Clinical outcome measures for Cutaneous Lupus Erythematosus

Joerg Albrecht1 and Victoria P. Werth2,3
1 Department of Medicine, Division of Dermatology, John Stroger Jr. Hospital of Cook County. Chicago, IL 60612
2 Philadelphia VA Medical Center, Philadelphia, PA, 19104 USA
3 University of Pennsylvania, School of Medicine, Philadelphia, PA, 19104 USA

Abstract

Cutaneous lupus erythematosus is a clinically heterogeneous group of rare skin diseases that only rarely have been subjected to controlled clinical trials. This may be have been partly due to a lack of suitable validated outcome instruments. Recently the FDA mandated that organ specific trials for lupus erythematosus need to use a combination of different outcome measures. The patient’s condition needs to be assessed in terms of quality of life, the patient’s global response and organ specific instruments that measure activity of the disease as well as damage due to the disease. For the skin the only formally validated and published instrument is currently the Cutaneous Lupus Erythematosis Disease Area and Severity Index (CLASI). This paper discusses the background of the development of the CLASI as well as issues related to its use and interpretation in the context of clinical research of CLE.

This issue of Lupus is dedicated to cutaneous lupus erythematosus (CLE). We want to discuss activity assessment tools that can be used to research cutaneous lupus. Questions of disease activity are largely based on our experience with designing the Cutaneous Lupus Erythematosus Disease Area and Severity Index (CLASI):

General desirable characteristics of outcome instruments

Finlay’s concisely defined what he thought were desirable qualities of an outcome instrument for clinical trials (1):

1. The method should be simple enough to use in a busy clinical setting.
2. The method should clearly separate scores derived from the observer and from the patient.
3. The signs chosen to be recorded should be amenable to change and should be unambiguous in their meaning and proven to be so. If the presence of two signs is highly correlated only one needs to be recorded.
4. Recording of area of involvement should be based on an assessment of the site of involvement rather than the virtually impossible task of determining an accurate total percentage involvement.
5. Validity testing including repeatability testing by the same and different observers must be carried out (1).

Corresponding author: Victoria P. Werth, MD, Professor of Dermatology and Medicine, Department of Dermatology, University of Pennsylvania, 2 Maloney, 3600 Spruce Street, Philadelphia, PA 19119, Tel.: 215-823-4439, Fax.: 215-823-8171, werth@mail.med.upenn.edu.
The understanding guiding these recommendations is that the outcome instrument is used in clinical trials, which may be multicenter.

These criteria are not necessarily needed for instruments used primarily to describe possibly new aspects of a disease, but they are needed to evaluate severity in any setting. Severity needs to be comparable or precisely described to be useful and unambiguous, e.g. mild pain is not a useful description while a pain level of 3 on a 0–10 visual analogue scale (VAS) is more reproducible. The need to assure intra- and inter-rater reliability is also true for the extension of existing instruments. If this evaluation is not done, this should be documented when the instrument is used and results have to be assessed with this limitation in mind.

Similarly the National Cancer Institute’s (NCI) Common toxicity criteria (2) measure the severity of adverse events and try to use stringently defined criteria to assure uniform coding.

**BACKGROUND**

In 2003 the FDA reacted to the dearth of new drugs licensed for SLE with a guidance instrument that was subsequently updated (3). It suggested that organ-specific therapies may be submitted to the FDA for approval. Clearly organs like the kidney require specific therapies that are adjusted as needed. Also the FDA was willing to consider therapy that reduces flares and activity but noted that there was a need for appropriate and validated outcome instruments that need to be developed to allow conduction of studies. These instruments did not exist at the time the FDA paper was written. Without these instruments studies would be prohibitively large and not reflect the treatment reality of the disease. The FDA also suggested that important clinical responses to therapy should be measured in terms of activity, damage, patient’s global response and health related quality of life (HRQL).

