Interrater agreement and reliability of outcome measurement instruments and staging systems used in hidradenitis suppurativa

Thorlacius, L; Garg, A; Riis, P T; Nielsen, S. M.; Bettoli, V; Ingram, J R; Del Marmol, V; Matusiak, L; Pascual, J C; Revuz, J; Sartorius, K; Tzillos, T; van der Zee, H H; Zouboulis, C C; Saunte, D M; Gottlieb, A B; Christensen, R; Jemec, G B E

Published in:
British Journal of Dermatology

DOI:
10.1111/bjd.17716

Publication date:
2019

Document version:
Accepted manuscript

Citation for published version (APA):

Go to publication entry in University of Southern Denmark's Research Portal

Terms of use
This work is brought to you by the University of Southern Denmark. Unless otherwise specified it has been shared according to the terms for self-archiving. If no other license is stated, these terms apply:
- You may download this work for personal use only.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying this open access version.

If you believe that this document breaches copyright please contact us providing details and we will investigate your claim. Please direct all enquiries to puresupport@bib.sdu.dk

Download date: 31. Oct. 2021
Interrater agreement and reliability of outcome measurement instruments and staging systems used in hidradenitis suppurativa

L. Thorlacius,1,2 A. Garg,3 P. T. Riis,4 S. M. Nielsen,5 V. Bettoli,6 J. R. Ingram,7 V. del Marmol,8 L. Matusiak,9 J. C. Pascual,3 J. Revuz,9 K. Sartorius,10 T. Tzellos,11 H. H. van der Zee,12 C. C. Zouboulis,13 D. M. Saunte,1 A. B. Gottlieb,14 R. Christensen,2,15 G. B. E. Jemec1

1) Department of Dermatology, Zealand University Hospital, Roskilde; Health Sciences Faculty, University of Copenhagen, Denmark.
2) Musculoskeletal Statistics Unit, The Parker Institute, Bispebjerg and Frederiksberg Hospital, The Capital Region of Denmark. Copenhagen, Denmark.
3) Department of Dermatology, Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, USA
4) Dermatology and Venereology, University of Ferrara, Italy
5) Department of Dermatology & Academic Wound Healing, Cardiff University, Cardiff, U.K.
6) Department of Dermatology, Hospital Erasme, Brussels, Université Libre de Bruxelles, Belgium
7) Department of Dermatology, Venereology and Allergology, Medical University, Wroclaw, Poland
8) Department of Dermatology, Hospital General Universitario de Alicante, Alicante, Spain
9) 11 chaussée de la mouette, 75016 Paris, France
10) Department of Dermatology, Stockholm South General Hospital (Södersjukhuset) and Department of Clinical Science and Education, Karolinska Institutet, Stockholm, Sweden
11) Department of Dermatology, Faculty of Health Sciences, University Hospital of North Norway, Troms, Norway
12) Department of Dermatology, Erasmus University Medical Center, Rotterdam, the Netherlands

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/BJD.17716

This article is protected by copyright. All rights reserved
What’s already known about this topic?

- Without valid and reliable instruments to measure outcomes, researchers and clinicians lack the necessary benchmarks to assess primary and secondary endpoints of interventional trials properly.
- Hidradenitis suppurativa (HS) is a chronic, inflammatory skin disease.
- Several outcome measure instruments exist for HS, but their validation is generally incomplete or of relatively low methodological quality.

What does this study add?
• In a prospective completely balanced design the study examined interrater reliability with HS-experienced dermatologists as the rater population of interest.
• The study did not find very good reliability for any included instrument or lesion counts.
• The study illustrates how difficult it is even for experts to agree on the type and number of lesions.
• The results question whether physical signs are best measured by a traditional physician lesion count instrument.

ABSTRACT

Background: Monitoring disease activity over time is a prerequisite for clinical practice and research. Valid and reliable outcome measurements instruments (OMIs) and staging systems provide researchers and clinicians with benchmark tools to assess the primary and secondary outcomes of interventional trials and to guide treatment selection properly.

Objectives: To investigate interrater reliability and agreement in instruments currently used in Hidradenitis Suppurativa (HS) with HS-experienced dermatologists being the rater population of interest.

