





Assessment of inter-rater reliability of clinical hidradenitis suppurativa outcome measures using ultrasonography

A. B. Lyons,¹  S. Narla,²  I. Kohli,^{1,3} R. Zubair,⁴  A. F. Nahhas,⁵ T. L. Braunberger,¹ M. K. Joseph,⁶ C. L. Nicholson,⁷ G. Jacobsen⁸ and I. H. Hamzavi¹ 

¹Multicultural Center, Department of Dermatology, Henry Ford Health System, Detroit, MI, USA; ²Department of Dermatology, St Lukes Hospital, Easton, PA, USA; ³Department of Physics and Astronomy, Wayne State University, Detroit, MI, USA; ⁴Department of Dermatology, Broward Hospital, Ft Lauderdale, FL, USA; ⁵Department of Dermatology, Beaumont Health-Farmington Hills, Farmington Hills, MI, USA; ⁶Department of Dermatology, University of Michigan, Ann Arbor, MI, USA; ⁷Department of Dermatology, Wayne State University, Detroit, MI, USA; and ⁸Department of Public Health Sciences, Henry Ford Hospital, Detroit, MI, USA

doi:10.1111/ced.14889

Summary

Background. Hidradenitis suppurativa (HS) staging and severity is typically based upon physical examination findings, which can result in misclassification of severity based on subclinical disease activity and significant variation between healthcare providers. Ultrasonography (US) is an objective tool to help evaluate subclinical disease and to more accurately classify disease severity.

Aim. To evaluate inter-rater reliability in HS disease severity assessment using clinical and US techniques.

Methods. In total, 20 subjects underwent clinical evaluation of HS, independently by two physicians, using clinical outcome measures, including Hurley, Sartorius, HS Physician Global Assessment (HS-PGA) and Hidradenitis Suppurativa Clinical Response (HiSCR). US was subsequently performed, and clinical assessments were repeated. Intraclass correlation coefficients (ICC) were obtained to evaluate inter-rater agreement of each outcome measure before and after US.

Results. Pre-US to post-US improvement in ICC was seen with the Sartorius, HiSCR nodule and abscess count, and the HiSCR draining fistula count. The scores went from having 'good' rater agreement for Sartorius and HiSCR nodule and abscess count, to 'poor' rater agreement for HiSCR draining fistula count, to 'excellent' rater agreement among these scores.

Conclusion. US improved inter-rater agreement and should be used in conjunction with physical examination findings to evaluate disease severity to ensure uniform staging of HS.

Introduction

Hidradenitis suppurativa (HS) is a debilitating skin disease characterized by chronic, recurrent, painful inflammatory abscesses, nodules and sinus tracts.¹ The prevalence of the disease has been reported to be between 0.00% and 4.1%, although it is estimated that the disease is often underdiagnosed and

misclassified overall.² Currently, the staging and severity of HS is determined clinically using a variety of methods dependent upon lesion counts and extent of area involvement, including Hurley staging, Sartorius and HS Physician Global Assessment (HS-PGA), among others.^{1,3–6} In addition, the Hidradenitis Suppurativa Clinical Response (HiSCR) is an outcome measure to evaluate HS severity and improvement following treatment, which uses the number of abscesses and inflammatory nodules (AN count) and number of draining fistulas.⁷ Although techniques of clinical evaluation have been helpful in the evaluation and

Correspondence: Dr Iltefat H. Hamzavi, Multicultural Dermatology Unit, 3031 W Grand Blvd, Suite 800, Detroit, MI 48202, USA
E-mail: ihamzav1@hfhs.org

Accepted for publication 12 August 2021

management of HS, these scoring systems can often underestimate disease severity if there is subclinical disease, and can result in significant variation between different healthcare providers. This is particularly important because the presence of more severe disease often changes management from medical to surgical.

The use of ultrasonography (US) is emerging as a valuable objective tool to assess HS stage more effectively. The Sonographic Scoring of Hidradenitis Suppurativa (SOS-HS) instrument was recently developed and utilizes the US evaluation of HS affected areas to evaluate HS severity.⁸ Previous studies on the use of US in the assessment of patients with HS have demonstrated that, compared with other methods, the technique can more accurately identify HS lesions and often results in upstaging of classification of the severity of disease, but have not assessed the question of whether the technique improves reliability of clinical assessments between different assessors. Consequently, this study aimed to evaluate inter-rater reliability in HS severity assessment using both clinical and US techniques and sought to assess the utility of US in supplementing the current 'gold-standard' practices of clinical assessment alone to determine clinical staging outcome measures for HS.

