Adult Vulvar Lichen Sclerosus: Can Experts Agree on the Assessment of Disease Severity?

Michal Sheinis, MD,[1] Nicole Green, BSc,[1] Pedro Vieira-Baptista, MD,[2,3,4] Carmine Carriero, MD, PhD,[5] Gayle Fischer, MBBS, FACD, MD,[6] Catherine Leclair, MD,[7] Nina Madnani, MD,[8] Micheline Moyal-Barracco, MD, and Amanda Selk, MD, MSc, FRCSC[1,9]

Objective: The objective of this study was to test the severity rating of the signs and architectural changes for interrater reliability among world experts via analysis of lichen sclerosus (LS) photographs.

Methods: A recent Delphi consensus exercise established a list of symptoms, signs, and architectural changes, which experts feel are important to include in a severity scale. Photographs of vulvar LS were manually extracted from patient charts and 50 photographs with a range of severity of signs and architectural changes were chosen. Lichen sclerosus experts were invited to take part in the study and 3 dermatologists and 3 gynecologists were selected for their expertise and geographic variety. Raters assessed the photographs for multiple signs and architectural changes as well as an overall impression of disease severity on a 4-point Likert scale. Intraclass correlation coefficients were calculated.

Results: The intraclass correlation coefficients were very poor for individual signs and architectural changes as well as for overall disease severity when analyzed for all 6 raters as well as when analyzed with dermatologists' and gynecologists' responses grouped separately. There were no statistically significant correlations found.

Conclusions: Global experts were unable to agree on any signs, architectural changes, or an overall global impression to assess vulvar LS disease severity based on analysis of vulvar photographs. Standardized descriptions regarding what constitutes mild, moderate, and severe signs and anatomical changes are required before further scale development can occur.

Key Words: lichen sclerosus et atrophicus, lichen sclerosis et atrophicus, lichen sclerosus, lichen sclerosis, severity scale, expert consensus

(*J Low Genit Tract Dis* 2020;24: 295–298)

Lichen sclerosus (LS) is a chronic inflammatory dermatosis that primarily affects the vulva in women. The condition commonly presents with itching or pain and can lead to debilitating changes in a woman's functioning.[1] Up to 40% of those affected are asymptomatic[2] but can still present with loss of vulvar architecture, which can in turn be associated with sexual difficulties and urinary problems. Furthermore, in approximately 5% of cases, LS is associated with squamous cell carcinoma of the vulva.[1]

Despite the burden of vulvar LS, there is little evidence supporting the ideal treatment program, particularly regarding maintenance therapy and long-term monitoring, and little agreement among international experts in these areas.[3] There has been one large nonrandomized follow-up study of maintenance therapy supporting maintenance therapy over as needed use of topical corticosteroids as a way to reduce the long-term ill effects of LS.[4] Because disease progression is unpredictable, women with mild disease on clinical examination may not be treated optimally and thus develop more severe disease. An objective severity scale may conceptually provide treatment guidance for the busy clinician. Previous studies evaluating LS have devised their own severity scales including Likert scales[5–8] and qualitative assessments.[9,10] For instance, in Lee et al.,[4] the varying severity of hyperkeratosis was used to guide treatment selection. However, in evaluating these previous studies, there was no validation of the severity scales. Consequently, comparing the results of these trials is challenging and undoubtedly contributes to the myriad of different approaches used for long-term treatment. Proceeding with high-quality treatment trials is difficult without validated scales to measure the severity of LS.

With the collaboration of a panel of international vulvar experts, a recent Delphi consensus exercise was conducted and established a list of 19 symptoms, signs, and architectural changes agreed to be important markers of the disease severity.[11] The objective of this research study is to test these items (signs and architectural changes) for interrater reliability by directing a small number of world experts to rate the severity of the disease using photographs of LS. It was expected that the original list of 19 items would be reduced to a small number of reliable items that would then be tested in a clinical setting.