Although there is hope that the observation of one organ-system may reflect the disease course of systemic lupus erythematosus, we feel that systemic LE is unpredictable and the disease activity in one organ system may improve, while another system like the skin flares (4).

While in the opinion of the FDA adequate outcome measures for licensing trials needed to be developed, a huge number of instruments have been available for a long time. The authors of one paper that focused on CLE counted 60 available SLE scores (5). However, a number of scores like the SLE Disease Activity Index (SLEDAI) (5), the Lupus Activity Criteria Count (LACC) (6), SLICC/ACR Damage Index for SLE (7) only document the presence or absence of symptoms, skin symptoms being counted amongst them. Parodi et al. selected the Systemic Lupus Activity Measure (SLAM) (8) as the best available tool to be used for the skin but beyond clinical follow-up found it inadequate for clinical research (9). If the interest is focused solely on the skin, their assessment of available validated scores was more negative.

As mentioned above there are numerous outcome instruments for skin diseases. However, we could identify none that would have been suitable for CLE. This includes the Dermatology Index of Disease Severity (DIDS) that was designed to be a universal outcome instrument for skin diseases (10;11).
Character of the disease

CLE differs in some aspects from other skin disease for which outcome instruments exist, like psoriasis and atopic eczema. These characteristics are important for the assessment of these research tools:

1. CLE is a relatively rare disease
2. The disease has a number of subtypes which are clinically very heterogeneous.
3. Some of these subtypes cause permanent damage, i.e. scarring while others do not

In our opinion this means that adequately powered comparative clinical trials are likely to use a population of patients, which is recruited from a variety of CLE subgroups. Therefore outcome instruments designed for the purpose of outcome studies should try to capture the clinical essence of the major types of CLE, while not trying to risk an unwieldy extension of the instrument in order to capture any form of lupus. For the CLASI this approach lead to concentration on the symptoms of acute lupus erthematosus (ACLE), subacute cutaneous lupus erythematosus (SCLE) and discoid lupus erythematosus (DLE). Due to the relatively small group of researchers working on CLE, it is our opinion that it would be desirable, if data is generated with a more comprehensive larger instrument based on the CLASI, which includes patients with other lupus subgroups but also patients with SCLE, ACLE and DLE, would also be presented with the core CLASI. This would allow comparison between studies but not pose undue burden on the investigator.

An undesirable example of the opposite approach is atopic dermatitis where 10 different outcome instruments (12) which have undergone different levels of validation, make comparison between trials unnecessarily difficult. We tried to avoid this through the design process of the CLASI and only time will tell whether this was successful. The involvement of all dermatologist interest in CLE was beyond our means, but we sought input from the American and European community of dermatologists interested in autoimmune diseases, who gave valuable and insightful comments. In order to assure fairly wide applicability of the instrument, it was validated with dermatologists and dermatology residents (13) and recently with rheumatologists as well (14). Since the CLASI performed well throughout, we hope that the instrument will be widely used by both groups.

General

As mentioned above the CLASI was designed primarily with ALE, DLE and SCLE in mind, therefore it assesses damage as well as activity. However, these scores have to be reported separately in order to avoid paradoxical stability of the score as damage increases while activity wanes. This phenomenon was seen in clinical practice (15). Other approaches can be sought and are reasonable. For instance in acne there are grading scale for activity (16.17) and separate instruments to assess scarring (18).

Symptoms

Even beyond the question of activity and scarring, there are further dimensions of CLE pathology that can be evaluated symptom by symptom, such as pain or pruritus, which are often assessed with visual analogue scales from 0–10. However, pruritus can also be assessed in terms of character, e.g. stinging and tickling etc. and in terms of affect e.g. annoying or bothersome (19). Different aspects of symptoms can be assessed e.g. with scales that assess e.g. the effect of pruritus on sleep (19).
Design of the Cutaneous LE Activity and Severity Index (CLASI)

