Behçet’s disease: evaluation of a new instrument to measure clinical activity

B. B. Bhakta, P. Brennan¹, T. E. James², M. A. Chamberlain, B. A. Noble³ and A. J. Silman¹

Rheumatology and Rehabilitation Research Unit, University of Leeds, ¹ARC Epidemiology Unit, University of Manchester, ²Department of Ophthalmology, St James University Hospital, Leeds and ³Department of Ophthalmology, The General Infirmary at Leeds, Leeds, UK

Abstract

Objective. Behçet’s disease (BD) is a rare multisystem disorder characterized by vasculitis. At present, there are no laboratory markers that correlate well with the clinical activity in BD. This has led to the development of an instrument (BD Current Activity Form) to measure activity. Scoring is based on the history of new clinical features present over the preceding 4 weeks prior to assessment. Standardized questions were developed for all parts of the form. The face validity of the proforma was determined following worldwide collaboration with physicians and ophthalmologists managing patients with BD. The aim of this study was to evaluate the interobserver reliability of this form.

Methods. Nineteen patients fulfilling the International Study Group criteria for BD were randomly allocated, questioned and examined independently on the same day by five physicians experienced in BD.

Results. There was good agreement between the physicians’ rating of oral [intraclass correlation coefficient (ICC) = 0.87] and genital (ICC = 0.95) ulceration, skin involvement (ICC = 0.62 for pustules and ICC = 0.66 for erythema nodosum), arthritis (ICC = 0.62), headache (ICC = 0.80), large vessel (kappa = 0.53), nervous system (kappa = 0.61) and eye involvement (kappa = 0.77). There was poor agreement for the question relating to the presence of bloody diarrhoea (ICC = 0.28). There was significant bias in the rating of fatigue by one of the physicians (F = 5.2, P = 0.001).

Conclusion. Overall, this instrument has good interobserver reliability for assessing general disease activity. We therefore suggest that this proforma has a place in routine clinical monitoring of patients with BD, as well as assessing outcome in therapeutic trials.

Key words: Behçet’s disease, Activity, Measurement, Reliability.

Behçet’s disease (BD) is a rare multisystem disorder characterized by vasculitis. The classification of this condition depends on the presence of clinical features as defined by the International Study Group (ISG) on Behçet’s Disease [1]. The availability of an international, standardized, reliable measure of disease activity is of prime importance in monitoring the natural history and effects of therapeutic intervention in BD [2]. At present, there are no laboratory markers that correlate well with clinical activity in BD. This has led to the development of a standardized proforma [Behçet’s Disease Current Activity Form (BDCAF)] to assess disease activity which is based on history of clinical features.

Previous work [3] compared two schemes that were available to assess disease activity [The Iranian Behçet’s Disease Dynamic Measure (IBDDAM) [4] and a European scheme initially developed in the UK]. Interobserver and intra-observer agreement between two clinicians using both forms were assessed in 13 patients with BD as defined by the ISG criteria. Reliability depends not only on the patient’s accurate recall of symptoms, but also on the clinician’s interpretation of them. This study suggested that agreement between clinicians in scoring of clinical features was greater when the standard period was 28 days, as in the European form, compared with the Iranian form in which a variable time period is taken [3]. Although there was greater variability in scoring when the Iranian form was used, the opinion of the clinicians was that both forms had good aspects and that an internationally accepted activity form could be derived from them without great difficulty.

As a result of this study, a prototype form was
developed incorporating aspects of both forms. This was circulated to all members of the International Scientific Committee for comments. A workshop was held in Leeds, UK, in 1994 to arrive at a consensus view about the contents of the activity form (face validity) with emphasis on the need for clarity and consistency for potential use by clinicians worldwide. The inclusion of laboratory parameters within the activity form was raised by various members of the International Scientific Committee on Behçet’s Disease. Although it was agreed that inclusion of erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP) measurements would not add significantly to overall measurement of disease activity [5], it was appreciated that where disease appeared to be clinically inactive, a raised ESR or CRP might prompt further investigation. There was general agreement that standardized questions should be developed for each organ system which could be readily translated for international use. The interobserver reliability of this new instrument developed from these discussions is presented here.

