K and SRI-50 solutions, respectively. We applied paired t tests and compared the mean SLEDAI-2K and SRI-50 scores from both rounds. P values ≤ 0.05 were considered significant.

We determined the interrater kappa for LS scores. According to Landis and Koch[19], agreement indexes were interpreted as follows: 0.81–1.00 = almost perfect, 0.61–0.75 = substantial agreement, 0.41–0.60 = moderate agreement, 0.21–0.40 = fair agreement, 0–0.20 = slight agreement, and ≤ 0 = poor agreement.

Sample size calculation. Sample size determined in this study was based on 3 estimates: reliability estimate, number of raters, and the confidence interval[9]. The sample size sufficient for an ICC of 0.80, a standard error of 0.05, and 10 raters is 31 patient profiles. Oversampling of 9 scenarios was done to allow for incomplete forms. Generally, samples of 40–50 are sufficient, and "going above 50 subjects in many situations is probably statistical overkill" (Streiner and Norman[9]). Indeed, the methodology adopted in our study to evaluate the intrarater reliability allowed us to double this number to 80 profiles. An ICC ≥ 0.75 is suggestive of good reliability and those below 0.75 poor to moderate reliability. For many clinical measurements, reliability should exceed 0.90 to ensure reasonable validity[16].

RESULTS

Patient demographic data. The patient profiles included 35 females and 5 males; 55% were Caucasian, 22% Black, 5% Asian, and 18% others. Age at diagnosis was 30.4 ± 12.7 years, age at the study date was 38.0 ± 13.5 years, and disease duration at study date was 7.6 ± 8.1 years. The mean SLEDAI-2K score at baseline visit was 11.90 ± 7.09 and the mean SRI-50 on followup visit was 5.98 ± 3.40[4,7]. The Systemic Lupus International Collaborating Clinics/ACR Damage Index (SDI) was 1.05 ± 1.45[20]. As described above the sampling strategy we adopted assured that each of the 24 descriptors of SLEDAI-2K was represented in at least 1 patient profile (Table 1).

Table 1. Distribution of clinical and laboratory descriptors of the SRI-50 in 40 patient profile scenarios.

Characteristic	N (%)
Seizure	1 (2.5)
Psychosis	3 (7.5)
Organic brain	3 (7.5)
Visual	3 (7.5)
Cranial nerve	2 (5.0)
Lupus headache	2 (5.0)
Cardiovascular accident	2 (5.0)
Vasculitis	7 (17.5)
Arthritis	11 (27.5)
Myositis	1 (2.5)
Casts	5 (12.5)
Hematuria	4 (10)
Proteinuria	6 (15)
Pyuria	5 (12.5)
Rash	16 (40)
Alopecia	8 (20)
Mucosal ulcers	3 (7.5)
Pleurisy	3 (7.5)
Pericarditis	1 (2.5)
Low complement	13 (32.5)
Increased anti-DNA antibody levels	18 (45)
Fever	1 (2.5)
Thrombocytopenia	1 (2.5)
Leukopenia	2 (5.0)

Common pitfalls. For SLEDAI-2K scoring, a total of 3 mis-scorings were found in the clinical descriptors compared to 27 in the laboratory descriptors in both rounds. For SRI-50 scoring, 12 mis-scorings were found in the clinical descriptors compared to 48 in the laboratory descriptors in both rounds. The mis-scorings were the result of the rater's failure to identify the appropriate relevant data available in the patient profile scenario or the wrong application (misunderstanding and unawareness) of the SLEDAI-2K or SRI-50 definitions. The most common pitfalls by raters in SLEDAI-2K scoring in both rounds were related to the 2 descriptors "casts" and "leukopenia." In scoring the SRI-50, the most common mis-scorings were related to complement, casts, pyuria, and leukopenia, and to a lesser extent to rash and fever. Almost all mis-scorings that were related to casts originated from one rater, who did not translate the number of casts from the case scenarios to the data retrieval form of the SRI-50. This resulted in wrong scoring in both SLEDAI-2K and SRI-50. The mis-scorings related to the complements were present only in the followup visit. This was related to mathematical miscalculation when determining whether there is a 50% improvement by the raters. Thus virtually all the mis-scorings were rater failures rather than instrument failures (Table 2).