The CLASI has two scores, as do many outcome instruments for SLE and as mandated by the FDA. The first describes the activity of the disease while the second describes the damage done by the disease (Illustration 1). Activity is scored as a summary score of erythema, scale/hypertrophy, mucous membrane involvement, acute hair loss and non-scarring alopecia. Damage is scored in terms of dyspigmentation and scarring, including scarring alopecia. Patients are asked whether dyspigmentation due to CLE lesions usually remains visible for more than 12 months, which is taken to be permanent. If so, the dyspigmentation score is doubled. Each score is calculated by simple addition based on the extent of the symptoms. The extent of involvement for each of the skin symptoms is documented according to specific anatomic areas that are scored according to the worst affected lesion within that area for each symptom (13). The scores for damage and activity are kept separate to avoid problems associated with weighing the different aspects of the disease against each other. Also the damage becomes apparent as the activity wanes and thus the score may remain paradoxically stable as the disease becomes less active. When we validated the clinical utility of the CLASI, this concern was shown to be more than theoretical (15). Attempts to combine patient reported symptoms and physician measured signs in one score have also been attempted in atopic dermatitis but were abandoned subsequently (20,21).

The subjective symptoms documented by the patient, like itch, pain and fatigue, are recorded separately on visual 0 to 10 analogue scales. The same is true for the global summary score if that is used to assess the patient’s status. Our clinical validation demonstrated that the summary score of the general skin condition as given by the physician and the patient correlated well with the CLASI, with the correlation being slightly better for the physician’s global assessment (15).

In the following text we will describe two design characteristics of the CLASI namely the large reliance on erythema and the lack of precise area assessment. These cornerstones of the instrument’s design were further validated through a similarly designed instrument for dermatomyositis, which performed very well compared to other instruments coincidently designed at about the same time (22).

Erythema

Erythema of the skin is an easily recognizable symptom, even on black skin. It reflects activity well since it is caused by inflammatory hyperemia. It is also reliably scored compared to instrumental measurement of bloodflow by laser doppler flowmeter, erythema meter or chromameter (17). This has been confirmed by studies using laser doppler techniques and visual assessments of erythema (23).

Area

There is a natural desire to quantify the area of skin involved in a disease as precisely as possible, however, scores that depend heavily on skin area assessment like the SCORAD have demonstrated limited reliability (24–26). CLE is even more difficult to measure since lesions often only involve small areas, similar to guttate psoriasis (27). Lesion counting, which is most commonly used for acne, does not have necessarily a better reliability than surface area assessment (28). SCLE may present with skin lesions of widely varying size. During therapy these may break up and paradoxically increase the total score.

For psoriasis it has been shown that the extent and severity of involvement of the visible skin has greater impact on the patient’s experience than involvement of the covered skin.
(29,30) and that the percentage of body surface area (BSA) involved is a poor indicator of associated psychological morbidity (31,32). Informal assessment of CLE patients perception confirms this. Table 1 compares the results for the CLASI with the PASI and the rule of 9 (27,33,34) to estimate surface areas in relation to total BSA. We believe that the relatively exaggerated weight of the head reflects the clinical picture of CLE.

Validation process

Guidance on the validation process for new instruments is easily available, e.g. from Streiner and Norman (35). Williams, however, points out that the complete validation of an instrument is never done in one published study, and thus full testing and launching of a new instrument is a stepwise process (36). We have described the validation process of the CLASI in particular in great detail before (11), but this limitation holds certainly true for the CLASI as well. We have confirmed the responsiveness of the instrument in a prospective study (15). To gather data about clinical responsiveness of the CLASI we designed a clinical study in which 9 CLE patients (5 DLE, 2 SCLE, 2 DLE/SLE) were followed to observe the clinical course after a change of therapy from baseline until day 56 (15). We then correlated serial observations using the CLASI with other clinical data like the physicians’ and the patients’ assessments of global skin health, and the patients’ assessments of pain and itch, all documented on a 0–10 visual analogue scale. We could demonstrate high correlation of the CLASI’s activity score with the changes in physicians’ assessment of skin health (r=0.87, p=0.005, n=8), and patients’ global health score (r=0.85, p=0.004, n=9) as well as pain (r=0.98, p=0.004, n=5). Three patients had therapy failures, which lead to a stable score throughout their course.

