Rheumatology 2009;48:1548–1552
Advance Access publication 24 September 2009

doi:10.1093/rheumatology/kep183

Numerical scoring for the Classic BILAG index

Lynne Cresswell[1], Chee-Seng Yee[2], Vernon Farewell[1], Anisur Rahman[3], Lee-Suan Teh[4], Bridget Griffiths[5], Ian N. Bruce[6], Yasmeen Ahmad[7], Athiveeraramapandian Prabu[2], Mohammed Akil[8], Neil McHugh[9], Veronica Toescu[2], David D'Cruz[10], Munther A. Khamashta[10], Peter Maddison[11], David A. Isenberg[3] and Caroline Gordon[2]

Objective. To develop an additive numerical scoring scheme for the Classic BILAG index.

Methods. SLE patients were recruited into this multi-centre cross-sectional study. At every assessment, data were collected on disease activity and therapy. Logistic regression was used to model an increase in therapy, as an indicator of active disease, by the Classic BILAG score in eight systems. As both indicate inactivity, scores of D and E were set to 0 and used as the baseline in the fitted model. The coefficients from the fitted model were used to determine the numerical values for Grades A, B and C. Different scoring schemes were then compared using receiver operating characteristic (ROC) curves. Validation analysis was performed using assessments from a single centre.

Results. There were 1510 assessments from 369 SLE patients. The currently used coding scheme (A = 9, B = 3, C = 1 and D/E = 0) did not fit the data well. The regression model suggested three possible numerical scoring schemes: (i) A = 11, B = 6, C = 1 and D/E = 0; (ii) A = 12, B = 6, C = 1 and D/E = 0; and (iii) A = 11, B = 7, C = 1 and D/E = 0. These schemes produced comparable ROC curves. Based on this, A = 12, B = 6, C = 1 and D/E = 0 seemed a reasonable and practical choice. The validation analysis suggested that although the A = 12, B = 6, C = 1 and D/E = 0 coding is still reasonable, a scheme with slightly less weighting for B, such as A = 12, B = 5, C = 1 and D/E = 0, may be more appropriate.

Conclusions. A reasonable additive numerical scoring scheme based on treatment decision for the Classic BILAG index is A = 12, B = 5, C = 1, D = 0 and E = 0.

Key words: SLE, Outcome measures, Disease activity, BILAG, Statistics, Global score, Regression model, Treatment decision.

Introduction

The Classic BILAG index is a comprehensive composite clinical index that has been validated for the assessment of SLE disease activity [1–3]. This index was developed on the principle of the physician's intention to treat. It is a transitional index that captures changing severity of clinical manifestations. It has an ordinal scale scoring system by design that produces an overview of disease activity across eight systems. The individual system scores were not intended to be summated into a global score. As such, Classic BILAG scores should be treated as ordinal data.

However, the accommodation of ordinal data and the multiplicity of systems do limit the statistical analyses that can be performed. In situations where a single summary (numerical) measure for the Classic BILAG index is desirable, such as when assessing laboratory data, the coding scheme of A = 9, B = 3, C = 1 and D/E = 0 is currently used [4]. This scheme was developed on an *ad hoc* basis during a period in the early 1990s when the Classic BILAG index was being compared on both real and paper patients with two global score indices, the SLEDAI and the SLAM. It was then felt that a Grade A score (the most active disease score) should be approximately three times numerically that of a Grade B score (the second most active score) (Isenberg DA, personal communication). However, this numerical coding scheme has not been validated. Over the years, there has been increasing concern that this is not optimal as it is not consistent with the premise on which the grading of the Classic BILAG index was defined in terms of the need for treatment. With the current coding scheme, eight system scores of Grade C (all systems with mild disease activity) would result in a total numerical score of 8 which is comparable with a single Grade A (numerical score of 9) and more than two Grade Bs (numerical score of 6) numerically. This is conceptually inappropriate as mild disease activity (Grade C) would not be treated in the same manner as moderate to severe disease activity (Grades A and B) with regards to the use of immunosuppressives and/or moderate to high-dose corticosteroids.