Reliability (interrater and intrarater). Table 3 lists the interrater reliability and the corresponding ICC (2,1) and ICC (2,k) values for each round separately and for all 80 patient profiles for SLEDAI-2K, SRI-50, and PGA. The ICC (2,k) performed on 80 patient profiles for interrater ranged from 1.00 for SLEDAI-2K and SRI-50 to 0.96 for PGA. The average intrarater ICC for SLEDAI-2K, SRI-50, and PGA were 0.99, 0.98, and 0.90, respectively[18].

Table 4 lists the intrarater reliability and the corresponding ICC (2,1) and ICC (2,k) for each rater separately for SLEDAI-2K, SRI-50 and PGA. The ICC (2,k) for

Table 2. Raters mis-scorings in rounds 1 and 2 in SLEDAI-2K/SRI-50.

Index	Descriptors	Round 1 (n = 400)*	Round 2 (n = 400)*	Total/Descriptor (n = 800)*
SLEDAI-2K	CD	0	3	3
	LD	11	16	27
	Total/round	11	19	30
SRI-50	CD	3	9	12
	LD	20	28	48
	Total/round	23	37	60

* Number of patients' scenarios. CD: clinical descriptors; LD: laboratory descriptors.

Table 3. Interrater reliability ICC (2,1) and ICC (2,k).

	ICC (2,1)	ICC (2,k)
	Round 1 + Round 2 (n = 800)	
SLEDAI-2K	0.94	1.00
SRI-50	0.99	1.00
PGA	0.69	0.96
	Round 1 (n = 400)	
SLEDAI-2K	0.99	1.00
SRI-50	0.97	1.00
PGA	0.63	0.94
	Round 2 (n = 400)	
SLEDAI-2K	0.99	1.00
SRI-50	0.98	1.00
PGA	0.60	0.94

SLEDAI-2K: Systemic Lupus Erythematosus-2000; SRI-50: SLEDAI-2K Responder Index-50; PGA: physician global assessment.

SLEDAI-2K and SRI-50 ranged from 0.97 to 1.00 among raters[9]. The PGA ICC (2,k) ranged from 0.86 to 1.00.

Categorical data for the SLEDAI-2K and SRI-50 are presented in Table 4. Of 400 patient profiles that were completed by 10 raters, 374 (93.5%) and 346 (86.5%) were concordant with the "Standard" results of SLEDAI-2K and SRI-50, respectively. The mean SLEDAI-2K scores were 11.83 ± 7.02, 11.83 ± 7.04, and 11.90 ± 7.09 in round 1 and round 2 and as per the "Standard," respectively. There was no statistically significant difference between round 1 compared to round 2 ($p = 0.82$). However, round 1 versus "Standard" (0.07 ± 0.71; $p = 0.05$) and round 2 versus "Standard" (0.08 ± 0.67; $p = 0.020$) showed results that were either statistically significant or borderline significant, but the actual differences from the "Standard" were not clinically significant.

The mean SRI-50 scores were 5.93 ± 3.34, 5.89 ± 3.33, and 5.98 ± 3.40 in round 1 and round 2 and per the "Standard," respectively. There was no statistically significant difference between round 1 versus round 2 ($p = 0.28$) or round 1 versus the "Standard" ($p = 0.12$). There was a statistically significant difference between round 2 compared to "Standard" of 0.08 ± 0.46 ($p = 0.02$), but this was not clinically significant. Substantial agreement was determined for interrater LS scores, with a kappa statistic of 0.57 (95% CI 0.49–0.66)[9,19].

DISCUSSION

Prior to use in clinical research or clinical practice, a health status measurement tool should be valid, reliable, and responsive for its intended use in its intended population[21]. We previously demonstrated that the SRI-50 is valid and is able to measure ≥ 50% improvement in disease activity of patients with lupus between visits[4]. In this study we have demonstrated that SRI-50 is reliable.

In our study, we evaluated both inter- and intraobserver reliability. To determine intrarater reliability, the rheumatologists reevaluated the same patient scenarios on 2 occasions, 14 days apart[9]. We developed patient scenarios to assure that all the descriptors were present, including some relatively rare manifestations of lupus. We used the valid standardized SRI-50 data retrieval form to help minimize other sources of variability[4].

