Inter- and intrarater reliability of Hurley staging for hidradenitis suppurativa*

Z.N. Ovadja,[1,2] M.M. Schuit,[1] C.M.A.M. van der Horst[1] and O. Lapid[1]

[1] *Department of Plastic, Reconstructive and Hand Surgery, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, the Netherlands*
[2] *Department of Plastic, Reconstructive and Hand Surgery, OLVG, Oosterpark 9, 1091 AC Amsterdam, the Netherlands*

Linked Comment: Thorlacius. *Br J Dermatol* 2019; **181**:243–244.

Correspondence
Oren Lapid.
E-mail: o.lapid@amc.uva.nl

Accepted for publication
21 December 2018

Funding sources
None.

Conflict of interest
None to declare.

**Plain language summary available online*

DOI 10.1111/bjd.17588

Summary

Background Hidradenitis suppurativa (HS) is a chronic, inflammatory and recurrent skin disease. Different staging instruments have been suggested, but none has achieved universal acceptance. Despite the fact that Hurley staging is one of the most widely applied HS disease severity staging instruments, it has not been validated.
Objectives To determine the inter- and intrarater reliability of the Hurley staging system.
Methods Fifteen raters (five plastic surgeons, five general surgeons and five dermatologists) independently staged 30 photos of patients with HS according to Hurley staging at two time points. Reliability was assessed using kappa (κ) statistics, and multivariable logistic regressions were used to determine independent risk factors for photos with discordant staging.
Results Inter-rater reliability was moderate for the three stages of HS [κ = 0·59, 95% confidence interval (CI) 0·48–0·70]. It was moderate for Hurley stage I (κ = 0·45, 95% CI 0·32–0·55) and stage II (κ = 0·51, 95% CI 0·31–0·71) and it was almost perfect for stage III (κ = 0·81, 95% CI 0·62–1·00). The intrarater reliability was substantial for all stages and all raters (κ = 0·65, 95% CI 0·58–0·72). For stage I it was moderate (κ = 0·50, 95% CI 0·38–0·62), for stage II it was substantial (κ = 0·62, 95% CI 0·51–0·73) and for stage III it was almost perfect (κ = 0·82, 95% CI 0·77–0·87). Hurley stages II and III were less likely to result in discordant staging than Hurley stage I (odds ratios 0·47, 95% CI 0·29–0·77 and 0·21, 95% CI 0·12–0·38, respectively). The mean time spent on staging a photo was 14 s.
Conclusions Hurley staging is reliable for rapid severity assessment of HS, with moderate inter-rater and substantial intrarater reliability for all stages. It is best for assessing Hurley stage III HS, which is an indication for surgery.

What's already known about this topic?

- Hidradenitis suppurativa is a relatively common disease without a universally accepted disease severity staging instrument.
- Hurley staging is one of the most widely applied disease severity staging instruments.

What does this study add?

- This study is the first to determine the inter- and intrarater reliability of Hurley staging.
- Hurley staging is reliable for rapid severity assessment of hidradenitis suppurativa. It is best for assessing Hurley stage III disease, which is an indication for surgery.

Hidradenitis suppurativa (HS), also known as 'acne inversa', is a chronic, inflammatory and recurrent skin disease affecting apocrine-gland-bearing areas of the body, in particular the axillary, inguinal and anogenital regions.[1,2] HS affects about 1% of the European adult population,[3] and a 3 : 1 female-to-male ratio has been reported.[4] It presents after puberty as painful, deep-seated, inflamed lesions, including nodules, sinus tracts, abscesses and scarring, with the diagnosis based on the clinical features of the skin lesions and their chronicity.[4] HS leads to impaired quality of life with increased risk of depressive complaints.[5,6]

Although the cause of HS remains unclear, there are some potential factors that affect HS, which can be grouped into genetic, environmental, endocrine and microbiological factors.[7–10]

Despite it being a relatively common disease, there is a lack of high-quality evidence for the best treatment options.[11] A multidisciplinary approach with a combination of medicinal and surgical treatment is often needed, performed by a group of cooperating dermatologists, general surgeons and/or plastic surgeons.[12] Because of the need for a multidisciplinary approach and because the quantification of disease severity supports the development of evidence-based treatments, a reliable, easy-to-use disease severity assessment instrument for HS is essential.