Materials and methods

Subjects
Twenty patients were recruited from the combined Rheumatology/Ophthalmology Behçet’s disease clinic at Leeds General Infirmary. All patients recruited fulfilled the ISG criteria for the classification of BD. Five physicians (four rheumatologists and one clinical immunologist) experienced in assessing clinical activity in BD acted as the assessors.

Instrument
The BDCAF scores oral and genital ulceration, skin, joint and gastrointestinal involvement, presence of fatigue and headache according to the duration of symptoms. The presence and type of large-vessel and central nervous system (CNS) involvement are documented. Eye activity was deemed present if there was a history of blurring of vision or if the eye was painful or red. All patients were examined by an ophthalmologist who completed the Behçet’s Oculopathy Index. This index was subject to a separate reproducibility evaluation in a subsequent study. In addition, patients were asked to rate on a seven-point scale how active they felt their BD disease had been over the preceding 4 weeks and on the day of assessment. The clinicians also completed a seven-rating scale to assess their opinion of overall activity. Only new symptoms over the preceding 4 weeks that the clinicians felt were due to BD were scored. Standardized questions were developed for all parts of the form. For use during routine clinical practice, changes to current medication could also be documented. The layout and instructions for scoring are shown in Fig. 1a and b, respectively.

Method
The proforma was circulated to the assessors prior to the reliability study to familiarize them with the layout and instructions. This study was conducted over 1 day. All subjects were randomly allocated to the assessors (1–5). All subjects were independently questioned and examined on the same day by all observers. Scoring was based on the patient’s response to a series of standard questions for each organ system subscale comprising the general disease activity form. All assessment forms were coded to prevent patient identification. At the end of each assessment, any qualitative difficulties in completing the forms were documented by the assessors.

Table 1. Interpretation of the kappa statistic

<table>
<thead>
<tr>
<th>Kappa statistic value</th>
<th>Level of agreement (interpretation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.20</td>
<td>Poor</td>
</tr>
<tr>
<td>0.21–0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41–0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61–0.80</td>
<td>Good</td>
</tr>
<tr>
<td>0.81–1.00</td>
<td>Very good</td>
</tr>
</tbody>
</table>

Statistical analysis
The test–retest reliability of the scoring was assessed with respect to two properties: bias and agreement. The presence of bias relates to the systematic deviation between observers in their scoring patterns, while the level of agreement reflects the extent of random differences in scoring between the five observers. A high level of reliability for the form constitutes a high level of agreement in the score for each organ system in the absence of bias.

The level of agreement for each organ system was assessed by calculating the kappa statistic. This is the chance-corrected measure of agreement with values close to one indicating a high level of agreement (Table 1).

For dichotomous variables, such as assessment of eye, nervous system and major vessel involvement, the calculation of the kappa statistic between two observers is a simple procedure. An overall kappa for all five observers was also calculated.

For those clinical features measured on a five-point scale, a weighted kappa statistic was calculated in order to incorporate the relative seriousness of the different amounts of disagreement. The weighted kappa statistic can be calculated between pairs of clinicians; however, the calculation of an overall weighted kappa value for all five clinicians is not possible using this method. Therefore, an overall kappa value for all five clinicians was estimated by using the intraclass correlation coefficient (ICC).

The possibility of bias was also considered separately for each organ system. The proportions of patients who were scored as having involvement within each of the dichotomous variables (eye, CNS and major vessel involvement) were compared using the $\chi^2$ statistic (values > 3.84 indicating bias). For the organ systems that were rated using a five-point scale, the distribution of scoring was compared using the ANOVA method to detect any clinician consistently scoring higher or lower for an organ system. The strength of any bias is given as a $P$ value. Finally, comparison of the clinicians’ and patients’
BEHÇET’S DISEASE CURRENT ACTIVITY FORM

Date of assessment: ____________________

Centre: ____________________

Clinician: ____________________

Name: ____________________

Sex: M/F

Tel: ____________________

DOB: ____________________

Self rating scale of wellbeing over last 28 days
(Patients to tick face chosen)