Methods

This study was approved by the Institutional Review Board at Henry Ford Hospital (institutional review board approval no. 11505). The International Conference of Harmonization (ICH) and Declaration of Helsinki Guidelines, and Good Clinical Practice (GCP) were followed during the implementation of this study, and informed consent was obtained before any study procedures.

Participants

Our hospital system has a large HS specialty dermatology clinic with a diverse patient population from which patients were asked to participate in the study. Participants were included if they were ≥ 18 years of age, were able to understand the requirements and risks of the study, were able to provide informed consent, and had a diagnosis of HS. Participants were excluded if they were pregnant, breastfeeding or allergic to any components of US gel.

Procedure

Participants completed one study visit in which two physicians separately assessed for HS severity using

Hurley staging, Sartorius Score, HS-PGA and HiSCR before and after performing high-frequency US imaging (variable-frequency probes with upper frequencies of 22 MHz; LOGIQ *e*; GE Healthcare, Milwaukee, WI, USA) to determine SOS-HS. The pre-US assessments were performed separately with only one physician in the room at each time. The physicians were dermatology clinical research fellows who had been extensively trained in US and the HS clinical severity outcome measurement tools. Both physicians were present for the US, but only one performed it. Both physicians interpreted the US imaging results separately. All clinical assessments and SOS-HS were graded separately, and the scores were kept from the other grading physician throughout the duration of the study.

Statistical analysis

Statistical analyses were performed by the Division of Biostatistics and Research Epidemiology at Henry Ford Health System. Intraclass correlation coefficients (ICC) to assess inter-rater reliability were obtained pre-US and post-US for each scoring system (Hurley stage, Sartorius, HS-PGA, HiSCR and HS-SOS), and the 95% CI around each ICC was calculated. Pre-US to post-US improvement in the ICC was considered statistically significant if the 95% CIs around their post-US ICCs did not encompass their pre-US ICCs. Given the sample size of 20 patients, it was determined that an ICC of 0.50 should have a 95% CI of no more than ± 0.34 while an ICC of 0.90 should have a 95% CI of no more than ± 0.09 . These levels of precision were determined to be adequate for the study.

Rater agreement indicates how similarly two raters scored the patients. An ICC of 1 indicates that the raters scored the patients identically while an ICC of 0 indicates there was no similarity in how they scored the patients. Correlation coefficients < 0.40 represent 'poor' agreement, between 0.40 and 0.69 represent 'good' agreement, and > 0.69 represent 'excellent' agreement. Pre-US to post-US improvement in the ICC was considered statistically significant if the 95% CIs around their post-US ICCs did not encompass their pre-correlation coefficients ($P < 0.05$).

Results

Participants and measures

In total, 20 subjects (13 women, 7 men) completed the study. However, only patients containing a complete set of scores for each outcome measure from both

raters were used ($n = 19$). One patient did not have a complete set of outcome measures due to inadvertent submission of incomplete forms by one of the physicians. The ICC results between raters pre-US and post-US assessment are summarized in Table 1.

Inter-rater agreement

The results presented in Table 1 indicate there was 'excellent' rater agreement for Hurley stage and 'poor' rater agreement for the HiSCR draining fistula count before US. For pre-US results, both the Sartorius and the HiSCR AN count had 'good' rater agreement. After US was performed, there was no statistical change in ICC for Hurley stage. However, the Sartorius, HiSCR AN count and HiSCR draining fistula count achieved 'excellent' rater agreement post-US with statistical significance achieved (i.e. the 95% CIs around their post-US ICCs did not encompass their pre-US ICCs). The HS-PGA demonstrated 'good' rater agreement both pre- and post-US with no significant change. In addition, the HS-SOS demonstrated 'good' rater agreement.