Methods: In a prospective completely balanced design, 24 HS patients underwent a physical examination by 12 raters (288 assessments) using nine instruments; analysed using generalised linear mixed models.

Results: For the staging systems, the study found good interrater reliability for Hurley staging in the axillae and gluteal region, moderate interrater reliability for Hurley staging in the groin and Physician’s Global Assessment, and fair interrater reliability for Hurley staging refined and International HS Severity Scoring System. For all the tested OMIs, the observed intervals for limits of agreement were very wide relative to the ranges of the scales.

Conclusion: The very wide intervals for limits of agreement imply that substantial changes are needed in clinical research in order to rule out measurement error. The results illustrate a difficulty, even for experienced HS experts, to agree on the type and number of lesions when evaluating disease severity. The apparent caveats call for global efforts, such as the Hidradenitis Suppurativa cORe outcomes set International Collaboration (HISTORIC) to reach consensus on how valid HS physical signs are best measured reliably in randomised trials.

Measuring health outcomes is a process in which a standardised attempt is made to observe a complex clinical picture. Hidradenitis Suppurativa (HS) is a complex, chronic, inflammatory skin disease with an estimated worldwide prevalence of 0.1-4%. The disease is characterised by painful, inflamed nodules in
the axillae, inguinal region, buttocks and inframammary region. These nodules may progress to form abscesses, tunnels (sinus tracts) and scarring.\textsuperscript{5,6} Interventions for HS are diverse and include medical and surgical therapy.\textsuperscript{7} The beneficial effects of biologics can be dramatic for some patients, but the overall response rate of the currently available biologics is around 60%,\textsuperscript{8} suggesting a continued large unmet need of treatment and, accordingly, a particular need for more trials on interventions for HS.

Evidence-based and consensus-endorsed outcome measurement instruments (OMIs) are required for the synthesis of trial results. However, OMIs in HS trials are diverse and abundant with 30 different OMIs identified in only 12 randomized controlled trials.\textsuperscript{9} This lack of consensus limits the possibility to perform evidence synthesis and may give rise to outcome reporting bias.\textsuperscript{11} A core outcome set (COS), which is an agreed minimum set of outcomes to be included in all clinical trials of a specific disease or trial population, has, therefore, been proposed for HS.\textsuperscript{12-17}

Once a COS is defined, OMIs should be selected based upon the measurement properties of the OMIs and consensus.\textsuperscript{18} For OMIs in HS, however, the validation has generally been found to be incomplete or of relatively low methodological quality.\textsuperscript{9} Interrater reliability and agreement are important measurement properties for physician-based OMIs. Nevertheless, these measurement properties have never been systematically investigated in HS, with only two studies investigating the interrater reliability of two different versions of the Sartorius Score.\textsuperscript{9,19,20} The methodological quality of both these studies was graded as ‘poor’ in a recent review.\textsuperscript{9}

For staging systems used to guide treatment selection in HS, interrater reliability and agreement are also unknown. We therefore aimed to investigate interrater reliability and agreement of both the most frequently used OMIs and staging systems in HS.

**METHODS**

The study was designed as a prospective study with a pre-specified protocol (see Supporting Information), and is reported in accordance with the ‘Guidelines for Reporting Reliability and Agreement Studies’ (GRRAS).\textsuperscript{32} The national Danish ethics committee was consulted and confirmed that no official ethical approval was required.

*Instruments under scrutiny*

Nine different instruments were included (Table 1, Figure S1). The instruments included recently-developed instruments and composite scores.

The first HS severity scoring system was supposed in 1989 by Hurley.\textsuperscript{21} The instrument classifies patients into three severity stages according to physical signs and was originally designed for
treatment selection.\textsuperscript{21} Hurley staging refined was recently suggested by a Dutch group of HS researchers for severity assessment and treatment guidance.\textsuperscript{22} The Physician’s Global Assessment (HS-PGA) is a lesion count based scale developed and used as an OMI in a phase II adalimumab trial.\textsuperscript{23} The instrument also classifies in stages from clear to very severe.