In response to these concerns, this analysis was performed to develop an additive numerical scoring scheme for the Classic BILAG index based on treatment decision, utilizing data collected in routine clinical practice, which would allow the system scores to be summed into a global numerical score.

Patients and methods

This was a multi-centre cross-sectional study involving eight centres across the UK in which the primary objective was the validation of the BILAG-2004 index that has been reported previously [5]. Patients with SLE who satisfied four or more of the revised ACR criteria for classification of SLE were recruited [6, 7]. Patients were excluded from the study if they were pregnant, <18 years of age or unable to give valid consent. This study received multi-centre research ethical approval from Hull and East Riding Research Ethics Committee as well as approval from the local research ethics committees of all participating

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, [2]Rheumatology Research Group, University of Birmingham, Birmingham, [3]Centre for Rheumatology, University College London, London, [4]Department of Rheumatology, Royal Blackburn Hospital, Blackburn, [5]Department of Rheumatology, Freeman Hospital, Newcastle-upon-Tyne, [6]ARC Epidemiology Unit, University of Manchester, Manchester, [7]Department of Rheumatology, North West Wales NHS Trust, Bangor, [8]Department of Rheumatology, Sheffield Teaching Hospitals NHS Trust, Sheffield, [9]Department of Rheumatology, Royal National Hospital for Rheumatic Diseases NHS Trust, Bath, [10]Lupus Research Unit, St Thomas' Hospital, London and [11]Department of Rheumatology, University of Wales, Bangor, UK.

Submitted 23 October 2008; revised version accepted 1 June 2009.

Correspondence to: Caroline Gordon, Rheumatology Research Group, School of Immunity and Infection, College of Medical and Dental Sciences, The Medical School, University of Birmingham, Birmingham B15 2TT, UK.
Email: p.c.gordon@bham.ac.uk

centres. Written consent was obtained from all patients. This study was carried out in accordance with the Declaration of Helsinki.

This study commenced in March 2005 and was completed in August 2006. At every assessment, data on disease activity (using SLEDAI 2000, Classic BILAG and BILAG-2004 indices) and treatment (current treatment and changes to treatment) were collected. The majority of patients recruited into this study had more than one assessment during this period. Only the data for the Classic BILAG index will be discussed in this article.

Classic BILAG index

This is an ordinal scale index with eight systems (general, mucocutaneous, neuropsychiatric, musculoskeletal, cardiorespiratory, vasculitis, renal and haematology) [1, 8]. Each system is assigned a disease activity grade, ranging from A to E. Grade A represents very active disease requiring immunosuppressive drugs and/or prednisolone dose of >20 mg daily (or equivalent). Grade B represents moderate disease activity requiring a lower dose of corticosteroids, anti-malarials or NSAIDs. Grade C indicates mild disease requiring symptomatic therapy, whereas Grade D implies no current disease activity but the system had previously been affected. Grade E indicates no current or previous disease activity.

Change in therapy

Change in therapy has previously been chosen as the reference standard for disease activity in the criterion validity analysis of the BILAG-2004 index [5]. This is based on the well-defined benchmark for active disease, which is the decision to treat. In line with this, change in therapy was regarded as an indicator of disease activity and used as the response (outcome) variable in this study.

Change in therapy was the difference in treatment after the patient was assessed, compared with the therapy that the patient was on prior to the assessment (or change in treatment following the assessment) [5]. The medications of interest included immunosuppressives, anti-malarials, glucocorticoids, biological therapy, topical glucocorticoids, topical immunosuppressives, intravenous immunoglobulins, plasmapheresis, prasterone, thalidomide and retinoids. NSAIDs were not included as they are commonly used to treat non-lupus indications (especially for pain relief) and some could be obtained over the counter as non-prescription medication. A robust definition for change in therapy was used, as in our previous study [5]. Three categories of change were defined, namely 'no change', 'increase in therapy' or 'decrease in therapy'.