- 2

- 1

0

1

2

3

4

5

6

7

8

9

10

(Patient to tick face chosen)

Self rating scale of wellbeing today

- 2

- 1

0

1

2

3

4

5

6

7

8

9

10

(Clinician to tick)

CLINICAL FEATURES

<table>
<thead>
<tr>
<th>ACTIVITY</th>
<th>(Other clinical details)</th>
</tr>
</thead>
</table>

Fatigue 0 1 2 3 4

Headache 0 1 2 3 4

Oral ulceration 0 1 2 3 4

Genital ulceration 0 1 2 3 4

Skin lesions erythema nodosum or superficial thrombophlebitis 0 1 2 3 4

Joints arthralgia 0 1 2 3 4

arthrosis 0 1 2 3 4

Gastrointestinal nausea or vomiting or abdominal pain diarrhoea with altered frank blood 0 1 2 3 4

Eye Is there eye activity? Yes / No

Possibly Ophthalmologic

CNS Is there new nervous system activity? Yes / No

(completed by Ophthalmologist)

Q 1. Yes / No Q 2. Yes / No Q 3. Yes / No Q 4. Yes / No Q 5. Yes / No

Major vessel Is there new major vessel activity? Yes / No

(completed by Ophthalmologist)

Q 1. Yes / No Q 2. Yes / No Q 3. Yes / No Q 4. Yes / No

Clinician’s impression of activity over last 28 days

Intention to initiate or change treatment? Yes / No

Current medication on arrival: ____________________

Change in medication: ____________________

Fig. 1. (a) The Behçet’s Disease Current Activity Form. (b) Scoring system.

perception of overall disease activity was calculated using the ICC.

Results

Of 20 patients initially recruited, one became ill during the study and was unable to complete it. Data on the remaining 19 patients are presented. Table 2 shows the bias and level of agreement between assessors when rating oral and genital ulceration, skin, joint and gastrointestinal involvement. Level of agreement is indicated by the ICC. Overall, the ICCs are > 0.60 for oral and genital ulceration, skin, joint involvement and the presence of headache, implying good agreement between assessors’ rating of these symptoms. There was, however, poor agreement for rating the presence of bloody diarrhoea. There was generally little bias between the assessors apart from rating fatigue, where a significant bias was identified. This bias related to Assessor 3 who consistently rated this symptom lower than the other four assessors. Repeat analysis omitting this assessor showed very good agreement and no systematic bias between the remaining four assessors (ICC = 0.82; F = 0.087, P = 0.967).

Table 3 shows the kappa values between the assessors’ rating of the presence of eye activity. The overall kappa is 0.77, implying good agreement between the assessors about the presence of activity. There was significant bias in the assessment of eye disease by Assessor 5 who consistently rated more patients as having active eye disease when compared to the other four assessors.

Table 4 shows the level of agreement between the assessors in identifying active CNS involvement. The overall kappa is 0.61, indicating good agreement between the assessors about the presence of new nervous system involvement.

Scoring system for Activity form

All scoring depends on the symptoms present over the preceding 4 weeks prior to assessment. Only clinical features that the clinician feels are due to Behçet’s Disease should be scored.

(1) To complete the self rating scale of overall wellbeing for the last 4 weeks, please ask the patient the following question:

There are some faces expressing various feelings, describing how you feel about your Behçet’s disease only, which of these faces describes how you have been feeling over the last 4 weeks?

To complete the self rating scale of wellbeing today, please ask the patient the following question:

There are some faces expressing various feelings, describing how you feel about your Behçet’s disease only, which of these faces describes how you feel today?

(2) Scoring for fatigue, headache, oral and genital ulceration, skin lesions, joint, symptoms, and gastrointestinal symptoms is based on duration of symptoms (rounded up to nearest week). Please ask the following question and tick in the blank with the reason system to be assessed:

Over the last 4 weeks, how many weeks was it total here you had ……….? 0

- no symptom

- 1 symptom for 1 week (1-7 days in total)

- 2 symptoms for 2 weeks (8-14 days in total)

- 3 symptoms for 3 weeks (15-21 days in total)

- 4 symptoms for 4 weeks (22-28 days in total)

(3) Eye involvement

Eye activity may be present if the following symptoms are present: (1) red eye, (2) blurred vision, (3) painful eye.