Various HS staging instruments are currently used,[13–15] but a universally accepted instrument is still lacking. Ingram *et al.* recently reviewed the available clinical measures for staging HS severity. They concluded that 90% of these instruments lack any evidence of validity.[14]

The Hurley staging system, first described in 1989, is one of the most widely used HS disease severity instruments according to the available studies. It stratifies patients into three stages and was originally designed for the selection of the appropriate treatment modality in a certain body location: medical therapy for Hurley stage I, local surgery for Hurley stage II and wide surgical excision for Hurley stage III. This makes it easy to use in the short time of a routine clinical visit (Table 1).[16]

Despite the wide use of Hurley staging and the fact that new staging instruments are often compared with it, there is surprisingly, to the best of our knowledge, no information on its inter- and intrarater reliability. Therefore, the main aim of this study was to determine the inter- and intrarater reliability of Hurley staging. The secondary aims were to evaluate whether reliability differed between the three stages of Hurley staging, to assess whether reliability was dependent on the specialization of the raters, and to assess the mean time spent on staging HS according to the Hurley staging system.

Table 1 Hurley staging

Stage	I	II	III
Abscess	Single or multiple	Single or multiple, widely separated, recurrent	Diffuse or near-diffuse involvement
Sinus tracts	–	+	Multiple interconnected
Cicatrization	–	+	+
Area			Entire area
Treatment	Medicinal therapy	Combined medicinal and local surgery	Combined medicinal and wide surgery

Materials and methods

This prospective study determined the intra- and inter-rater reliability of Hurley staging among 15 raters. The study was performed using 30 photos of individual cases, selected by the investigators (Z.N.O. and O.L.).

Selection of photos

Photos were selected from websites licensed for educational purposes and they were therefore already staged according to Hurley, which will be referred to as the reference stage (examples shown in Fig. 1). The investigators reassessed the photos. The photos selected included lesions in the most involved regions of the body: the axilla, groin and mons pubis, and the gluteal, genital and perineal or perianal regions. Thirty photos of the highest available quality were chosen. The selected photos included cases of all three Hurley stages, and each stage was represented by 10 photos. The photos were placed in a random order in an online questionnaire.

Selection of raters

We tried to include a representative group of raters, which resulted in the selection and participation of five plastic surgeons, five general surgeons and five dermatologists. They worked in different hospitals and clinics in the Netherlands, had at least 2 years of experience in the treatment of patients with HS, and were familiar with the use of Hurley staging in daily clinical practice. The investigators were not involved as raters.

Rating process

The online questionnaire was sent to the raters. Each rater independently staged all 30 photos according to Hurley staging (at T1). The photos were provided in random order, and the raters were given unlimited time for assessment. The raters were blinded to clinical information and the source of the photos, and were not allowed to discuss their observations with other raters. Per rater, the questionnaire could be filled in only once. It was not possible to return to a previous question and correct the answer. The raters were not informed that the photos were divided into three groups (Hurley stage I, II and III) with exactly the same number of photos in each group.

To ensure unambiguous application of the Hurley staging, an overview of this staging instrument (Table 1) was available to the raters while they were completing the questionnaire. In order to assess intrarater reliability, the photos were assessed a