Increase in therapy was defined as any increase in the medications of interest regardless of any concomitant reduction in other medications. Decrease in therapy was defined as any decrease in the medications of interest without any concomitant increase in other medications. However, change in therapy was not just a simple change in the dose of the medications. The following special circumstances had to be taken into account as described previously [5]:

(i) dosing levels based on body weight;
(ii) step-down change of immunosuppressive therapy;
(iii) gradual escalation of immunosuppressive therapy following initiation;
(iv) increase in immunosuppressive therapy for steroid-sparing effect; and
(v) reduction or discontinuation of therapy due to side effects.

For some immunosuppressives, different dosing levels based on body weight were used in the definition of change in therapy (Table 1). A change in therapy was deemed to have occurred when there had been a change in the dosing level of these medications. These levels were based on clinical judgement and experience. For medications that were not listed in Table 1, a simple change in dose did constitute a change in therapy.

Table 1. Dosing levels of medications used for definition of change in therapy

Medications	Level 1	Level 2	Level 3
AZA, mg/kg/day	<1	1–2.4	⩾2.5
Mycophenolate mofetil, g/day	<2	2	3
Cyclosporin A, mg/kg/day	<2	2–3	>3
Tacrolimus, mg/kg/day	<0.10	0.10–0.15	>0.15
MTX, mg/week	<10	10–15	>15
Oral cyclophosphamide, mg/kg/day	<1	1–2	>2

A switch in immunosuppressive therapy was generally considered as an increase in therapy except in the situation of changing cyclophosphamide to AZA, MTX or cyclosporin. This is because it is a common practice to make such a change once the disease is under control, as prolonged cyclophosphamide therapy is associated with significant toxicity (step-down phase). In fact, this step-down phase would be equivalent to a reduction in therapy as the discontinuation of cyclophosphamide was considered as a decrease in therapy, whereas the initiation of the other immunosuppressive was not considered as an increase in therapy. In the case of the change from cyclophosphamide to mycophenolate mofetil, the situation was clarified with the local investigator as to whether the change was the result of failed cyclophosphamide (indicating an increase in therapy) or as a step-down phase (indicating a decrease in therapy).

As most immunosuppressives have potential toxicity, it is common practice to start at a low dose and gradually escalate to the target dose. To take this into account, any increase in the dose of immunosuppressives within the first 3 months of initiation was considered to be part of an escalation plan to achieve the target dose and not as an increase in therapy. Similarly, it is also a common practice to reduce the glucocorticoid dose gradually during this period, as immunosuppressives will have a steroid-sparing effect. Therefore, any concomitant reduction in the glucocorticoid dose during the escalation phase was not considered as a reduction in therapy.

If an immunosuppressive agent was started only for its steroid-sparing effect, this was not considered to be an increase in therapy. If any medication was decreased or discontinued due to side effects, this was not considered to be a reduction in therapy.

For this analysis, change in therapy was restricted into two categories, namely 'increase in therapy' and 'no increase in therapy'. Therefore, 'no increase in therapy' represents a combination of 'no change' and 'decrease in therapy'. Increase in therapy is chosen as the marker for active disease in the analysis, rather than decrease in therapy, as increase in treatment is very likely to occur with active disease and is unlikely to occur with inactive disease. The reverse does not hold true as inactive disease does not necessarily lead to a decrease in therapy, particularly if the patient is already on low-dose therapy (such as low-dose glucocorticoids).

Validation analysis

A validation exercise was performed using assessments from a single centre (Birmingham) from January 2003 to December 2004. In this centre, data on disease activity (using the Classic BILAG index) and treatment were collected prospectively at every assessment. However, the data collection on the treatment at every assessment is slightly different to that of the sample used in the primary analysis. As a result, the same definition of change in therapy could not be applied in the validation sample. Here, change in therapy was deemed to have occurred when there had been a change in the dose or change in the immunosuppressive drugs. There was no allowance for special circumstances such as dosing levels based on body weights, escalation phase, step-down phase, increase in therapy for only steroid-sparing effect and reduction or discontinuation of therapy due to side effects.