Please ask the following question (tick if symptom present):

Over the last 4 weeks have you had a red eye, painful eye, or blurred vision? 0

If any of these symptoms are present, or if you feel there may be eye activity refer patient to ophthalmologist who will determine the eye score (Behçet’s Ophthalmopathy Index)

(4) Nervous system

Please ask the following question (tick if symptom present):

Over the last 4 weeks have you had any numbness, difficulty with speech or hearing, double vision, weakness or loss of feeling in the face, arm or leg, memory loss, or loss of balance? 0

If the answer to all of these is “no” then move to Q 1.5 for diarrhoea involvement; otherwise determine the following:

Q 1. Are there new symptoms or signs consistent with meningeal involvement? 0

Q 2. Are there new symptoms or signs consistent with isolated cranial nerve involvement? 0

Q 3. Are there new symptoms or signs consistent with brainstem or cerebellar involvement? 0

Q 4. Are there new symptoms or signs consistent with cerebral hemispheric involvement? 0

Q 5. Are there new symptoms or signs consistent with spinal cord involvement? 0

(5) Major vessel involvement include intracranial involvement

Please ask the following question (tick if symptom present):

Over the last 4 weeks have you had shortness of breath, coughing up blood, or had any pain, swelling or discomfort of either the face, arm, or leg? 0

If the answer to all of these is “no” then move to Q 1.4 for diarrhoea involvement; otherwise determine the following:

Table 3 shows the kappa values between the assessors’ rating of the presence of eye activity. The overall kappa is 0.77, implying good agreement between the assessors about the presence of activity. There was significant bias in the assessment of eye disease by Assessor 5 who consistently rated more patients as having active eye disease when compared to the other four assessors.

Table 4 shows the level of agreement between the assessors in identifying active CNS involvement. The overall kappa is 0.61, indicating good agreement between the assessors about the presence of new nervous system involvement.
evaluation of a new Behçet’s disease instrument

731

T. Ass, bias and agreement between observers

Intraclass correlation
95% CI for variance due to

<table>
<thead>
<tr>
<th>Clinical feature</th>
<th>Bias between assessors</th>
<th>coefficient</th>
<th>random error and bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fatigue</td>
<td>( F = 5.21; P = 0.001 )</td>
<td>0.60</td>
<td>± 2.28</td>
</tr>
<tr>
<td>Headache</td>
<td>( F = 0.71; P = 0.58 )</td>
<td>0.80</td>
<td>± 1.20</td>
</tr>
<tr>
<td>Oral ulceration</td>
<td>( F = 2.09; P = 0.09 )</td>
<td>0.87</td>
<td>± 0.88</td>
</tr>
<tr>
<td>Genital ulceration</td>
<td>( F = 1.46; P = 0.22 )</td>
<td>0.95</td>
<td>± 0.54</td>
</tr>
<tr>
<td>Skin lesions</td>
<td>( F = 2.14; P = 0.08 )</td>
<td>0.66</td>
<td>± 1.44</td>
</tr>
<tr>
<td>Erythema nodosum or superficial thrombophlebitis</td>
<td>( F = 1.49; P = 0.22 )</td>
<td>0.62</td>
<td>± 2.04</td>
</tr>
<tr>
<td>Pustules</td>
<td>( F = 0.34; P = 0.85 )</td>
<td>0.73</td>
<td>± 1.84</td>
</tr>
<tr>
<td>Joint involvement</td>
<td>( F = 1.64; P = 0.17 )</td>
<td>0.62</td>
<td>± 1.64</td>
</tr>
<tr>
<td>Gastrointestinal involvement</td>
<td>( F = 1.88; P = 0.12 )</td>
<td>0.60</td>
<td>± 1.39</td>
</tr>
<tr>
<td>Nausea or vomiting or abdominal pain</td>
<td>( F = 2.42; P = 0.06 )</td>
<td>0.28</td>
<td>± 1.07</td>
</tr>
</tbody>
</table>

The level of agreement in identifying new large-vessel involvement is shown in Table 5. Although there is complete agreement between Assessors 2 and 4 and 3 and 5, the overall kappa is 0.53, implying only moderate agreement between all the assessors about the presence of new large-vessel involvement. Significant bias was present in the rating given by Assessor 1 who consistently found more new large-vessel activity. There was again considerable difficulty in categorizing the type of large-vessel involvement. Four patients were deemed to have new large-vessel involvement by one or more of the assessors, but in only two patients was the type of involvement identified.