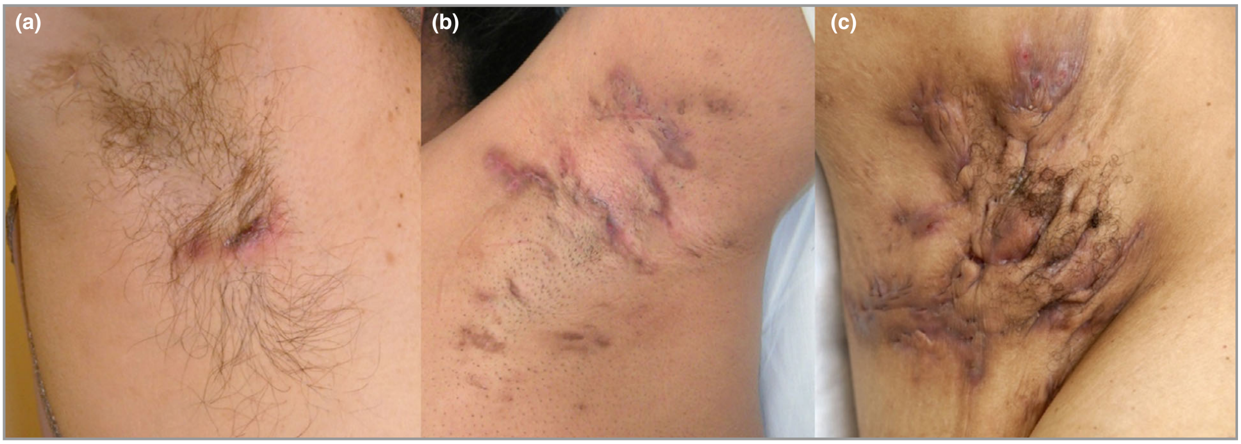

Fig 1. Examples of photos used in the online questionnaire. (a) Hurley stage I: axillary hidradenitis suppurativa (HS) with one or possible two abscesses, without evident sinus tracts or cicatrization in a smaller area. (b) Hurley stage II: axillary HS with separated abscesses, without evident sinus tracts, with cicatrization in a more diffuse area. (c) Hurley stage III: axillary HS with diffuse and connected abscesses, with sinus tracts and cicatrization in the entire area.

second time, 3 weeks later (T2), in order to prevent recall bias. On the second photo assessment for the intrarater reliability, the same photos were provided in a different random order.

Outcome 1: inter-rater and intrarater reliability

Inter-rater reliability was determined between all raters at T1. Intrarater reliability was determined within each rater at T1 vs. T2.

Outcome 2: risk factors for photos with discordant staging

We determined independent risk factors for photos with discordant staging between raters and the reference stage at T1. Discordance was based on correct staging (yes vs. no). Studied variables were Hurley reference stages (I, II or III) and profession (plastic surgeon, general surgeon or dermatologist).

Outcome 3: time spent on staging

The raters were given unlimited time for all assessments. The time spent on staging was recorded for each photo at T1. The raters were not informed that the time spent on staging the photos was recorded.

Statistics

The data were analysed using SPSS version 21 (IBM, Armonk, NY, U.S.A.) and the software package R version 3·1·1 (R Foundation, Vienna, Austria). All continuous data were assessed for normality distribution by analysing frequency histograms and Q–Q plots. Homogeneity of variance was assessed by the Levene test. The inter- and intrarater reliability will be described as the kappa-value (κ) and 95% confidence interval (CI).

Inter- and intrarater reliability were analysed using kappa statistics as described by Fleiss and Cohen, respectively. The kappa-values for intrarater reliability were calculated for each of the individual raters at T1 vs. T2, before calculation of the mean kappa-value with 95% CI. Interpretation of the values was carried out according to the guidelines of Landis and Koch, which suggest that values < 0 represent poor reliability, 0·00–0·20 slight reliability, 0·21–0·40 fair reliability, 0·41–0·60 moderate reliability, 0·61–0·80 substantial reliability and 0·81–1·00 almost perfect reliability.[17] A subgroup analysis was conducted for the (mean) kappa-values of inter-rater and intrarater reliability for the photos of each individual Hurley reference stage and each profession.

We performed sample-size calculations using the kappaSize package in R.[18] We assumed that the true value of kappa is 0·7 for both inter- and intrarater reliability and that 33% of images fall in Hurley stage I, 33% in Hurley stage II and 33% in Hurley stage III. In the inter-rater study, each of the 15 raters would rate a total of 30 images. This would give a lower boundary of the 95% CI for the estimated kappa of 0·573. In the intrarater study, each rater would rate a total of 30 images twice. This would give a lower boundary of the 95% CI for the estimated kappa of 0·486 for each rater.

Univariable and multivariable logistic regression models were used to determine risk factors for photos with discordant staging. A one-way ANOVA was used to test the statistical significance of differences in time spent on staging photos. Statistical significance was defined as $P < 0·05$. The Guidelines for Reporting Reliability and Agreement Studies were used to report this study.[19]

Results

The photos of the 30 patients were staged twice by a multidisciplinary panel of 15 experts from three medical disciplines.