Statistical analysis

All statistical analyses were performed using Stata (Stata Corporation, TX, USA). Logistic regression was used to relate the probability of an increase in therapy (outcome variable) to the total number of Grades A, B, C and D/E (explanatory variables) obtained across the eight systems at each assessment. Grades D and E were combined together for this analysis as they both indicate inactivity. Therefore, four categorical scores were possible (A, B, C and D).

What will be termed as the total counts model corresponds to the logistic regression model which takes the form of:

$$\text{logit}(P) = \alpha + \beta_A x_A + \beta_B x_B + \beta_C x_C + \beta_D x_D \qquad (1)$$

where P is the probability of an increase in therapy; α is the intercept term; x_A, x_B, x_C and x_D are explanatory variables representing the number of Grades A, B, C and D/E scores, respectively, at each assessment; β_A, β_B, β_C and β_D are the coefficients for the corresponding explanatory variables x_A, x_B, x_C and x_D.

Grade D/E was used as the reference category in the model which meant that the coefficient β_D for Grade D/E was assigned the value of 0. Furthermore, as Grade D/E indicated inactivity, it was decided prior to analysis that both Grades D and E would be assigned the numerical value of 0. The values of the coefficients for the other explanatory variables (β_A, β_B and β_C) were estimated and used to derive the numerical values for Grades A, B and C. Estimation was based on generalized estimating equations with an independent working correlation matrix to account for the correlation between multiple assessments from the same patient. This generated a robust estimate for the variance matrix of the maximum likelihood estimates.

The aim is to provide a relative weighting of Grades A, B, C and D/E that can be used independently of the above logistic regression model. Therefore, the ratios of the estimates of these coefficients (denoted by $\hat{\beta}_A$, $\hat{\beta}_B$ and $\hat{\beta}_C$) provided the basis for the formulation of possible numerical values for Grades A, B and C. For simplicity, the numerical value for Grade C was fixed at 1. Therefore, the numerical value for Grade A should be close to $\hat{\beta}_A/\hat{\beta}_C$, and similarly, the numerical value for Grade B should approximate $\hat{\beta}_B/\hat{\beta}_C$.

The fitting of the total counts model suggested some possible coding schemes. For each proposed coding scheme, the corresponding numerical global score at each assessment can be calculated. This is achieved by converting the ordinal score of each system into its suggested numerical equivalence and summating the numerical score of the eight systems. Therefore, the numerical global score is given by the following formula:

$$\text{Numerical global score } (x_S) = A_S x_A + B_S x_B + x_C$$

where x_A, x_B and x_C represent the number of Grades A, B and C, respectively, at each assessment, and A_S and B_S represent the numerical values assigned to Grades A and B, respectively, by the particular coding scheme.

Further logistic regression models were used to determine how well these proposed coding schemes and the currently used scheme (A = 9, B = 3, C = 1 and D/E = 0) fit with the total counts model. These single variable models were in the form of:

$$\text{logit}(p) = \alpha + \beta_S x_S \qquad (2)$$

where P is the probability of an increase in therapy; α is the intercept term; x_S is the numerical global score obtained using a particular coding scheme; β_S is the coefficient for the numerical global score x_S.

Wald tests were used to check if there was a demonstrable difference in fit between a single variable model and the total counts model. A comparable fit between the single variable model of a particular coding scheme and the total counts model would indicate that the coding scheme suggested for Grades A, B, C and D/E is reasonable.

Receiver operating characteristic (ROC) curves, plots of sensitivity *vs* 1 − specificity, together with the area under these curves (AUCs) were used to compare the performance (predictive power) of the different coding schemes [9]. Larger AUC values are preferable, since the higher the AUC the better the scheme is at predicting increase in treatment.