The use of patient profile scenarios as compared to live case scenarios has been reported. Case scenarios were previously adopted in the initial development and validation of the SLEDAI, the SDI, and the ACR response criteria for SLE clinical trials[5,20,22]. A recent study showed that the use of paper case scenarios to determine the interrater reliability of triage scales in the emergency department is an efficient method that approximates that of live cases. Further, the authors concluded that if the results are found to be within an acceptable performance range, further testing of interrater reliability using live cases may be unnecessary[23]. In our study, the results of the ICC for SRI-50 exceeded 0.90, ensuring reasonable reliability[16].

For test-retest and interrater reliability, indexes of agreement are required as opposed to tests of association. The ICC deals with continuous data and is sensitive to systematic biases between observers or administration times and, more importantly, it is sensitive to both association and agreement[9,16,21,24]. The kappa statistic deals better with categorical data. In this study, we adopted the ICC in determining the reliability of SRI-50, SLEDAI-2K, and PGA and the kappa statistic in determining the reliability of the LS results. We observed high ICC for interrater and intrarater, confirming the reliability of SRI-50 along with SLEDAI-2K. These findings are in agreement with studies that also have demonstrated that the original SLEDAI and its updated version, SLEDAI-2K, are reliable indices[25,26,27,28,29]. Further, when we converted the results of SLEDAI-2K and SRI-50 into categorical data, we found no clinically significant difference compared with the "Standard." Our study thus provides evidence that rheumatologists from different centers and different countries are able to assess disease activity by SRI-50 along with SLEDAI-2K 30 days in a particular patient in a similar way. This information is useful for collaborative studies of patients with SLE that include the assessment of disease activity.

Table 4. Intrarater reliability and the corresponding ICC (2,1) and ICC (2,k) for each rater separately for SLEDAI-2K, SRI-50, and PGA, and the categorical data for SLEDAI-2K and SRI-50.

Raters		Continuous Data		Categorical Data*, Number of Profiles (%)	
	SLEDAI-2K	SRI-50	PGA	SLEDAI-2K = Standard SLEDAI-2K	SRI-50 = Standard SRI-50
1 ICC (2,1)	0.98	0.97	0.75	78 (97.5)	73 (91.3)
ICC (2,k)	0.99	0.99	0.86		
2 ICC (2,1)	1.00	1.00	0.85	76 (95.0)	73 (91.3)
ICC (2,k)	1.00	1.00	0.92		
3 ICC (2,1)	0.99	0.97	0.86	68 (85.0)	66 (82.5)
ICC (2,k)	1.00	0.98	0.93		
4 ICC (2,1)	0.99	0.96	0.91	77 (96.3)	73 (91.3)
ICC (2,k)	0.99	0.98	0.95		
5 ICC (2,1)	1.00	1.00	0.84	79 (98.8)	77 (96.3)
ICC (2,k)	1.00	1.00	0.91		
6 ICC (2,1)	1.00	0.99	0.88	77 (96.3)	74 (92.5)
ICC (2,k)	1.00	0.99	0.94		
7 ICC (2,1)	1.00	1.00	0.90	78 (97.5)	73 (91.3)
ICC (2,k)	1.00	1.00	0.95		
8 ICC (2,1)	1.00	0.94	0.67	78 (97.5)	76 (95.0)
ICC (2,k)	1.00	0.97	0.81		
9 ICC (2,1)	1.00	1.00	0.99	78 (97..5)	77 (96.3)
ICC (2,k)	1.00	1.00	0.99		
10 ICC (2,1)	1.00	1.00	0.99	80 (100)	66 (82.5)
ICC (2,k)	1.00	1.00	1.00		
Total no. of equal scores				374 (93.5)	346 (86.5)

* Number (percentage) of right answers for both SLEDAI-2K and SRI-50 scores versus the "Standard" SLEDAI-2K and SRI-50 solutions, respectively.

Model 2 of the ICC (2,1) was adopted in our study. This model partitions the total variance into effects due to differences between subjects, differences between raters, and error variance[16]. In this model, patients are evaluated by the same raters, and these raters are considered representative of a large population of similar raters. More important, we chose this model for our study because we were interested in establishing the SRI-50 intrarater and interrater reliability and documenting that SRI-50 has a broad application[16]. Our results confirmed that the SRI-50 can be used with confidence and equally by all rheumatologists despite heterogeneity in the level of training[16].