A physician rated global HS severity Visual Analog Scale (VAS) was also included although it has only been used once previously.\textsuperscript{24} We did this to have a global anchor and to test how the instrument would perform, knowing that the physician global assessment was a potential core domain for the HS COS.\textsuperscript{14}

The International HS Severity Scoring System (IHS4) was developed through expert discussion and Delphi exercise by the European HS Foundation (EHSF) in addition to a multicentre prospective study.\textsuperscript{25} IHS4 is a lesion count based instrument aiming to serve both as an OMI and a severity staging system.

The Sartorius score was first suggested in 2003\textsuperscript{26} and was later modified as the tested Modified Sartorius Score (MSS).\textsuperscript{27} MSS is an OMI incorporating different items within the domain of physical signs (lesion count, region count, surface area and Hurley staging).

The HS Clinical Response (HiSCR) was developed retrospectively from phase II adalimumab trial data.\textsuperscript{28} The HiSCR is a lesion count based OMI. We tested a 'baseline assessment', defined as the sum of abscesses and inflammatory nodules, AN.

The two included composite HS severity scores were the Acne Inversa Severity Index (AISI), recently suggested by an Italian group,\textsuperscript{29} and the HS Severity Index (HSSI), used in two publications on infliximab and adalimumab for HS.\textsuperscript{30,31}

\textit{Power and sample size considerations}

In a two one-sided tests (TOST) analysis for additive equivalence of paired means with bounds -5 and 5 HiSCR-units for the mean difference of 0 and a significance level of 0.05, a common standard deviation of 8 and correlation between pairs of 0.5, a sample size of 24 pairs yields a power of 0.815. The assumed correlation (between pairs) of 0.5 was based on a conservative estimate; in comparison, if we assumed that the correlation between measures was higher, the statistical power would increase as well (e.g. correlation = 0.6 → power = 90.6\%). For the other OMIs we pragmatically assumed that a sample size of 24 pairs would be sufficient. As described by Donner & Rotondi (2010), sample size requirements that achieve a prespecified expected lower limit for a 95\% confidence interval (95\%CI) for the case of multiple raters are available.\textsuperscript{33} For example, suppose that we want to assure with 95\% confidence that a substantial level of interobserver agreement is achieved (e.g. reliability = 0.8) then the number of paired samples (participants)
required to achieve a value of no less than 0.6 (i.e. both the point estimate as well as the precision around it), representing good reliability, would be 24 patients (Figure S2).

**Sampling method and randomisation**

HS patients were identified from the Department of Dermatology, Zealand University Hospital, Roskilde, Denmark’s the outpatient clinic. Patients with a clinically confirmed diagnosis of HS\(^5\), older than 18, were eligible. To ensure the inclusion of all severities, patients were divided in three groups based on the reported number of boils in their HISREG (Register for Hidradenitis Suppurativa\(^3\)) questionnaire\(^4\) in the last four weeks (group 1: 0–4 reported boils, group 2: 5–9 reported boils and group 3: 10 or more reported boils). Questionnaires from the last six months before the sampling were used. Four patients from group 1, 10 patients from group 2 and 10 patients from group 3 were recruited at random. In addition, four group 2 patients were recruited to participate as a substitute in case of non-attendance among the original participants.

Raters were Dermatologists from different countries with a clinical background of at least 10 years of experience with HS and documented interest in the disease by peer-reviewed publications on HS.

Patients and raters were randomised based on a computer-generated list of random numbers, with a varying block size from 3 to 4. The order in which each rater met each patient was based on a prespecified patient-by-rater visit algorithm (see Supporting Information).

**Rating process**

The ratings took place at the outpatient clinic at the Department of Dermatology, Zealand University Hospital, Roskilde, Denmark. The study protocol (see Supporting Information) including the tested instruments was sent to all raters by email one week before the study. Before the ratings, all raters were gathered for an introductory session where the individual instruments, ratings and rotation were reviewed. It was reviewed in detail how scoring had to be carried out for each instrument, but no joint bedside or classroom training were performed.