Since then we are aware of the use of the instrument in a cross-sectional survey and case series evaluating medical treatment (37,38), but controlled clinical trials have not been performed. The probably most serious test of an outcome instrument would be the use of the CLASI for licensing purposes. This vetting by the FDA would certainly be the most stringent test for the instrument, but we are optimistic that this would go well.

How do we can we measure cutaneous lupus? What does a score mean?

As mentioned above, CLE is multidimensional and it is our impression that the different dimensions of the disease have to measured and assessed separately. The cutaneous aspects of the disease can be measured by the CLASI, the pain with visual analogue scales, the quality of life with different quality of life measures, preferably those focused on skin symptoms like the Skindex, rather than with other instruments like the SF-36 (39) which are likely to underestimate the specific quality of life issues related to skin diseases. What actually constitutes a meaningful clinical difference is notoriously difficult to assess. Based on experience with a pain scale it seems that on a 0–10 visual analogue scale a change of 2 points is clinically meaningful (40). For most important instruments there is a body of evidence that facilitates the interpretation of study outcomes, which are then taken up by the FDA. However, still the translation of what is generally considered a clinical meaningful outcome is often conceptually difficult. Is a score of 50 of the OSAAD (Objective Severity Assessment of Atopic Dermatitis) twice as bad as a score of 25 (i.e. is the score linear)? What does a 75% reduction of the Psoriasis Activity and Severity Index (PASI), short PASI 75, mean for the patient? The answer for the individual patient will depend on the location of the lesions and their visual prominence as much as on the restrictions that the patient feels in his normal life due to them. Some patients will always wear long sleeves if any skin blemish can be noted and restrict their leisure activities, while others are largely unfazed by clearly obvious and prominent lesions. Thus the answer for clinical trials is an evaluation that correlates quality of life instruments with clinical outcome instruments.
A general criticism of outcome instrument is that they do not adequately reflect the improvement of the patients. This is correct, but misleading because the criticism assumes that the disease process is a measurable conceptual entity that can be expressed as a clinical summary score. Even taking into account the differences of disease perception between patients, instruments are designed to measure only limited aspects of a disease. A thermometer measures body temperature but not what it feels like to have fever, or whether the fever is dangerous; similarly an instrument like the PASI reflects the affected skin area but not the psychological or the physiological effect of psoriasis. This is due to the nonlinear nature of many disease processes, which resists summarization in one number. A frequently used solution in therapeutic research is to express change in terms of differential, i.e. the relative reduction of the score. Again however, this is no panacea. The same PASI 75 may signal dramatic improvement in some patients while others feel that their disease has hardly changed because hand and facial lesions are still present. Thus dramatic improvement in Quality of Life instruments may be associated with changes in the disease activity based on the location of the lesions. As a consequence HRQL often does not parallel clinical disease severity as assessed by the clinician, though it may arguably be one of the most important clinical guides to treatment choice (41). HRQL, however, cannot be the only measure for clinical trials for several reasons, amongst them:

1. For early trials during clinical development, it is important to show that a treatment is physiologically active and can change the disease process, i.e. to demonstrate efficacy. For these trials the utilization of HRQL outcomes as a primary outcome may be far too early.

2. Particularly for scarring diseases treatment that halts the disease process is tremendously important. However, if scarring cannot be completely avoided, the patient may be traumatized by the wounds s/he has suffered and quality of life may not return to normal. Quality of life may not improve or even worsen with development of more damage, even when the activity of the disease has improved.

3. The diagnosis of a severe skin disease may be life altering in itself and lead to changes, like strict sun avoidance, which in themselves can be life altering.