For the validation analysis, ROC curves for the developed coding scheme and the currently used coding scheme were produced and compared with those for the original (primary analysis) sample.

Results

There were 369 SLE patients and they contributed 1510 assessments for the analysis. Of the 369 patients, 88.6% had more than one assessment during the study period. The demographics of the patients are summarized in Table 2.

Of the 1510 assessments, an increase in therapy was recorded 342 (22.6%) times. Summary of the Classic BILAG index scores attained at each assessment, depending on whether or not an increase in therapy was observed, is given in Table 3. There were eight systems, so the maximum possible occurrence of a particular grade at each assessment was eight.

TABLE 2. Demographics of patients recruited into the study ($n = 369$)

Patient characteristics	
Female sex, %	92.7
Age, mean ± S.D., years	41.6 ± 13.2
Race, %	
Caucasian	59.9
Afro-Caribbean	18.4
South Asian	18.4
Oriental	1.4
Others	1.9
Disease duration, mean ± S.D., years	8.8 ± 7.7
Number of assessments, %	
One	11.4
Two	12.5
Three	19.8
Four	18.7
Five	15.2
Six	10
Seven or more	8.1

TABLE 3. Summary of the Classic BILAG index scores at each assessment, by change in therapy ($n = 1510$)

	Increase in therapy	No increase in therapy
No. of visits with ⩾1 Grade A	76	19
No. of visits with ⩾1 Grade B and 0 Grade A	210	236
No. of visits with ⩾1 Grade C, 0 Grade B and 0 Grade A	53	781
No. of visits with just Grades D or E recorded	3	143
Total	342	1179
Number of Grade A at each visit, mean (range)[a]	0.3 (0–4)	<0.1 (0–2)
Number of Grade B at each visit, mean (range)[a]	1.1 (0–5)	0.3 (0–5)
Number of Grade C at each visit, mean (range)[a]	1.8 (0–6)	1.6 (0–6)
Number of Grades D or E at each visit, mean (range)[a]	4.9 (1–8)	6.2 (2–8)

[a] This is the mean number of Grades A, B, C and D/E at each assessment with a possible range of 0–8.

Primary analysis

Maximum likelihood estimation of the total counts model gave the coefficient estimates $\hat{\beta}_A = 2.73$, $\hat{\beta}_B = 1.54$ and $\hat{\beta}_C = 0.24$. When a baseline score of 1 was assigned to a Grade C score, the estimated coefficients suggested that the numerical value for Grade A should be approximately 2.73/0.24 = 11.4, and the numerical value for Grade B should be approximately 1.54/0.24 = 6.4. This suggested the following possible coding schemes:

(i) A = 11, B = 6, C = 1 and D/E = 0;
(ii) A = 12, B = 6, C = 1 and D/E = 0;
(iii) A = 11, B = 7, C = 1 and D/E = 0.

There was significant evidence that a model using the current coding scheme (A = 9, B = 3, C = 1 and D/E = 0) for the Classic BILAG index did not fit the observed data as well as the total counts model (Wald test $P < 0.001$) and that a more appropriate coding scheme should be considered. The three new coding schemes did fit the data as well as the total counts model, with non-significant (Wald test minimum $P = 0.77$) difference between the models using the new coding schemes and the total counts model. This was also reflected in the AUC measures, with the AUC for the three proposed schemes being ~0.864 and the AUC for the current coding scheme (A = 9, B = 3, C = 1 and D/E = 0) being 0.857. To summarize, the results suggested that any of the three proposed schemes would be a reasonable choice, and represented an improvement on the current coding scheme (A = 9, B = 3, C = 1 and D/E = 0). For ease of interpretability in application, the coding scheme of A = 12, B = 6, C = 1 and D/E = 0 was considered the preferred scheme.