Each patient was then placed in separate rooms and the raters walked from room to room on rotation in two rounds. All rooms had sufficient lighting and a standard examination couch. The patient reported outcome measures needed for the composite scores were filled out before the ratings and were thus identical for all raters. Raters received information on the patients’ age and gender but not on health status, identity or history. Raters were aware that their assessments would be compared with those of others. They were not allowed to discuss their scores with other raters during the process. Rotation was set

\(^1\) https://unn.no/fag-og-forskning/medisinske-kvalitetsregistre/hisreg-register-for-hidradenitis-suppurativa

This article is protected by copyright. All rights reserved
to take place every tenth minute, but prolongation was possible.

The raters noted their scores on scoring sheets, with each of the instruments on a separate page. The sheets were shuffled like a stack of cards before assembly, randomising the sequence from round to round. Exceptions were the lesion count sheet (located second to last, so that raters would not reuse this information for individual scores) and the physician global VAS (located last as a summary score).

**Statistical analysis**

For the quantification of the interrater reliability, the intra-class correlation coefficient (ICC) type ICC (2,1) was used. ICC (2,1) is based on a two-way random effects model estimating agreement, assuming the raters to be a random sample from the population of raters and each rater assess all of the subjects, and the reliability of a single rater’s assessment being of interest.\(^{35}\) ICC was applied to all data, including the ordinal data, since weighted kappa for ordinal data can be interpreted as an ICC (2,1), and the ICC (2,1) can be used for multiple raters.\(^{36}\) Results were interpreted using the Altman’s Kappa Benchmark Scale (values of 0–0.20 represent poor reliability, 0.21–0.40 represent fair reliability, 0.41–0.60 represent moderate reliability, 0.61–0.80 represent good reliability and 0.81–1.00 represent very good reliability).\(^{37}\) In order to critically interpret the observed reliability, the lower limits of the 95% CI’s were used for interpretation, rather than the point estimate, to respect the limited sample size of 24 patients.\(^{38}\)

For quantification of interrater agreement for variables with >4 categories, a modified Bland-Altman method was applied. The Bland-Altman method provides insight into the distribution of differences in relation to mean values. In the original Bland-Altman approach, agreement is quantified by calculating the mean difference between two sets of observations and the standard deviation (SD) for this difference. We applied a modified form that allows multiple raters, by calculating the mean score for each patient (x-axis) and the differences from the mean score for each patient (y-axis).\(^{39}\) The 95% limits of agreement were defined as ±1.96× SD, where the SD is the SD of the differences from the mean scores.

For variables with ≤4 categories, the observed agreement was calculated as the number of patients for which all the raters agreed.

The minimal detectable change (MDC) was calculated from SEM×1.96×\(\sqrt{2}\), where SEM was calculated from SD×\(\sqrt{(1-ICC)}\), and here the SD is the SD of all scores from all subjects.\(^{40}\)

Reliability and agreement calculations were based on the assumption of normally distributed data, even though not all data followed a strict normal distribution. This aspect was investigated with a preliminary sensitivity analysis showing no significant impact on our conclusions. Further sensitivity analyses included imputing missing values with the grand mean of all scores, and complete cases analysis, i.e., removing all subjects with missing values for any of the raters (Table S1 and S2).
All calculations were carried out using the statistical software R (version 3.2.3)\textsuperscript{41} with the package “psych”\textsuperscript{42} and “irr”.\textsuperscript{43}

RESULTS

Twelve raters from different countries (Author no. 2, 5-14, 18) and 24 patients were included. Patient characteristics are found in Table 2. The total number of replicate observations conducted was 288 for each instrument as well as the lesion counts. Percentages of missing data are found in Table 3 and a description of how missing data were handled is found in statistics. Only data from 11 raters were available for Hurley staging refined, since a scoring sheet was lost.

Interrater reliability and agreement

Table 3 shows the interrater reliability and agreement results, and Figure 1 shows two examples of the modified Bland-Altman plots. The remaining plots are provided as Figures S3 and S4. Intervals for limits of agreement are shown on all plots and in Table 3.