Guidelines for acceptable ICC values vary. Streiner and Norman suggest that a tool with good reliability when studying groups of people should have an ICC exceeding 0.85, and Tammemagi, *et al* lower the cutoff value to > 0.75 to be acceptable[30,31]. McHorney and Tarlov, among others, required to have a coefficient > 0.90 when interpreting individual data rather than group data[32]. In our study, the test-retest and intrarater coefficients exceeded 0.9. The raters' recall bias for test-retest reliability was eliminated with the methodology adopted in our study, where patients were reevaluated after at least 14 days[9]. The reliability for PGA exceeded 0.9 and LS scores showed substantial agreement for interrater LS scores with kappa statistics.

Several factors can improve the reliability of a measurement and, to improve the reliability of SRI-50, we intended to ensure the presence of the following factors: (1) using more clearly written descriptors with universally understood words; and this was confirmed to be present in both SRI-50 definitions and SRI-50 data retrieval forms[4]; (2) selecting clear detailed definitions to cover all the aspects within each descriptor; and (3) using categorical and numerical rating scales in each of the descriptors, whenever applicable, instead of dichotomous response choices. As examples, numerical scales are used to determine if there is an improvement in headache, pleurisy, cranial nerve disorder, alopecia, pericarditis; and categorical scales to determine the improvement in myositis, alopecia, and rash[3].

Overall, the performance of the SRI-50 was excellent, despite the mis-scorings that occurred during this study. Virtually all the mis-scorings were rater failures rather than instrument failures. Indeed, the mis-scorings that resulted from the scoring of the laboratory descriptors and the calculation of the 50% improvement could be avoided by more accurate readings of the cases. It is very important that all rheumatologists familiarize themselves with the definitions of SLEDAI-2K initially and then learn the SRI-50 to ensure better performance. In research centers and clinical trials, the laboratory data that include lupus serology (complements and anti-dsDNA), white blood cell counts and platelets, and urinalysis variables are entered and analyzed systematically in the database after being reviewed by rheumatologists. The review by rheumatologists is not just for the purpose of patient safety; it is also to assess whether abnormalities are due to SLE and in some cases (such as drug toxicities) might override scoring of some of these on the SLEDAI. Using the SRI-50 data retrieval form would help to minimize mistakes when transferring the data from laboratory reports.

The training of all rheumatologists to accomplish this task is crucial. An SRI-50 manual has been developed for this purpose, along with an electronic version of the SRI-50. The dedicated website for SRI-50 is under construction at this time. This will include training and examination modules, after which certification will be granted for successful completion of the examination module.

Our study shows that the SRI-50 is reliable in detecting ≥ 50% improvement in disease activity between visits in patients with lupus[4]. Thus SRI-50 can be adopted as a responder index in clinical and research settings and in clinical trials.