METHODS

Patient charts from the vulva clinic at Women's College Hospital from 2016 to 2018 were manually searched for patients with a diagnosis of LS and a photo of the condition. Photos were extracted if there was consent to use the photos for research. A total of 336 photos met eligibility criteria and were deidentified and extracted by the information technology department. The authors (A.S. and N.G.) manually reviewed the photos for quality. Fifty photos were then selected as they were considered to display a spectrum of severity of all of the signs and architectural changes included in the final Delphi consensus.[11]

A strong password-protected Excel spreadsheet was used to combine these photographs along with a 4-point Likert rating scale for each sign and architectural change as well as overall assessment of disease severity. "0" indicated absence of that sign/architectural change in the photo, "1" indicated mild presence, "2" indicated moderate presence, and "3" indicated severe presence. If their answer was left blank, this indicated an unsure assessment.

Expert raters were recruited via an e-mail invitation through the International Society for the Study of Vulvovaginal Disease. Experts, who had previously participated in the Delphi Consensus and who stated they wanted to be involved in further work on the severity scale, were invited to participate. Of those interested, the final raters were selected for their expertise in vulvar LS, global distribution, and an even distribution of 3 dermatologists

[1]Department of Obstetrics and Gynecology, University of Toronto, Toronto, Canada; [2]Hospital Lusíadas Porto, Porto, Portugal; [3]LAP – Laboratório de Anatomia Patológica, Unilabs, Porto, Portugal; [4]Lower Genital Tract Unit, Centro Hospitalar de São João, Porto, Portugal; [5]Department of Gynecology-Obstetrics, University of Bari, Italy; [6]University of Sydney, Sydney, Australia; [7]Department of Obstetrics and Gynecology, OHSU School of Medicine, Portland, OR; [8]P.D. Hinduja National Hospital, Mumbai, India; [9]Women's College Hospital, Toronto, Ontario, Canada

Correspondence to: Amanda Selk, MD, MSc, FRCSC, University of Toronto, Women's College Hospital Toronto, 76 Grenville St, Toronto, Ontario, M5S 1B2, Canada. E-mail: amanda.selk@utoronto.ca

P.V.-B. is currently involved in a project sponsored by Seegene.

The Research Ethics Board approval was granted from Women's College Hospital: 2018-0045-E. The other authors have declared they have no conflicts of interest.

© 2020, ASCCP
DOI: 10.1097/LGT.0000000000000534

TABLE 1. Expert Rater Demographics

Rater	Sex	Specialty	Country	Years in practice
Dr. Pedro Vieira Baptista	Male	Gynecology	Portugal	11–15 y
Dr. Carmine Carriero	Male	Gynecology	Italy	>20 y
Dr. Gayle Fischer	Female	Dermatology	Australia	>20 y
Dr. Catherine Leclair	Female	Gynecology	United States	16–20 y
Dr. Nina Madnani	Female	Dermatology	India	>20 y
Dr. Micheline Moyal-Barracco	Female	Dermatology	France	>20 y

and 3 gynecologists for a total of 6 raters. The raters were sent the files and given 2 weeks to complete the rating of the 50 photographs.

A statistician used SAS (SAS Institute Inc, Cary, NC) to calculate intraclass correlation coefficients (ICCs) among the 6 raters. Although the ICC was originally developed for continuous data, it has been adapted for use in evaluating interrater reliability for categorical (including dichotomous) responses as well (12). In evaluating the interrater agreement, we calculated the ICC for a recorded dichotomous response variable. The predominant response (the most common one) to each question was considered the “true” answer, and all responses were subsequently recoded to “true” or “false.” We then used the NLMIXED procedure in SAS to estimate the ICC for the agreements in the recoded rating responses to each question/item for evaluating the consistency in responses among the 3 gynecologist raters, the 3 dermatologist raters, and the 6 combined raters. Higher ICCs indicate greater agreement among raters. Originally, it was planned to repeat the exercise 2 weeks later to calculate intrarater reliability but this was not done because of initial results regarding interrater reliability.