Reliability and agreement are conceptually different parameters of reproducibility. Agreement assesses how close the results of the repeated measurements are, i.e., the measurement error; while reliability assesses how well patients can be distinguished from each other despite measurement error.\textsuperscript{44,45} Agreement predominantly reflect instrument characteristics, while reliability depends on the heterogeneity of the study sample. Therefore, Reliability is more appropriate for distinguishing purposes (for classification/staging), while agreement is preferred to measure changes in health status (for OMIs).\textsuperscript{44}

The purpose of the Hurley staging system, Hurley staging refined and the 3-component part of IHS4 is classification, meaning that the reliability results are the most important parameters for these instruments. For the axillae and gluteal regions, Hurley staging showed good interrater reliability, but for the groin the results were only moderate. Fair interrater reliability was found for Hurley staging refined and IHS4 classification. HSSI showed good ICC results, but the score is a composite score and HSSI is not designed for staging.

In contrast, the main purpose of MSS, HiSCR, HS-PGA, HSSI, IHS4 total and AISI is to measure changes in health status, implying that the agreement results are of most importance for these instruments. In general, our observed intervals for limits of agreement were very wide compared to the ranges, or the usual ranges, of the scales. For example, the MSS and HiSCR baseline AN limits of agreement were $\pm 75.25$ and $\pm 12.28$, respectively. The low agreement between the raters is also reflected in the Minimal Detectable Change (MDC) results. MSS and HiSCR baseline AN MDC were 110.71 and 17.85, respectively, implying that substantial changes would be needed in order to rule out measurement error.
over real change.

**Subgroup analysis for different severity groups**

The above results indicate a difficulty in reliably evaluating disease severity in HS by existing instruments and by individual lesion count. Logically, this would appear particularly evident in patients with numerous and more confluent lesions. Therefore, we assessed whether interrater reliability and agreement differed between patients with mild or moderate disease compared to those with severe disease in a subgroup analysis (Table S3 and S4). The intervals for limits of agreement are indeed wider in groups 2 and 3 compared with group 1 (most evident for MSS and HiSCR baseline AN), implying that the measurement error does increase with increasing HS severity.

**DISCUSSION**

In this interrater reliability and agreement study, 24 HS patients underwent a physical examination by 12 HS experienced raters using nine instruments. For the staging systems, we found good interrater reliability for Hurley staging in the axillae and gluteal region and moderate results in the groin. Accordingly, Hurley staging seems to be an acceptable instrument for classification in HS, with ICC results similar to those previously reported for e.g. atopic eczema and acne instruments.46,47 The physician global VAS is not a formally developed instrument for HS staging, but with an ICC of 0.72 (95%IC 0.59–0.84), the instrument may also be suited for distinguishing purposes.

In the literature, ICC results are frequently reported and highlighted, even for OMIs, although experts in field emphasise that agreement parameters are preferred when evaluating instruments designed to measure changes in health status.44,45 As parameters of agreement are always expressed in the unit of measurement, it is not possible to find general guidance about what values are acceptable.45 Nevertheless, our observed limits of agreement intervals were very wide compared to the ranges, or the usual ranges, of the OMI scales. As an example, the limits of agreement for MSS was ± 75.25, implying that even with a change of e.g. 50 in MSS, there is a reasonable chance that this change could be due to measurement error. For the remaining included OMIs, a similar excise results in the same overall picture. Therefore, we cannot immediately recommend any of the included OMIs for measuring physical signs in HS under the given conditions.

We used HS experienced dermatologist as raters. While the individual scoring systems were reviewed in an introductory session, the training did not include a bedside training session.19 This design was chosen to evaluate `real-world’ interrater reliability and agreement among experts. However, it would be valuable for future studies to investigate if a training session can improve interrater reliability and
agreement.

For logistical reasons, we could not examine intrarater reliability and agreement (repeated assessments by the same rater). Nevertheless, intrarater reliability could be speculated to be higher than the interrater reliability based on previous reported test-retest reliability for HiSCR AN count.\textsuperscript{28} Accordingly, we, for now, recommend that baseline and endpoint assessments are done by the same rater if lesion count based instruments are used as endpoints in HS trials.

Limitations to this study includes, firstly, the assumption of normally distributed data in reliability and agreement calculations, as described in the statistics section. Secondly, all patients were from a single European centre, and most patients were Caucasians. Thirdly, the study was designed exclusively to investigate reliability and agreement, and slightly different results may be obtained in studies designed to investigate interventional effects.