REFERENCES

1. Gladman DD. Indicators of disease activity, prognosis, and treatment of systemic lupus erythematosus. Curr Opin Rheumatol 1994;6:487-92.
2. Strand V, Gladman D, Isenberg D, Petri M, Smolen J, Tugwell P. Outcome measures to be used in clinical trials in systemic lupus erythematosus. J Rheumatol 1999;26:490-7.
3. McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993;31:247-63.
4. Touma Z, Gladman DD, Ibañez D, Urowitz MB. Development and initial validation of SLEDAI-2K (Systemic Lupus Erythematosus Disease Activity Index 2000) Responder Index-50 (SRI-50). J Rheumatol 2011;38;275-84.
5. Bombardier C, Gladman DD, Urowitz MB, Caron D, Chang CH. Derivation of the SLEDAI. A disease activity index for lupus patients. The Committee on Prognosis Studies in SLE. Arthritis Rheum 1992;35:630-40.
6. Gladman DD, Ibanez D, Urowitz MB. Systemic Lupus Erythematosus Disease Activity Index 2000. J Rheumatol 2002;29:288-91.
7. Touma Z, Urowitz MB, Gladman DD. SLEDAI-2K for a 30-day window. Lupus 2010;19:49-51.
8. Touma Z, Urowitz MB, Ibañez D, Gladman DD. SLEDAI-2K 10 days versus SLEDAI-2K 30 days in a longitudinal evaluation. Lupus 2010;20:67-70.
9. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. In: Streiner DL, Norman GR, editors. Reliability, generalizability theory and validity. New York: Oxford University Press; 2008:167-276.
10. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials 1991;12 Suppl:142-58.
11. Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. Arthritis Rheum 1997;40:1725.
12. Tan EM, Cohen AS, Fries JF, Masi AT, McShane DJ, Rothfield NF, et al. The 1982 revised criteria for the classification of systemic lupus erythematosus. Arthritis Rheum 1982;25:1271-7.
13. Wollaston SJ, Farewell VT, Isenberg DA, Gordon C, Merrill JT, Petri MA, et al. Defining response in systemic lupus erythematosus: a study by the Systemic Lupus International Collaborating Clinics group. J Rheumatol 2004;31:2390-4.
14. Wolfe F, Michaud K, Pincus T, Furst D, Keystone E. The Disease Activity Score is not suitable as the sole criterion for initiation and evaluation of anti-tumor necrosis factor therapy in the clinic: Discordance between assessment measures and limitations in questionnaire use for regulatory purposes. Arthritis Rheum 2005;52:3873-9.
15. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-8.
16. Portney LG, Watkins MP. Statistical measures of reliability. In: Portney LG, Watkins MP, editors. Foundations of clinical research: Applications to practice. International edition. Upper Saddle River, NJ: Pearson Higher Education; 2008:557-85.
17. Winer BJ, Brown DR, Michels KM. Statistical principles in experimental design. New York: McGraw Hill; 1971.
18. Agarwal GG, Awasthi S, Walter SD. Intra-class correlation estimates for assessment of vitamin A intake in children. J Health Popul Nutr 2005;23:66-73.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
20. Gladman D, Ginzler E, Goldsmith C, Fortin P, Liang M, Urowitz M, et al. The development and initial validation of the Systemic Lupus International Collaborating Clinics/American College of Rheumatology damage index for systemic lupus erythematosus. Arthritis Rheum 1996;39:363-9.
21. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol 1997;50:79-93.
22. Liang MH, Fortin PR, Schneider M, Abrahamowicz M, Alarcon GS, Bombardieri S, et al. The American College of Rheumatology response criteria for systemic lupus erythematosus clinical trial for measures of overall disease activity. Arthritis Rheum 2004; 50:3418-26.
23. Worster A, Sardo A, Eva K, Fernandes CM, Upadhye S. Triage tool inter-rater reliability: a comparison of live versus paper case scenarios. J Emerg Nurs 2007;33:319-23.
24. Beaton DE, Bombardier C, Smith P, Mahood Q, Hogg-Johnson S, van der Velde G, et al. Course notes: measurement skills workshop. Toronto: University of Toronto Health Policy, Management and Evaluation; 2010.
25. Petri M, Hellmann D, Hochberg M. Validity and reliability of lupus activity measures in the routine clinic setting. J Rheumatol 1992;19:53-9.
26. Gladman DD, Goldsmith CH, Urowitz MB, Bacon P, Bombardier C, Isenberg D, et al. Crosscultural validation and reliability of 3 disease activity indices in systemic lupus erythematosus. J Rheumatol 1992;19:608-11.
27. Hawker G, Gabriel S, Bombardier C, Goldsmith C, Caron D, Gladman D. A reliability study of SLEDAI: a disease activity index for systemic lupus erythematosus. J Rheumatol 1993;20:657-60.
28. Yee CS, Isenberg DA, Prabu A, Sokoll K, Teh LS, Rahman A, et al. BILAG-2004 index captures systemic lupus erythematosus disease activity better than SLEDAI-2000. Ann Rheum Dis 2008;67:873-6.
29. FitzGerald JD, Grossman JM. Validity and reliability of retrospective assessment of disease activity and flare in observational cohorts of lupus patients. Lupus 1999;8:638-44.
30. Fleiss JL. The design and analysis of clinical experiments. New York: John Wiley and Sons; 1986.
31. Tammemagi MC, Frank JW, Leblanc M, Artsob H, Streiner DL. Methodological issues in assessing reproducibility — a comparative study of various indices of reproducibility applied to repeat ELISA serologic tests for Lyme disease. J Clin Epidemiol 1995;48:1123-32.
32. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995;4:293-307.