However, when used to assess the final effectiveness of treatment, HRQL is an important outcome. That treatment may not be successful in changing HRQL by itself is not necessarily a sign of treatment failure. We validated the CLASI to demonstrate clinical responsiveness in small group of nine patients with CLE (5 DLE, 2 SCLE, 2 DLE/SLE). The results of our study were surprising in that complete resolution of the disease in two SCLE patients in particular was only associated with minimal improvement in their Skindex scores as they dropped from 68 to 67 and 77 to 70. Combined we only found moderate correlation (Spearman r=0.55) between improvements of the CLASI activity scale and the Skindex scores. In 3 DLE patients we found mild to moderate improvement of the Skindex even though the skin condition as measured by the CLASI did not improve. In three patients the Skindex and the CLASI correlated, in one the skin condition got worse and the patients Skindex increased, in another the skin remained unchanged as did the Skindex and in the third the skin improved as did the Skindex.(42) These results of a small study with a single investigator are only preliminary. But they confirm the clinical experience that arrest of the activity, particularly in scarring conditions, while preventive, does not solve the HRQL problems that the patients face. In order to address these problems the therapeutic approach has to include reconstructive and cosmetic treatment. The patient may also continue to be faced with limitations like sun avoidance that can seriously restrict social and leisure activities like sports. In addition CLE may be only part of the patient’s affliction that is compounded by the systemic symptoms that the patient may have.
As with the CLASI itself it would be desirable if investigators in this relatively rare disease could try to use the same HRQL instrument throughout. We have used the Skindex, but others may find other instruments more responsive and helpful. Ideally those interested in clinical research of lupus would come together to try to establish a standard that does not need to be the Skindex to facilitate comparison between studies. This has been done successfully in oncology.

**Conclusion**

Clinical trials in CLE are still rare but may become more frequent. We certainly think that the development of the CLASI may help to facilitate the design of precise trials that are necessary to assess a rare disease like lupus. Some, who are interested in subgroups of CLE may find it necessary to increase the precision of the instrument through well validated modifications. If these are used, comparability of the results should be attempted through documentation of the core CLASI as well.

At the same time CLE is multidimensional and assessment of the skin condition only through the CLASI is certainly not enough to measure it fully. Quality of life assessment, the patient’s global response, and pain assessment are helpful to give a fuller picture of the clinical condition the patient is in and the effect that treatment has on their condition.

**Acknowledgments**

This material is based upon work supported by the National Institutes of Health, including NIH K24-AR 02207 (Werth). This work was also partially supported by a Merit Review Grant from the Department of Veterans Affairs Veterans Health Administration, Office of Research and Development, Biomedical Laboratory Research and Development.

**Reference List**


Figure.
Table 1

Contribution of body areas to the total score in different cutaneous outcome instruments (13)

<table>
<thead>
<tr>
<th>Area</th>
<th>CLASI (activity)</th>
<th>CLASI (damage)</th>
<th>PASI</th>
<th>Rule of 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>28%</td>
<td>32%</td>
<td>10%</td>
<td>9%</td>
</tr>
<tr>
<td>Breast</td>
<td>14%</td>
<td>15%</td>
<td>5%</td>
<td>9%</td>
</tr>
<tr>
<td>Abdomen</td>
<td>7%</td>
<td>8%</td>
<td>5%</td>
<td>9%</td>
</tr>
<tr>
<td>Back</td>
<td>14%</td>
<td>15%</td>
<td>10%</td>
<td>18%</td>
</tr>
<tr>
<td>Legs</td>
<td>14%</td>
<td>15%</td>
<td>40%</td>
<td>36%</td>
</tr>
<tr>
<td>Arms/hands</td>
<td>14%</td>
<td>15%</td>
<td>30%</td>
<td>18%</td>
</tr>
<tr>
<td>Mucous membrane</td>
<td>7%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>