The HS COS consensus study established that physicals signs should be measured, as one of the core domains, in all future clinical trials for HS.\textsuperscript{12} Accordingly, the HIdradenitis SuppurataTiva cORe outcomes set International Collaboration (HISTORIC) will have to select the most suited OMI for measuring physical signs. This process involves an evaluation of all measurement properties, feasibility aspects and a consensus process,\textsuperscript{18} and this falls outside the scope of this study. Nevertheless, our results illustrate how difficult it is, even for experts, to measure HS physical signs of HS by a simple rater counting and, therefore, question whether HS physical signs are best measured by a traditional physician lesion count instrument. Consequently, other assessment methods of physical signs, such as ultrasound evaluation, do need consideration for the HS COS.

In conclusion, in terms of interrater agreement, we cannot recommend any of the included OMIs when used without a preceding bedside training session. More research is needed to examine intrarater agreement and to examine if interrater agreement can be improved e.g. by training. For staging, Hurley staging and a physician global VAS proved to be acceptable instruments.

ACKNOWLEDGEMENT

We thank all the included patients and raters for their willingness to participate, time and patience and all others who voluntarily helped with running the study. We thank IDEOM for additional funding to cover raters’ and patients’ travel expenses and patient gifts.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper.
FIGURE TITLES & LEGENDS

Fig 1. Examples of modified Bland-Altman plots

Agreement plots for and Modified Sartorius Score (MSS) and Hidradenitis Suppurativa Clinical Response (HiSCR) baseline inflammatory lesion count (sum of abscesses and inflammatory nodules, AN). Horizontal lines indicate the lower and upper 95% limits of agreement and the line of no difference. Rates are presented by different symbols. Please note some symbols are superimposed on top of each other.

REFERENCES


This article is protected by copyright. All rights reserved


18 Prinsen CA, Vohra S, Rose MR *et al.* How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. *Trials* 2016; **17**: 449.


38 Bland JM. The tyranny of power: is there a better way to calculate sample size? *BMJ (Clinical research ed.)* 2009; 339: b3985.


43 Gamer M. Various Coefficients of Interrater Reliability and Agreement. R package version 0.84. [https://cran.r-project.org/package=irr](https://cran.r-project.org/package=irr) 2012.


### TABLES

**Table 1** Instruments under scrutiny

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Range</th>
<th>Characteristics</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hurley staging system</td>
<td>(0)-3*</td>
<td>Classifies patients into three severity stages according to physical signs</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Designed to be used per region separately (one stage per region)</td>
<td>HS severity classification/staging Treatment guidance</td>
</tr>
</tbody>
</table>
Hurley staging refined 1-7 Classifies patients into seven severity stages according to physical signs Designed to be used on the patient globally (one stage per patient)

HS-PGA 0-5 Lesion count Physical signs OMI HS severity classification/staging
Physician global VAS 0-10 Overall HS severity grading by physician Physician global assessment OMI HS severity classification/staging

IHS4 0-unlimited/1-3 Lesion count Physical signs OMI/Severity classification/staging
MSS 0-unlimited Lesion and region count plus surface area determination Physical signs OMI
Incorporates Hurley staging
HiSCR (baseline AN tested) 0-unlimited Lesion count Physical signs OMI
HSSI 0-19 Lesion and region count plus surface area determination HS severity OMI (combined physical signs, drainage and pain)
Incorporates patient reported measures for drainage and pain
AISI 0-unlimited Lesion identification and region count plus patient global VAS HS severity OMI (combined physical signs and patient global assessment)

*The Hurley staging system does not include any clear stage in the original version, but in this study ‘0’ was added to represent normal judged skin in the area.