Validation analysis

There were 405 SLE patients contributing 2360 assessments to the validation analysis: 93.2% were females, 21% Caucasian, 19% Asian, 54.3% Afro-Caribbean, 1% Orientals and 12.4% others. The mean age at first assessment was 42.4 years, with mean disease duration of 12.4 years.

Of the 2360 assessments, increase in therapy occurred in 439 (18.6%) assessments. When the coding scheme of A = 12, B = 6, C = 1 and D/E = 0 was applied, the AUC measure for the ROC curve (0.769) was less, as would be expected, than that for the original sample from which the coding was derived (0.864). The application of the two other coding schemes, discussed in the previous section, resulted in the same difference between AUC measures.

Interestingly, the currently used coding scheme (A = 9, B = 3, C = 1 and D/E = 0) also generated a decline in AUC measures from 0.857 to 0.769, which was identical to that for the newly proposed coding scheme (Fig. 1). The comparability of the ROC curves for the two coding schemes using the validation sample also reflected the fact that, for this sample, a formal improvement from use of the new coding scheme was not demonstrated.

Further formal analysis was done using a logistic regression model containing the global numerical score (x_S) together with total counts of Grades A and B (x_A and x_B). This analysis (data not shown) revealed that the comparable performance of the current coding scheme (A = 9, B = 3, C = 1 and D/E = 0) with the proposed coding scheme (A = 12, B = 6, C = 1 and D/E = 0) in the validation sample was due to the weight assigned to Grade B relative to Grade A (9/3 = 3 in the current coding scheme compared with 12/6 = 2 in the proposed coding scheme). This suggested that a coding scheme with a lower relative weighting of Grade B to A was more appropriate. Hence, a new coding scheme (A = 12, B = 5, C = 1 and D/E = 0) with a lower relative weighting of Grade B to A of approximately 2.5 was considered. This new coding scheme (A = 12, B = 5, C = 1 and D/E = 0) had comparable performance with the proposed coding scheme of A = 12, B = 6, C = 1 and D/E = 0 in both the original or validation samples (Fig. 1).

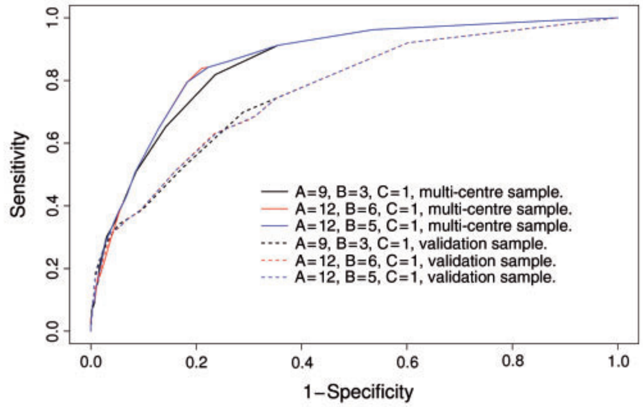

Fig. 1. ROC curves to compare sensitivity and specificity of the different coding schemes for the original (multi-centre) and validation samples.

Discussion

This is the first study that has attempted to determine formally an additive numerical coding scheme for the Classic BILAG index using data collected from actual clinical practice. It is clear from the results that the current coding scheme (A = 9, B = 3, C = 1 and D/E = 0) for this index is inappropriate, which was not wholly unexpected. From the regression model (total counts model), three new coding schemes were found to be reasonable, and represented an improvement on the currently used coding scheme (that was not derived empirically and was never validated) [4]. For practical application, the coding scheme of A = 12, B = 6, C = 1 and D/E = 0 was proposed as the preferred choice. However, the validation exercise did not detect a formal improvement of the proposed coding scheme (A = 12, B = 6, C = 1 and D/E = 0) when compared with the current coding scheme (A = 9, B = 3, C = 1 and D/E = 0). Further formal analysis suggested a compromise between the currently used coding scheme and the proposed coding scheme, in terms of the relative weight of a Grade B to A (for which the optimal value remains unclear), in the form of A = 12, B = 5, C = 1 and D/E = 0. This new coding scheme has been shown to be appropriate to both the original and validation samples. With this new coding scheme, an assessment with all eight systems scoring Grade C will not have a numerical score that would be almost equivalent to a Grade A or greater than that of two Grade Bs which is the case with the current coding scheme (A = 9, B = 3, C = 1 and D/E = 0). Therefore, the new coding scheme is intuitively more acceptable within clinical practice than the current coding scheme.