Discussion

HS is a chronic and debilitating disease that impairs quality of life (QoL) and has limited effective treatment options.^{9,10} Thus, it is an important and rapidly

expanding area of ongoing dermatological research, but consistent clinical outcome measures remain an obstacle. These clinical outcome measures are further limited by how they are assessed currently, i.e. solely through physical examination. Physical examination techniques, such as visualization and palpation of HS lesions, have low sensitivity, and healthcare providers may miss deep or torturous fistulae or abscesses, as one study found clinically unrecognized fluid collections in 76% of patients with HS undergoing US evaluation.⁸ It is also difficult to differentiate between a draining abscess and a draining fistula. Furthermore, a clinically palpable lesion could correspond to either a nodule, abscess, fistula or scar, making it difficult to evaluate HS severity fully by physical examination alone.⁸

In this study, before the US assessment, there was 'poor' rater agreement for the draining fistula count, and the Hurley stage was the only outcome measure with 'excellent' rater agreement pre-US. Lack of inter-rater reliability is a major concern in many HS clinical outcome measures such as Hurley staging, Sartorius, HS-PGA and HiSCR. Further, a recent study by Thorlacius *et al.* investigated inter-rater reliability between 12 international HS experts (with > 10 years of experience) for several HS outcome measures, and found good inter-rater reliability for Hurley staging but found very wide limits of agreement for most of the other outcome measures they assessed.¹¹ Consequently, they did not recommend any of the other outcome measures (apart from Hurley stage and Physician Global Visual Analogue Scale) for measuring clinical severity of HS. However, despite the good inter-rater reliability for Hurley staging for disease severity, it often does not adequately capture disease activity. It does not incorporate patient-reported outcome components or QoL measurements and is therefore insufficient for evaluating response to treatment (e.g. one patient with Hurley stage 3 might be in immense discomfort due to pain, inflammation and drainage, while another patient with Hurley stage 3 might not). Thus, other outcome measures are needed to assess and quantify disease activity adequately.

US interpretation skills may vary between raters, but despite this, raters had 'good' inter-rater agreement for HS-SOS. In addition, US examination significantly reduced inter-rater variability between raters for multiple outcome measures, showing its utility in HS assessment. The current study demonstrated that the use of US resulted in statistically significant pre-US to post-US scoring improvement in the inter-rater agreement for Sartorius, HiSCR AN count and HiSCR

Table 1 Intra-class correlation results for rater agreement.

HS clinical outcome measure	ICC	95% CI
Pre-US assessment		
Hurley	0.71 ^a	0.39–0.87
Sartorius	0.59	0.21–0.81
HS-PGA	0.53	0.13–0.79
HiSCR AN	0.69	0.37–0.86
HiSCR draining fistula count	0.20	0.00–0.58
US assessment		
HS-SOS	0.63	0.27–0.83
Post-US assessment		
Hurley	0.61	0.24–0.82
Sartorius ^b	0.89 ^a	0.75–0.96
HS-PGA	0.59	0.20–0.81
HiSCR AN count ^b	0.94 ^a	0.82–0.97
HiSCR draining fistula count ^b	0.75 ^a	0.47–0.89

AN, abscesses and inflammatory nodules; HiSCR, Hidradenitis Suppurativa Clinical Response; HS, hidradenitis suppurativa; HS-PGA, Hidradenitis Suppurativa Physician Global Assessment; ICC, intraclass coefficient; SOS-HS, Sonographic Scoring of Hidradenitis Suppurativa; US, ultrasonography. ^aExcellent rater agreement. ^bPre-US to post-US ICC improvement that was statistically significant; significant change is indicated when the pre-US ICC does not encompass the 95% CI around the post-US ICC.

draining fistula count. This was due to better visualization of HS lesions with increased ability to distinguish sinus tracts, nodules, abscesses, inflammation and scarring. This emphasizes the utility of using US along with physical examination to obtain more accurate clinical staging and improve inter-rater reliability. High-frequency and ultrahigh-frequency US can provide additional detailed information beyond what clinical examination can provide, including the presence of subclinical disease activity, response to treatment, inflammation, and depth and margins of affected areas.^{12–16} Furthermore, a multicentre study conducted in patients with HS comparing Hurley staging when graded clinically and with US demonstrated intrarater and inter-rater US agreements of 94.9% and 81.7%, respectively.¹⁷ In contrast to our study, that study utilized only Hurley staging and had US performed by dermatologists with 10% of the cases reviewed by a radiologist as an external consultant.