ICCs were calculated to determine the agreement between assessors about the amount of overall activity over the preceding 28 days. There were some differences of opinion, with an ICC value of 0.48 suggesting only moderate agreement. The patient’s rating of disease activity over the preceding 28 days was compared with the clinician’s rating of disease activity over the preceding 28 days. On average, the patient tended to score well-being lower than the clinician. The amount of bias was \(-0.94\), suggesting that the patients tended to rate their well-being approximately one point (one face) lower than the clinician’s impression. Comparison between the patient’s self-rating score of well-being today and over the preceding 4 weeks showed a bias of 0.46. This suggests that patients tended to rate their well-being better on the day of assessment than over the preceding 28 days.

Discussion

The problem of scoring new disease activity and disaggregating it from established damage and resultant functional loss has exercised clinicians treating several multisystem diseases. Collaborative efforts have led to the establishment of several systemic lupus erythematosus indices [6–8]. A similar approach has been used in the development of the BD activity proforma. As with these indices, measurement of BD activity has to encompass several organ systems. Disease ‘severity’, disease ‘activity’ and established damage should be considered separately. The scoring system must (a)
adequately span the range between no activity and maximum activity in each system, (b) be free of systematic bias among different clinicians, (c) be valid, (d) be reproducible, (e) be comprehensive enough to incorporate geographical variability in clinical features [9] and (f) be simple enough to use in routine clinical practice.

Several instruments have been developed worldwide incorporating broadly similar organ system subscales. In some instruments, the measurement of disease activity relies solely on clinical features [2], while others include laboratory investigations, and changes in body weight and temperature. Although changes in body weight and temperature may indicate systemic activity, we have found them neither specific nor sensitive enough as markers for disease activity. Similarly, haemoglobin and ESR do not correlate well with activity [5].

In common with many other rheumatic diseases, the clinical features in BD vary considerably over time. In order to document this variation, new clinical features present over the preceding 28 days are scored using the BDCAF. This represents a compromise between assessing disease activity based on (a) clinical features on the day of assessment, which may be unrepresentative of overall disease activity and (b) clinical features present over a longer time period, as in the IBDDAM, which reduces reliability in terms of accurate recall of symptoms by the patient.

The disease activity rating for oral and genital ulceration, and skin lesions using the BDCAF relies solely on the duration of symptoms and does not take into account the size or number of lesions present, which might also reflect activity. Unfortunately, although documentation of the latter may be more representative of activity, its reliability is likely to be poor because of the difficulty for patients in recalling these symptoms accurately (the number and size of ulcers/skin lesions). There was good agreement between assessors in the scoring of oral and genital ulceration, and joint and skin manifestations using duration of symptoms alone. Although duration alone may not encapsulate all the features of activity in these organ systems, this scoring appears to be relatively simple to use, reliable and free of bias. Fatigue is a common problem in patients with BD, although it is not known how it correlates with the other clinical features. Assessor 3 consistently rated the fatigue symptoms lower than the other four assessors. The results from this study confirm that clinicians may differ in whether they attribute fatigue to BD or to other conditions that may co-exist, such as fibromyalgia. Although the presence of fibromyalgia was not specifically identified in this study, further studies are needed to quantify fibromyalgic symptoms in patients with BD.