HS-PGA, The Physician’s Global Assessment; OMI, Outcome Measurement Instrument; VAS, Visual Analog Scale; IHS4, International Hidradenitis Suppurativa Severity Scoring System; MSS, Modified Sartorius Score; HiSCR baseline AN, The Hidradenitis Suppurativa Clinical Response baseline sum of abscesses and inflammatory nodules; HSSI, HS severity index; AISI, Acne Inversa Severity Index

Table 2 Baseline characteristics

<table>
<thead>
<tr>
<th>Variables</th>
<th>Patients (n=24)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td></td>
</tr>
<tr>
<td>Age, years</td>
<td>42.7 (11.9)</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>15 (63%)</td>
</tr>
</tbody>
</table>

This article is protected by copyright. All rights reserved
Present nicotine users, n (%) 16 (67%)
Former and present nicotine users, n (%) 23 (96%)
Weight, kg 106 (33.5)
BMI, kg/m² 34.7 (9.5)
<25 kg/m², n (%) 5 (21%)
25 to 30 kg/m², n (%) 3 (13%)
≥30 kg/m², n (%) 16 (67%)
Disease duration*, years 20.4 (11.7)
Receiving oral clindamycin/rifampicin or tetracycline, n (%) 7 (29%)
Former or present biological treatment, n (%) 8 (33%)
Number of boils last 4 weeks* 8.7 (6.2)
0-4, n (%) 5 (21%)
4-9, n (%) 10 (42%)
10 or more, n (%) 9 (38%)
VAS-measured skin pain (0-10), cm 4.0 (2.7)
DLQI score (0-30) 13.3 (7.6)
0-9, n (%) 10 (42%)
10-19, n (%) 7 (29%)
20-30, n (%) 7 (29%)
Instruments†
HSSI (0-19), median (IQR), 12 (10;16)
Hurley axillae‡ (0-1), median (IQR) 1 (0;2)
Hurley gluteal‡ (0-1), median (IQR) 0 (0;1)
Hurley groin‡ (0-1), median (IQR) 2 (0;2)
IHS4 total§ (0-unlimited), median (IQR) 8 (3;22)
IHS4 classification§ (1-3), median (IQR) 2 (1;3)
Physician global VAS (0-10), median (IQR) 4.3 (3.0;7.0)
MSS (0-unlimited), median (IQR) 43 (20;80)
HS-PGA (0-5), median (IQR) 3 (2;4)
Hurley staging refined (1-7), median (IQR) 5 (2;7)
HiSCR baseline AN (0-unlimited), median (IQR) 4 (1;11)
AISI (0-unlimited), median (IQR) 21.0 (14.4;35.1)
Lesion counts†
Total tunnel, median (IQR) 2 (0;6)
Draining tunnel, median (IQR) 0 (0;2)
Total nodules, median (IQR) 5 (2;13)
Inflammatory nodules, median (IQR) 3 (1;10)
Non-draining tunnel, median (IQR) 1 (0;2)
Non-inflammatory nodules, median (IQR) 1 (0;3)
Abscesses, median (IQR) 0 (0;1)
Data is presented as mean (SD) unless otherwise stated.

* Missing values for one patient
† Medians are grand medians, i.e. of all observations from all raters. The amount of missing data is presented in Table 2.
‡ Hurley was rated separately for left and right region, but the observations were collected in the same analysis. Hurley staging does not include any clear stage in the original version, but in this study ‘0’ was added to represent normal judged skin in the area.
§ The final version of IHS4 was agreed upon after the ratings took place, scores were therefore calculated from the lesion counts.

SD, standard deviation; BMI, Body Mass Index; VAS, Visual Analog Scale; DLQI, Dermatology Life Quality Index; IQR, interquartile range; HSSI, HS severity index; IHS4, International Hidradenitis Suppurativa Severity Scoring System; MSS, Modified Sartorius Score; HS-PGA, The Physician’s Global Assessment; AISI, Acne Inversa Severity Index; HiSCR baseline AN, The Hidradenitis Suppurativa Clinical Response baseline sum of abscesses and inflammatory nodules.