US can detect subclinical anatomical information in HS that may significantly alter the severity, staging and treatment options, sometimes even from a medical to surgical approach. Loo *et al.*¹⁸ found that 56.9% of patients with HS had subclinical disease seen on US. A study performed by Napolitano *et al.*¹⁹ showed that 28.7% of patients had more severe HS measured by US SOS-HS when compared with the clinical Hurley staging system. Similarly, a study by Martorell *et al.*¹⁷ revealed that for patients diagnosed with Hurley stage I disease, staging changed to a more severe stage in 44.7% of patients after evaluation with US. Another study by Lacarrubba *et al.*²⁰ found that 27% of patients had worse HS measured by US when compared with clinical assessment of HS-PGA. In addition, Wortsman *et al.*⁸ found that US findings modified the disease management in 82% of adult patients with HS, and management was changed from medical to surgical in 24% of patients. A subsequent study in children (< 15 years of age) with HS revealed that US findings resulted in modification of medical management of the disease in 92% of cases.²¹

Wortsman recently called for US to become a standard of care for all patients with HS.²² It was further suggested that the ideal situation to perform US would be while making a baseline examination in all patients with HS and then intermittently to monitor the degree of severity. US is generally widely available in many clinical and emergency departments worldwide and is part of radiology residency training programmes. Moreover, there are a growing number of publications on the use of US in HS and a number of training courses offered through international US societies such

as the American Institute of Ultrasound in Medicine or the European Federation of Societies for Ultrasound in Medicine and Biology.²² Training to use US for the examination of patients with HS is obtainable, further supporting its utility along with clinical examination to improve inter-rater reliability.

The limitations of this study include a relatively small sample size, single-centre design and multiple raters, as research fellowships overlapped during the study. Of note, the same set of raters performed the pre-US and post-US evaluation for any individual participant. In addition, US is highly user-dependent, and it could have been useful for both physicians to have performed, as well as interpreted, the US individually. The nomenclature for HS lesions visualized during US is still being developed, and further work needs to be done to provide better interpretation of US features of HS. The limitations of US for evaluation of HS include the inability to detect lesions < 0.1 mm in size and decreased resolution when using lower frequencies to image deeper lesions, which can limit clear visualization of edges of deep sinus tracts in those who are obese. The strengths of this study include statistical improvement in ICCs pre- and post-US shown for multiple HS scoring systems despite having several different raters.

Conclusion

As demonstrated in this study, US can help improve inter-rater reliability for assessing HS disease activity and severity. The use of clinical grading alone often underestimates the true extent of disease. US should accompany clinical examination to decrease variation in staging and severity between providers to provide the appropriate treatment recommendations and to evaluate treatment response. Future studies should examine differences in treatment responses in patients who have been evaluated with US vs. those who were not, to determine if US evaluation affects patient outcomes. These studies should also increase the number of raters and patients to ensure the changes noted in inter-rater reliability can be verified.

Acknowledgement

We thank General Electric (GE) for providing the US used for this study.

Conflict of interest

ABL, SN, RZ and IK are subinvestigators for Lenicura and General Electric. IHH is the President of the HS

Foundation, an investigator for Lenicura and General Electric, a consultant for Incyte, and is on AbbVie Advisory Board (unpaid). AFN, TLB, MKJ, CLN and GJ declare that they have no conflicts of interest.

What's known about this topic?

- US can be utilized in patients with HS to help evaluate for subclinical disease and more accurately classify severity of disease.

What does this study add?

- US improved inter-rater agreement in this study and should be used in conjunction with physical examination findings to evaluate disease severity to ensure uniform staging.