The proposed coding scheme did not fit the validation data as well as the original data, which is not unexpected. Notably, there is a comparable decline in the AUC measures, between the original sample and validation sample, for both the currently used and proposed coding schemes. This finding suggests that there are inherent differences between the samples analysed. It is very likely that the difference in the definition of change in therapy between these two samples is a factor. In the validation sample, the definition of change in therapy did not allow for circumstances such as dosing levels based on body weights, escalation phase, step-down phase, increase in therapy only for steroid-sparing effect, and reduction or discontinuation of therapy due to side effects. With this difference in definition, we might have expected a higher proportion of increase in therapy in the validation sample, but that was not the case (18.6% in the validation sample and 22.6% in the original sample). Hence, there were other factors that contributed to the differences between the two samples. It should be noted that the original data were collected from a variety of centres, whereas the validation sample was from a single centre, albeit with a broad ethnic mix. Apart from that, detailed training on how to use the index was provided to the physicians involved in the original sample (prior to the start of the study), which was more formal and extensive than the one provided to the physicians involved in the validation sample. This may have resulted in an operational difference in the index between these two samples. Although the validation sample is not ideal, it is the best available sample with prospective data collection (on disease activity and treatment) that is suitable for the validation analysis.

Change in therapy was used as the reference standard for disease activity in the absence of a better alternative. To date, the best benchmark to define active disease is the decision as to whether the disease activity should be treated. Physician's global assessment has been used previously as a gold standard, but this has been shown to perform unsatisfactorily with poor agreement between physicians in several studies [10–13].

One of the limitations of this study is in the cross-sectional design whereby only the disease activity at the time of assessment is taken into account. This does not take into consideration other factors that will have an impact on the treatment decision such as prior disease activity, current therapy, previous therapy (and its response), presence of comorbidities and patient's opinion (in particular, refusal to change therapy as advised). It is not possible to model all these factors into the analysis of this study. Furthermore, the reference for disease activity in this study of 'actual change in therapy' is different to that of 'intention to treat', which is the premise for the scoring of the Classic BILAG index. Actual change in therapy involves consideration of many factors (as discussed above), whereas the main consideration for intention to treat is disease activity. It is recognized that there is a variation between physicians in their threshold in changing treatment for a certain level of disease activity. Therefore, it was not surprising that increase in treatment did not occur in all assessments with active disease (at least one Grade A or B) in this study as demonstrated in Table 3.

Another drawback of this study is that the coding scheme is not derived from data with a full range of possible scores. For example, there was only one assessment with more than two Grade As recorded, and only 15 assessments with more than two Grade Bs recorded along with no Grade A. Thus, the data are too sparse to investigate the possibility of ceiling effects on the number of Grade As and Bs that contribute to increase in therapy. Nevertheless, the dataset is representative of those seen in routine clinical practice. In our experience, an assessment with three or more Grade As is a rarity (almost invariably in a very ill patient) and it would be very difficult to undertake another larger prospective multi-centre study than the one used for this analysis.

The key point to note is that the coding scheme of A = 12, B = 5, C = 1 and D/E = 0 is appropriate for both samples and is consistent with the principles on which the index is based. Therefore, in situations where a single summary numerical measure of the Classic BILAG index is desired, this coding scheme provides a reasonable way of achieving this. It should not be assumed that the same numerical coding scheme is applicable to the BILAG-2004 index. It has to be emphasized that the Classic BILAG index is ordinal scale with eight system scores by design, for which we strongly recommend that it should be treated as such in studies. Any coding scheme resulting in a single numerical value will not be able to capture all the information available from this index. Hence, there will be many circumstances in which the use of a single score will be inappropriate and the ordinal system scores should be used.