Research ethics board approval was attained from Women's College Hospital: 2018-0045-E. This study was not funded by any sponsor.

RESULTS

The 6 raters selected varied in their geographical location, specialty, and sex although most of the raters had more than 20 years in practice (see Table 1). Of the 6 raters, 100% completed the ratings of all 50 photographs for each of the 13 signs and 6 architectural changes associated with LS (see Table 2). Intraclass correlation coefficients of responses showed poor agreement when analyzed for all 6 raters as well as when analyzed according to profession (3 dermatologists' and 3 gynecologists' responses grouped separately). With all 6 raters, ICCs ranged between 0.07 and 0.29 with all p values greater than 0.05. In the dermatologists alone group, ICCs ranged between 0.03 and 0.29 with all p values greater than 0.05. In the gynecologists alone group, ICCs ranged between 0.03 and 0.29 with all p values greater than 0.05 (see Table 2).

DISCUSSION

The objective of this research study was to evaluate the interrater reliability of signs and architectural changes in photographs of LS among global experts. The goal is to develop a validated and objective severity scale of disease burden as a way to follow clinical and research outcomes. The results demonstrate a

TABLE 2. Intraclass Correlation Coefficient Calculations for Individual Signs/Architectural Changes and Subjective Overall Disease Severity

Item	Combined raters (ICC)	Dermatologist raters (ICC)	Gynecologist raters (ICC)
Fissures	0.07	0.03	0.09
Whitening	0.12	0.09	0.14
Crinkly	0.26	0.29	0.10
Extent	0.11	0.10	0.10
Erosions	0.10	0.10	0.10
Ulcerations	N/A	N/A	N/A
Hyperkeratosis	0.29	N/A	N/A
Excoriations	0.14	N/A	0.09
Lichenification	0.16	0.13	N/A
Elasticity	N/A	0.28	N/A
Sclerosis	N/A	0.29	0.08
Petechiae	0.07	0.09	0.03
Clitoral hood fusion	0.29	0.29	0.29
Labial Fusion	0.11	0.11	0.27
Narrowing introitus	0.25	0.28	0.16
Anterior changes	0.25	0.29	0.09
Perianal involvement	N/A	0.14	0.29
Posterior commissure bands	0.29	0.29	0.29
Overall	0.11	0.03	0.09

*There was no statistical significance found.

N/A indicates not available.

complete lack of consensus regarding perception of severity for individual signs and architectural changes as well as for overall disease severity using photographs. Previous LS severity scales devised by researchers in the past[5–10] including those that used photographs to guide raters[4,8] were not tested for reliability.

Though disappointing, the demonstrated lack of interrater reliability in this study is valuable because it demonstrates that optimal evaluation of disease severity may require a clinical examination through inspection and palpation. Photographs, even in an ideal setting, may not convey all necessary information for full assessment. Severity scales devised by researchers in the past[5–10] including those that used photographs to guide raters[4,8] may be subject to the same challenges.

No definitions for mild, moderate, and severe were provided to the raters as these are difficult to define. This was intentionally done as the original Delphi list of signs and architectural changes is too long for day-to-day clinical use. One goal of this photo rating was to identify the signs with a range of severity that experts could easily identify, and then, these signs would merit further testing for scale development among general gynecologists and dermatologists. However, no consensus was reached for any one sign or architectural change. Even overall perception of disease severity varied widely between different expert reviewers. The Cochrane review on LS treatment from 2011[12] briefly refers to the various scoring systems but ultimately only takes into account the "investigator-related global degree of improvement," possibly because of the inconsistencies in the rating systems.

In other realms of dermatology, there has been some success with severity rating scales based on photographs, for example, the Wrinkle Severity Rating Scale.[13] However, in other complex diseases such as acne[14] and psoriasis,[15] reviews have found the same dilemma that exists in LS with a lack of validated scales that measure the full range of disease severity, which makes good quality research a significant challenge.