References

- Gill L, Williams M, Hamzavi I. Update on hidradenitis suppurativa: connecting the tracts. *F1000prime Rep* 2014; **6**: 112.
- Miller IM, McAndrew RJ, Hamzavi I. Prevalence, risk factors, and comorbidities of hidradenitis suppurativa. *Dermatol Clin* 2016; **34**: 7–16.
- Hurley H. Axillary hyperhidrosis, apocrine bromhidrosis, hidradenitis suppurativa, and familial benign pemphigus: surgical approach. *Dermatol Surg* 1989; **133**: 1506–11.
- Alikhan A, Lynch PJ, Eisen DB. Hidradenitis suppurativa: a comprehensive review. *J Am Acad Dermatol* 2009; **60**: 539–61; quiz 562–3.
- Sartorius K, Lapins J, Emtestam L, Jemec GB. Suggestions for uniform outcome variables when reporting treatment effects in hidradenitis suppurativa. *Br J Dermatol* 2003; **149**: 211–13.
- Zouboulis CC, Del Marmol V, Mrowietz U *et al*. Hidradenitis suppurativa/acne inversa: criteria for diagnosis, severity assessment, classification and disease evaluation. *Dermatology* 2015; **231**: 184–90.
- Kimball AB, Sobell JM, Zouboulis CC *et al*. HiSCR (Hidradenitis Suppurativa Clinical Response): a novel clinical endpoint to evaluate therapeutic outcomes in patients with hidradenitis suppurativa from the placebo-controlled portion of a phase 2 adalimumab study. *J Eur Acad Dermatol Venereol* 2016; **30**: 989–94.
- Wortsman X, Moreno C, Soto R *et al*. Ultrasound in-depth characterization and staging of hidradenitis suppurativa. *Dermatol Surg* 2013; **39**: 1835–42.
- Alikhan A, Sayed C, Alavi A *et al*. North American clinical management guidelines for hidradenitis suppurativa: a publication from the United States and Canadian Hidradenitis Suppurativa Foundations: part I: diagnosis, evaluation, and the use of complementary and procedural management. *J Am Acad Dermatol* 2019; **81**: 76–90.
- Alikhan A, Sayed C, Alavi A *et al*. North American clinical management guidelines for hidradenitis suppurativa: a publication from the United States and Canadian Hidradenitis Suppurativa Foundations: part II: topical, intralesional, and systemic medical management. *J Am Acad Dermatol* 2019; **81**: 91–101.
- Thorlacius L, Garg A, Riis PT *et al*. Inter-rater agreement and reliability of outcome measurement instruments and staging systems used in hidradenitis suppurativa. *Br J Dermatol* 2019; **181**: 483–91.
- Oranges T, Vitali S, Benincasa B *et al*. Advanced evaluation of hidradenitis suppurativa with ultra-high frequency ultrasound: a promising tool for the diagnosis and monitoring of disease progression. *Skin Res Technol* 2020; **26**: 513–19.
- Wortsman X, Calderon P, Castro A. Seventy-MHz ultrasound detection of early signs linked to the severity, patterns of keratin fragmentation, and mechanisms of generation of collections and tunnels in hidradenitis suppurativa. *J Ultrasound Med* 2020; **39**: 845–57.
- Kelekis NL, Efstathopoulos E, Balanika A *et al*. Ultrasound aids in diagnosis and severity assessment of hidradenitis suppurativa. *Br J Dermatol* 2010; **162**: 1400–2.
- Wortsman X, Jemec G. A 3D ultrasound study of sinus tract formation in hidradenitis suppurativa. *Dermatol Online J* 2013; **19**: 18564.
- Wortsman X, Castro A, Figueroa A. Color Doppler ultrasound assessment of morphology and types of fistulous tracts in hidradenitis suppurativa (HS). *J Am Acad Dermatol* 2016; **75**: 760–7.
- Martorell A, Alfageme Roldán F, Vilarrasa Rull E *et al*. Ultrasound as a diagnostic and management tool in hidradenitis suppurativa patients: a multicentre study. *J Eur Acad Dermatol Venereol* 2019; **33**: 2137–42.
- Loo CH, Tan WC, Tang JJ *et al*. The clinical, biochemical, and ultrasonographic characteristics of patients with hidradenitis suppurativa in Northern Peninsular Malaysia: a multicenter study. *Int J Dermatol* 2018; **57**: 1454–63.
- Napolitano M, Calzavara-Pinton PG, Zanca A *et al*. Comparison of clinical and ultrasound scores in patients with hidradenitis suppurativa: results from an Italian Ultrasound Working Group. *J Eur Acad Dermatol Venereol* 2019; **33**: e84–7.

- 20 Lacarrubba F, Dini V, Napolitano M *et al.* Ultrasonography in the pathway to an optimal standard of care of hidradenitis suppurativa: the Italian Ultrasound Working Group experience. *J Eur Acad Dermatol Venereol* 2019; **33**: 10–14.
- 21 Wortsman X, Rodriguez C, Lobos C *et al.* Ultrasound diagnosis and staging in pediatric hidradenitis suppurativa. *Pediatr Dermatol* 2016; **33**: e260–4.
- 22 Wortsman X. Color Doppler ultrasound: a standard of care in hidradenitis suppurativa. *J Eur Acad Dermatol Venereol* 2020; **34**: e616–17.