Inter-rater reliability

The inter-rater reliability was moderate for all stages and all raters (κ = 0·59, 95% CI 0·48–0·70) (Table 2). It was moderate for Hurley stage I (κ = 0·45, 95% CI 0·32–0·55) and Hurley stage II (κ = 0·51, 95% CI 0·31–0·71), and was almost perfect for Hurley stage III (κ = 0·81, 95% CI 0·62–1·00). General surgeons had better overall inter-rater reliability (κ = 0·65, 95% CI 0·51–0·79) than plastic surgeons (κ = 0·58, 95% CI 0·44–0·72) and dermatologists (κ = 0·61, 95% CI 0·46–0·76). Especially for Hurley stage III, general surgeons (κ = 0·88, 95% CI 0·72–1·00) had better inter-rater reliability than plastic surgeons (κ = 0·76, 95% CI 0·52–1·00) and dermatologists (κ = 0·79, 95% CI 0·51–1·00). For Hurley stage I, general surgeons and dermatologists both had the best inter-rater reliability (κ = 0·55, 95% CI 0·30–0·80), compared with plastic surgeons (κ = 0·49, 95% CI 0·26–0·72). There was a small difference for Hurley stage II between the three groups, in which the general surgeons had the best inter-rater reliability (κ = 0·52, 95% CI 0·25–0·79), followed by plastic surgeons (κ = 0·49, 95% CI 0·31–0·67) and dermatologists (κ = 0·49, 95% CI 0·21–0·77).

Intrarater reliability

The mean intrarater reliability was substantial over all stages and all raters (κ = 0·65, 95% CI 0·58–0·72) (Table 3). The mean kappa-value of all raters was lowest for Hurley stage I (moderate; κ = 0·50, 95% CI 0·38–0·62), followed by Hurley stage II (substantial; κ = 0·62, 95% CI 0·51–0·73), and it was highest for Hurley stage III (almost perfect; κ = 0·82, 95% CI 0·77–0·87). This difference in intrarater reliability between the Hurley stages was also seen in the various professions separately (Table 3).

Risk factors for photos with discordant staging

In total 450 photos were staged by our panel of raters at T1. The raters' staging of photos differed from the reference stage in 112 cases (25%). An independent risk factor for discordant staging was photos labelled as Hurley reference stage I, which were more likely to result in discordant staging than photos labelled as Hurley reference stage II (odds ratio 0·47, 95% CI 0·29–0·77) or III (odds ratio 0·21, 95% CI 0·12–0·38) (Table 4). The medical specialty of the raters did not significantly affect the risk in the univariable or multivariable analysis.

Time spent on staging

The mean time spent on staging the 30 photos at T1 for all raters was 441 s, with a mean of 14 s per photo. There was a difference between plastic surgeons, general surgeons and dermatologists in the time spent staging all the photos. The general surgeons spent the least time on the staging, with a mean time of 13 s per photo, followed by the plastic surgeons and dermatologists (mean time per photo of 15 and 16 s,

Table 2 Inter-rater reliability of all raters and separately for the three disciplines with subgroup analysis

	All raters (n = 15)	Plastic surgeons (n = 5)	General surgeons (n = 5)	Dermatologists (n = 5)
All stages (n = 30)	0·59 (0·48–0·70) Moderate	0·58 (0·44–0·72) Moderate	0·65 (0·51–0·79) Substantial	0·61 (0·46–0·76) Substantial
Hurley I (n = 10)	0·45 (0·32–0·55) Moderate	0·49 (0·26–0·72) Moderate	0·55 (0·30–0·80) Moderate	0·55 (0·30–0·80) Moderate
Hurley II (n = 10)	0·51 (0·31–0·71) Moderate	0·49 (0·31–0·67) Moderate	0·52 (0·25–0·79) Moderate	0·49 (0·21–0·77) Moderate
Hurley III (n = 10)	0·81 (0·62–1·00) Almost perfect	0·76 (0·52–1·00) Substantial	0·88 (0·72–1·00) Almost perfect	0·79 (0·51–1·00) Substantial