Surrogate indicators are required for lesions that are less visible. Routine direct scoring of gastrointestinal (GI) activity is difficult. The scale to assess GI tract activity is based on two questions designed to identify upper and lower GI tract inflammation. The scoring represents a compromise between ease of monitoring activity and the accuracy with which the answers to the questions reflect inflammation. The lack of agreement in rating the presence and duration of bloody diarrhoea suggests that it is not often easy to determine from history alone whether this symptom relates to mucosal inflammation or is merely the result of a combination of other conditions such as drug-related diarrhoea (e.g. colchicine) and bleeding haemorrhoids. This item was included in the BD proforma as a result of its use in a proforma for measuring clinical activity in inflammatory bowel disease by the gastroenterology service. Patients in whom mucosal inflammation was suspected were referred to the gastroenterology service for further investigation and advice.

The need for surrogate indicators of activity also applies to large-vessel and nervous system involvement. Currently, MRI does not have a role in the monitoring of neurological disease activity as MRI lesions can be identified in patients with [10] and without [11, 12] clinical evidence of CNS involvement. Therefore, assessment of CNS involvement in BD is based on clinical features. As CNS involvement may be a solitary event (e.g. stroke) or relapsing (e.g. aseptic meningitis), it seems appropriate to score activity based on the site of lesion (e.g. meningeal, hemispheric, basal ganglia, brain stem and spinal cord), bearing in mind that the pathophysiology of CNS involvement remains unclear. This system would have merit as the location of presumed pathology could be determined by clinical features and radiological imaging. Currently, there is not sufficient knowledge of the natural history of lesions at various sites to enable the severity of the type of involvement to be graded and, therefore, the BDCAF aims to record and categorize neurological events during the course of the disease. This study shows that while there was good agreement in terms of identifying new CNS involvement, there remained disagreement between the assessors on the probable location of lesions based on clinical history and examination. From a practical perspective, further opinion is sought from a neurologist when neurological symptoms or signs are present (particularly if they are new), which may then require the CNS score to be amended.

Although the skin pathergy reaction is highly specific for BD, there is considerable variation in the rate of positivity in patients from different geographical areas, which limits its clinical usefulness. A positive pathergy reaction is common in patients from Iran, Turkey and Japan, but rare in those from the UK, the USA and France. Although this test has diagnostic importance, there is no evidence to suggest that its presence correlates with disease activity. The main difficulties with using this test as a measure of disease activity are the lack of consensus on the procedure (e.g. the optimal number of needle pricks needed [13]) and the practicalities, in a routine out-patient clinic, of grading the reaction 2 days after administration. For these reasons, it was not included as part of the disease activity assessment in the BDCAF.

The reliability of scores obtained for the patient and physician perception of overall activity was only moderate and associated with bias. While an overall perception
of activity is important, moderate reproducibility and the presence of bias may limit its usefulness.

The reliability and ease of use of this activity form have to be balanced against the validity of the questions being asked. If more serious manifestations can be predicted by readily assessed symptoms such as oral and genital ulceration, skin lesions, arthritis and superficial thrombophlebitis, then these organ system sub-scoring could be expanded to provide a more accurate representation of activity. This study has highlighted the difficulty in reliably scoring uncommon manifestations such as large-vessel involvement, GI tract inflammation and nervous system involvement.

This new instrument offers an easy-to-complete and reliable method of assessing and documenting clinical activity in patients with BD for use in routine clinical practice. The proforma takes 5–10 min to complete (longer if detailed examination of the vascular tree or nervous system is required). For the purpose of research, treatment trials targeted at a specific organ system would require a more comprehensive measure of activity within that organ system. For instance, a study of treatment aimed at oral ulceration (e.g. thalidomide) would not only require assessment of the duration of symptoms, but also other features which may relate to activity (number of ulcers, size of ulcers, number of crops of ulcers, site of ulcers, etc.). However, it is recommended that such detailed assessment is accompanied by this validated general disease activity instrument to alert the clinician conducting any trials to any advantageous or deleterious effects of the trial drug on other organ systems.

Acknowledgements

The authors are grateful to Dr Colin G. Barnes, Professor Hasan Yazici and Dr John M. Bamford for their advice during the development of this disease activity proforma. We would also like to thank the nursing staff of the ophthalmic outpatient department for their assistance during the reproducibility study. This research was supported by a grant from the Arthritis and Rheumatism Council.

References