Although clinicians attempt to rate disease severity with objective measures such as architectural changes, the patient's assessment of disease burden may be drastically different. For instance, a patient's self-perception of her vulvar appearance and sexual function may signify different priorities. Inclusion of quality of life measures would capture the patient perspective and incorporate it into a final severity score. Several quality of life measures such as the Dermatology Life Quality Index,[16] the Skindex-29,[17] and the Female Sexual Function Index[18] already exist; however, they are not validated in the LS population. Recently, the Female Genital Self-image Scale was used to look at body image in LS.[19] There is current work ongoing eliciting patient input for this scale through qualitative and quantitative research methods.

In addition to not defining severity of the signs and architectural changes, there are several limitations to the study methodology. First, the photographs themselves were not standardized for distance or lighting as their primary purpose was clinical care and not research. Some of the photographs did not include the entire vulva and perianal skin, making it difficult to evaluate extent of disease. However, all raters were given the same photos, so this should not have affected ratings. In addition, signs such as hyperkeratosis, loss of elasticity, and sclerosis are difficult to assess visually and are optimally assessed with palpation and manipulation of the tissue. In retrospect, it would have been advantageous to attempt to agree on severity grading before assessing the photographs. Finally, this severity scale does not take into account emerging evidence regarding the genetic predictors of onset and progression of the disease[20,21]; however, as these data are so new, extensive research will need to be done before use of genetics in daily clinical practice.

The goal of a severity scale is to be able to put cases in order from least to most severe. Scales that only characterize findings with yes/no or presence/absence of an item have no ability to discriminate compared with Likert scales such as rating an item on a scale from 1 to 5 (least to most severe). Ultimately, if disease activity is to be assessed, the components of a scale need to show a range of severity. They need to show responsiveness, which means they can change over time and in response to intervention. Raters need to be able to assess items similarly between themselves (interrater reliability), and over time, the rater should be able to rate consistently the same way if given the same case twice (intrarater reliability). Future scale development will need to keep these principles in mind, which may mean focusing more on patient symptoms and less on architectural changes that do not usually change over time (i.e., scarring is generally considered irreversible).

In the case of LS, patient, perception of severity is often different than the clinician's and may be more related to function than appearance. A major goal of clinician rated severity is to determine treatment choice. In the case of LS, this involves topical therapy and surgical interventions. Therefore, in assessing severity, the things that can be modified by treatment are the most relevant. Loss of labia minora is often not noticed by patients and cannot be changed by treatment. However, for instance, severe hyperkeratosis, fissuring, and ulceration can cause significant discomfort and are also highly amenable to treatment. Fusion that splits recurrently is symptomatic and also treatable surgically. Thus, in assessing severity, one could stratify according to a combination of likelihood to cause symptoms and amenity to treatment.

CONCLUSIONS

The lack of validity in severity scales for adult vulvar LS is a significant barrier preventing high-quality clinical trials for disease treatment. This study demonstrates the lack of interrater reliability based on photographic analysis for predetermined signs and architectural changes among internationally renowned experts. Future studies with photographic analysis will need to focus on including a handful of items with definitions of what is mild, moderate, and severe. It needs to focus on including items that change over time and can be modified by treatment. These steps will reduce the heterogeneity and bias currently inherent in the clinical and research outcomes in the study of LS. Finally, LS severity cannot be assessed by visual inspection alone and needs to include symptom assessment and quality of life measures as well.

ACKNOWLEDGMENTS

The authors thank Zhong-Cheng Luo for the statistical analysis.