The data are shown as the kappa-value (95% confidence interval). Interpretation of the strength of reliability is according to the guidelines of Landis and Koch.[17]

Table 3 Intrarater reliability of all raters and separately for the three disciplines with subgroup analysis

	All raters (n = 15)	Plastic surgeons (n = 5)	General surgeons (n = 5)	Dermatologists (n = 5)
All stages (n = 30)	0·65 (0·58–0·72) Substantial	0·70 (0·64–0·76) Substantial	0·55 (0·38–0·72) Moderate	0·69 (0·61–0·77) Substantial
Hurley I (n = 10)	0·50 (0·38–0·62) Moderate	0·55 (0·46–0·64) Moderate	0·40 (0·22–0·71) Fair	0·55 (0·36–0·74) Moderate
Hurley II (n = 10)	0·62 (0·51–0·73) Substantial	0·73 (0·56–0·90) Substantial	0·46 (0·22–0·70) Moderate	0·67 (0·56–0·78) Substantial
Hurley III (n = 10)	0·82 (0·77–0·87) Almost perfect	0·82 (0·76–0·88) Almost perfect	0·79 (0·67–0·91) Substantial	0·85 (0·76–0·94) Almost perfect

The data are shown as the kappa-value (95% confidence interval). Interpretation of the strength of reliability is according to the guidelines of Landis and Koch.[17]

Table 4 Univariable and multivariable analysis of risk factors for photos with discordant Hurley staging between raters and the reference stage

Variable	Univariable OR (95% CI)	P-value	Multivariable OR (95% CI)	P-value
Reference stage				
Hurley I	1	< 0·001	1	< 0·001
Hurley II	0·47 (0·29–0·77)	0·003	0·47 (0·28–0·77)	0·003
Hurley III	0·21 (0·12–0·38)	< 0·001	0·21 (0·12–0·38)	< 0·001
Profession				
Plastic surgeon	1	0·21	1	0·19
General surgeon	1·14 (0·69–1·90)	0·60	1·16 (0·68–1·96)	0·59
Dermatologist	0·71 (0·41–1·22)	0·22	0·70 (0·40–1·22)	0·20

OR, odds ratio; CI, confidence interval.

respectively). However, this difference was not significant (P = 0·64) (Table 5). There was a statistically significant difference in the mean time spent on staging photos between the Hurley stages: Hurley I, 329 s for 10 photos (95% CI 209–449); Hurley II, 182 s (95% CI 151–212) and Hurley III 152 s (95% CI 114–189) (P = 0·006). Games–Howell post hoc analysis revealed that the mean time spent on staging was statistically significantly longer for Hurley stage I than for Hurley stage III (P = 0·04).

Table 5 Mean time the raters spent on staging the photos at the first appraisal

	Time (s)	P-value
Reference stage		
Hurley I (n = 10)	329 (209–449)	0·006
Hurley II (n = 10)	182 (151–212)	
Hurley III (n = 10)	152 (114–189)	
Profession		
Plastic surgeon (n = 5)	453 (312–593)	0·64
General surgeon (n = 5)	387 (292–482)	
Dermatologist (n = 5)	485 (303–667)	

The data are shown as the mean time (95% confidence interval).

Discussion

Assessment of the disease severity of HS is a challenge in daily clinical practice owing to the lack of a standard, accepted assessment instrument and the wide variability in the clinical appearance of HS. Considering the need for a multidisciplinary approach in the treatment of HS and the need for more research, an optimal staging instrument is needed that can be easily and quickly implemented in clinical routine and provide an accurate, responsive and clinically relevant representation of the disease severity.[14] One major impediment is the poor reliability of more detailed staging instruments.[11]

Although Hurley staging was the first described HS severity instrument and its usability has previously been questioned,[16] it remains the most widely used staging system in the research of HS, as well as in clinical practice.[13,14] However, to the best of our knowledge, its inter- and intrarater reliability have never been determined. Therefore, the main purpose of this study was to determine whether Hurley staging is a reliable scoring instrument for the staging of HS.