Rheumatology key message

- The unvalidated numerical coding for the Classic BILAG index (A = 9, B = 3, C = 1, D/E 0) is not appropriate. A12, B5, C1 and D/E 0, based on data, is more reasonable.

Acknowledgements

We would like to thank the Arthritis Research Campaign, Medical Research Council, Lupus (UK), the nurse specialists of all participating centres, the Wellcome Trust Clinical Research Facility (Birmingham) and Sandwell and West Birmingham Hospital NHS Trust for their support.

Funding: This study was supported by grants from Arthritis Research Campaign (Grant No. 16081), Medical Research Council (U.1052.00.009) and Aspreva Pharmaceuticals (unrestricted educational grant). Funding to pay the Open Access publication charges for this article was provided by arthritis research campaign.

Disclosure statement: C.-S.Y. is funded by an unrestricted educational grant from Aspreva Pharmaceuticals and has received honoraria from Roche and Merck Serono. C.G. is a consultant for and has received honoraria and grant/research support from Roche Pharmaceuticals, UCB, Bristol-Myers Squibb, Vifor Pharama, Amgen and Genentech. All other authors have declared no conflicts of interest.

References

1 Hay EM, Bacon PA, Gordon C *et al.* The BILAG index: a reliable and valid instrument for measuring clinical disease activity in systemic lupus erythematosus. Q J Med 1993;86:447–58.

2 Brunner HI, Feldman BM, Bombardier C, Silverman ED. Sensitivity of the Systemic Lupus Erythematosus Disease Activity Index, British Isles Lupus Assessment Group Index, and Systemic Lupus Activity Measure in the evaluation of clinical change in childhood-onset systemic lupus erythematosus. Arthritis Rheum 1999;42:1354–60.

3 Ward MM, Marx AS, Barry NN. Comparison of the validity and sensitivity to change of 5 activity indices in systemic lupus erythematosus. J Rheumatol 2000;27:664–70.

4 Stoll T, Stucki G, Malik J, Pyke S, Isenberg DA. Association of the Systemic Lupus International Collaborating Clinics/American College of Rheumatology Damage Index with measures of disease activity and health status in patients with systemic lupus erythematosus. J Rheumatol 1997;24:309–13.

5 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.

6 Tan EM, Cohen AS, Fries JF *et al.* The 1982 revised criteria for the classification of systemic lupus erythematosus. Arthritis Rheum 1982;25:1271–7.

7 Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. Arthritis Rheum 1997;40:1725.

8 Isenberg DA, Gordon C. From BILAG to BLIPS–disease activity assessment in lupus past, present and future. Lupus 2000;9:651–4.

9 Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561–77.

10 Guzman J, Cardiel MH, rce-Salinas A, Sanchez-Guerrero J, Alarcon-Segovia D. Measurement of disease activity in systemic lupus erythematosus. Prospective validation of 3 clinical indices. J Rheumatol 1992;19:1551–8.

11 Gladman DD, Goldsmith CH, Urowitz MB *et al.* Crosscultural validation and reliability of 3 disease activity indices in systemic lupus erythematosus. J Rheumatol 1992;19:608–11.

12 Gladman DD, Goldsmith CH, Urowitz MB *et al.* Sensitivity to change of 3 Systemic Lupus Erythematosus Disease Activity Indices: international validation. J Rheumatol 1994;21:1468–71.

13 Wollaston SJ, Farewell VT, Isenberg DA *et al.* Defining response in systemic lupus erythematosus: a study by the Systemic Lupus International Collaborating Clinics group. J Rheumatol 2004;31:2390–4.