REFERENCES

1. Tasker GL, Wojnarowska F. Lichen sclerosus. *Clin Exp Dermatol* 2003;28: 128–33.
2. Goldstein AT, Marinoff SC, Christopher K, et al. Prevalence of vulvar lichen sclerosus in a general gynecology practice. *J Reprod Med* 2005;50: 477–80.
3. Selk A. A survey of experts regarding the treatment of adult vulvar lichen sclerosus. *J Low Genit Tract Dis* 2015;19:244–7.
4. Lee A, Bradford J, Fischer G. Long-term management of adult vulvar lichen sclerosus: a prospective cohort study of 507 women. *JAMA Dermatol* 2015;151:1061–7.
5. Tamburino S, Lombardo GA, Tarico MS, et al. The role of nanofat grafting in vulvar lichen sclerosus: a preliminary report. *Arch Plast Surg* 2016;43: 93–5.
6. Borghi A, Corazza M, Minghetti S, et al. Topical tretinoin in the treatment of vulvar lichen sclerosus: an advisable option? *Eur J Dermatol* 2015;25: 404–9.

7. Virgili A, Borghi A, Toni G, et al. First randomized trial on clobetasol propionate and mometasone furoate in the treatment of vulvar lichen sclerosus: results of efficacy and tolerability. *Br J Dermatol* 2014;171:388–96.

8. Corazza M, Maietti E, Toni G, et al. Combining topical tretinoin with mometasone furoate in the treatment of vulvar lichen sclerosus: results of dermoscopic assessment. *Dermatol Ther* 2018;31:e12735.

9. Ventolini G, Swenson KM, Galloway ML. Lichen sclerosus: a 5-year follow-up after topical, subdermal, or combined therapy. *J Low Genit Tract Dis* 2012;16:271–4.

10. Paslin D. Treatment of lichen sclerosus with topical dihydrotestosterone. *Obstet Gynecol* 1991;78:1046–9.

11. Sheinis M, Selk A. Development of the adult vulvar lichen sclerosus severity scale-a Delphi consensus exercise for item generation. *J Low Genit Tract Dis* 2018;22:66–73.

12. Chi CC, Kirtschig G, Baldo M, et al. Topical interventions for genital lichen sclerosus. *Cochrane Database Syst Rev* 2011;CD008240.

13. Day DJ, Littler CM, Swift RW, et al. The wrinkle severity rating scale: a validation study. *Am J Clin Dermatol* 2004;5:49–52.

14. Agnew T, Furber G, Leach M, et al. A comprehensive critique and review of published measures of acne severity. *J Clin Aesthet Dermatol* 2016;9:40–52.

15. Feldman SR, Krueger GG. Psoriasis assessment tools in clinical trials. *Ann Rheum Dis* 2005;64(suppl 2): ii65–8; discussion ii69–73.

16. Finlay AY, Khan GK. Dermatology Life Quality Index (DLQI)—a simple practical measure for routine clinical use. *Clin Exp Dermatol* 1994;19:210–6.

17. Chren MM, Lasek RJ, Flocke SA, et al. Improved discriminative and evaluative capability of a refined version of Skindex, a quality-of-life instrument for patients with skin diseases. *Arch Dermatol* 1997;133:1433–40.

18. Rosen R, Brown C, Heiman J, et al. The Female Sexual Function Index (FSFI): a multidimensional self-report instrument for the assessment of female sexual function. *J Sex Marital Ther* 2000;26:191–208.

19. Hodges KR, Wiener CE, Vyas AS, et al. The Female Genital Self-image Scale in adult women with vulvar lichen sclerosus. *J Low Genit Tract Dis* 2019;23:210–3.

20. Haefner HK, Welch KC, Rolston AM, et al. Genomic profiling of vulvar lichen sclerosus patients shows possible pathogenetic disease mechanisms. *J Low Genit Tract Dis* 2019;23:214–9.

21. Rotondo JC, Borghi A, Selvatici R, et al. Association of retinoic acid receptor β gene with onset and progression of lichen sclerosus-associated vulvar squamous cell carcinoma. *JAMA Dermatol* 2018;154:819–23.