The overall inter-rater reliability was moderate, and the intrarater reliability was substantial. A notable outcome of this study is the difference in reliability between the three Hurley stages. The highest inter-rater reliability was for Hurley stage III (almost perfect), compared with Hurley stages I and II (both moderate). This was also the case for intrarater reliability, which was almost perfect for Hurley stage III, compared with moderate and substantial for Hurley stages I and II, respectively. The Hurley reference stage was an independent risk factor for discordant staging. Photos with mild cases of HS (Hurley I) had higher odds of discordant staging than those of more severe cases (Hurley II or III). The medical specialization of the rater did not affect discordant staging. These outcomes suggest that the Hurley staging is particularly suitable for defining severe cases of HS in and across the professions most involved in managing HS.

Considering the photographic assessment in this study and the high reliability outcome for severe cases in and across professions, the Hurley classification could be useful for assessment in telemedicine and for research purposes, specifically for severe, stage III disease.

The mean time spent on staging a photo of a patient with HS was 14 s for all raters. This is much shorter than the time reported for other staging instruments (Acne Inversa Severity Index, 46·4 s and the Revuz version of the original Sartorius score, 83·2 s).[20,21] However, we must take into account that Hurley staging describes the finding in a particular region and not the total disease burden for the patient.[16]

Assuming that staging HS lesions in real clinical practice by palpating lesions and looking at them from different angles is easier than staging photos, this outcome suggests that Hurley staging is a time-efficient staging instrument. An earlier study described Hurley staging as useful for rapid staging of HS severity in clinical practice.[22]

It was remarkable that all raters spent less time on staging photos of Hurley stage III, and also had the best inter- and intrarater reliability on those photos. This confirms the conclusion that Hurley staging was especially useful for recognizing severe cases of HS. An explanation for our observation that Hurley staging seems to be less reliable for mild and moderate cases of HS could be that Hurley stages I and II are more difficult to recognize on a photo than in real patients, as mentioned above. In addition, raters are more certain that a severe-looking case must fit into Hurley stage III.

One of the strengths of this study is that the raters were from the three different disciplines that are most involved in the treatment of HS. Furthermore, the panel consisted of a large number of raters with a wide spread of experience, varying from 2 years to decades. This reflects clinical practice in a multidisciplinary setting and therefore we consider the panel to be representative. Another strength of the study was the flexibility around completing the questionnaire, whereby the raters were allowed to pause the questionnaire and finish at another moment. This reduced the chance of pressure bias.

However, our study has certain limitations. We chose to use available photos of patients with HS from websites licensed for educational purposes, in order to have prelabelled cases according to Hurley staging, causing limited availability. The selection of pictures we used may have not represented all possible scenarios seen in clinical practice. In addition, it is possible they did not precisely represent the normal anatomical distribution. However, lesion appraisal is dependent not on location, but on clinical skin features, which are similar in all regions. The use of available photos from educational sources could also have caused rating bias, when raters have seen and remembered the photos before the assessment. Firstly, we accepted this chance of bias because we used different sources, blinded for raters, which makes it is less probable that the raters could have recognized some of the used photos. Secondly, if the raters had recognized some of the included photos after gathering information on how to use the Hurley staging, it is assumable that some photos divided over all stages would be recognized. Finally, choosing the right stage does not directly affect inter- and intrarater reliability, it is about choosing the same stage as others and one's own previous assessment.

The use of photos and not patient examination may be seen as a weakness of our study. However, Hurley staging does not require palpation of the lesions. Considering the trend towards photographic staging and telemedicine,[23] this limitation simulates daily communication between specialists and can in fact be seen as an advantage. The mean time the raters spent on completing the questionnaire was short. We attribute this to the ease of using this staging instrument for severe cases of HS, but it is also possible that the raters did not take sufficient time to fill in the questionnaire as accurately as possible.

In conclusion, Hurley staging is reliable for rapid assessment of HS, and it is best for assessing Hurley stage III disease for all of the most involved professions, meaning whether or not a patient should be operated on. Our study proves that the staging can be done on photos and is therefore appropriate for telemedicine.

References

1 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.

2 Kurzen H, Kurokawa I, Jemec GB *et al.* What causes hidradenitis suppurativa? *Exp Dermatol* 2008; **17**:455–6.

3 Revuz JE, Canoui-Poitrine F, Wolkenstein P *et al.* Prevalence and factors associated with hidradenitis suppurativa: results from two case–control studies. *J Am Acad Dermatol* 2008; **59**:596–601.

4 Revuz J. Hidradenitis suppurativa. *J Eur Acad Dermatol Venereol* 2009; **23**:985–98.

5 Zouboulis CC, Desai N, Emtestam L *et al.* European S1 guideline for the treatment of hidradenitis suppurativa/acne inversa. *J Eur Acad Dermatol Venereol* 2015; **29**:619–44.

6 Alavi A, Kirsner RS. Local wound care and topical management of hidradenitis suppurativa. *J Am Acad Dermatol* 2015; **73** (5 Suppl. 1):S55–61.

7 Fitzsimmons JS, Guilbert PR, Fitzsimmons EM. Evidence of genetic factors in hidradenitis suppurativa. *Br J Dermatol* 1985; **113**:1–8.

8 Sartorius K, Emtestam L, Jemec GB, Lapins J. Objective scoring of hidradenitis suppurativa reflecting the role of tobacco smoking and obesity. *Br J Dermatol* 2009; **161**:831–9.

9 von der Werth JM, Williams HC. The natural history of hidradenitis suppurativa. *J Eur Acad Dermatol Venereol* 2000; **14**:389–92.

10 Sartorius K, Killasli H, Oprica C *et al.* Bacteriology of hidradenitis suppurativa exacerbations and deep tissue cultures obtained during carbon dioxide laser treatment. *Br J Dermatol* 2012; **166**:879–83.

11 Ingram JR, Woo PN, Chua SL *et al.* Interventions for hidradenitis suppurativa: a Cochrane systematic review incorporating GRADE assessment of evidence quality. *Br J Dermatol* 2016; **174**:970–8.

12 Jemec GB. Clinical practice. Hidradenitis suppurativa. *N Engl J Med* 2012; **366**:158–64.

13 Hessam S, Scholl L, Sand M *et al.* A novel severity assessment scoring system for hidradenitis suppurativa. *JAMA Dermatol* 2018; **54**:330–5.

14 Ingram JR, Hadjieconomou S, Piguet V. Development of core outcome sets in hidradenitis suppurativa: systematic review of outcome measure instruments to inform the process. *Br J Dermatol* 2016; **175**:263–72.

15 Zouboulis CC, Tzellos T, Kyrgidis A *et al.* Development and validation of the International Hidradenitis Suppurativa Severity Score System (IHS4), a novel dynamic scoring system to assess HS severity. *Br J Dermatol* 2017; **177**:1401–9.

16 Hurley HJ. Axillary hyperhidrosis, apocrine bromhidrosis, hidradenitis suppurativa, and familial benign pemphigus: surgical approach. In: *Dermatologic Surgery: Principles and Practice* (Roenigk RK, Roenigk HH, eds). New York: Marcel Dekker, 1989; 729–39.

17 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**:159–74.

18 Rotondi MA. kappaSize: sample size estimation functions for studies of interobserver agreement. Available at: https://cran.r-project.org/package=kappaSize (last accessed 12 February 2019).

19 Kottner J, Audigé L, Brorson S *et al.* Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol* 2011; **64**:96–106.

20 Chiricozzi A, Faleri S, Franceschini C *et al.* AISI: a new disease severity assessment tool for hidradenitis suppurativa. *Wounds* 2015; **27**:258–64.

21 Canoui-Poitrine F, Revuz JE, Wolkenstein P *et al.* Clinical characteristics of a series of 302 French patients with hidradenitis suppurativa, with an analysis of factors associated with disease severity. *J Am Acad Dermatol* 2009; **61**:51–7.

22 van der Zee HH, Jemec GB. New insights into the diagnosis of hidradenitis suppurativa: clinical presentations and phenotypes. *J Am Acad Dermatol* 2015; **73** (5 Suppl. 1):S23–6.

23 Liddy C, Moroz I, Mihan A *et al.* A systematic review of asynchronous, provider-to-provider, electronic consultation services to improve access to specialty care available worldwide. *Telemed J E Health* 2018; https://doi.org/10.1089/tmj.2018.0005.