Table 3: Reliability and agreement for included instruments and total lesion counts

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of patients examined, n</th>
<th>Number of raters, n</th>
<th>Observed agreement, n pts (%)</th>
<th>ICC (95%CI)</th>
<th>MDC % of scale</th>
<th>% of missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification instruments:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hurley axillae† ((0)-1-3)</td>
<td>24</td>
<td>12</td>
<td>15 (63%)</td>
<td>0.72 (0.63 to 0.81)</td>
<td>1.61</td>
<td>40.3</td>
</tr>
<tr>
<td>Hurley gluteal† ((0)-1-3)</td>
<td>24</td>
<td>12</td>
<td>10 (42%)</td>
<td>0.72 (0.62 to 0.80)</td>
<td>1.59</td>
<td>39.8</td>
</tr>
<tr>
<td>Hurley groin† ((0)-1-3)</td>
<td>24</td>
<td>12</td>
<td>3 (13%)</td>
<td>0.55 (0.44 to 0.67)</td>
<td>1.97</td>
<td>49.3</td>
</tr>
<tr>
<td>IHS4 classification ‡ (1-3)</td>
<td>24</td>
<td>12</td>
<td>1 (4%)</td>
<td>0.54 (0.39 to 0.71)</td>
<td>1.58</td>
<td>52.7</td>
</tr>
<tr>
<td>Hurley staging refined (1-7)</td>
<td>24</td>
<td>11§</td>
<td>-2.86 to 2.86*</td>
<td>0.51 (0.35 to 0.68)</td>
<td>4.29</td>
<td>61.3</td>
</tr>
<tr>
<td>Classification/OMIs:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician global VAS‡ (0-10)</td>
<td>24</td>
<td>12</td>
<td>-2.53 to 2.53*</td>
<td>0.72 (0.59 to 0.84)</td>
<td>3.69</td>
<td>33.5</td>
</tr>
<tr>
<td>HS-PGA (0-5)</td>
<td>24</td>
<td>12</td>
<td>-1.28 to 1.28*</td>
<td>0.64 (0.50 to 0.79)</td>
<td>1.86</td>
<td>31.0</td>
</tr>
<tr>
<td>Physical signs OMIs:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSS (0-unlimited)</td>
<td>24</td>
<td>12</td>
<td>-75.25 to 75.25*</td>
<td>0.66 (0.52 to 0.80)</td>
<td>110.71</td>
<td>19.5</td>
</tr>
<tr>
<td>IHS4 totals (0-unlimited)</td>
<td>24</td>
<td>12</td>
<td>-36.47 to 36.47*</td>
<td>0.47 (0.32 to 0.65)</td>
<td>53.08</td>
<td>26.3</td>
</tr>
<tr>
<td>HiSCR baseline AN (0-unlimited)</td>
<td>24</td>
<td>12</td>
<td>-12.28 to 12.28*</td>
<td>0.44 (0.29 to 0.63)</td>
<td>17.85</td>
<td>38.0</td>
</tr>
<tr>
<td>Composite OMIs:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSSI (0-19)</td>
<td>24</td>
<td>12</td>
<td>-3.75 to 3.75*</td>
<td>0.78 (0.66 to 0.88)</td>
<td>5.40</td>
<td>27.0</td>
</tr>
<tr>
<td>AISI (0-unlimited)</td>
<td>24</td>
<td>12</td>
<td>-41.35 to 41.35*</td>
<td>0.40 (0.25 to 0.59)</td>
<td>60.09</td>
<td>31.4</td>
</tr>
<tr>
<td>Lesion counts:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total tunnel</td>
<td>24</td>
<td>12</td>
<td>-9.79 to 9.79*</td>
<td>0.55 (0.41 to 0.72)</td>
<td>14.26</td>
<td>20.4</td>
</tr>
<tr>
<td>Draining tunnel</td>
<td>24</td>
<td>12</td>
<td>-7.69 to 7.69*</td>
<td>0.46 (0.31 to 0.64)</td>
<td>11.20</td>
<td>26.0</td>
</tr>
<tr>
<td>Total nodules</td>
<td>24</td>
<td>12</td>
<td>-15.31 to 15.31*</td>
<td>0.44 (0.29 to 0.62)</td>
<td>22.29</td>
<td>30.5</td>
</tr>
<tr>
<td>Inflammatory nodules</td>
<td>24</td>
<td>12</td>
<td>-12.54 to 12.54*</td>
<td>0.40 (0.26 to 0.59)</td>
<td>18.24</td>
<td>38.8</td>
</tr>
<tr>
<td>Non-draining tunnel</td>
<td>24</td>
<td>12</td>
<td>-6.42 to 6.42*</td>
<td>0.28 (0.17 to 0.47)</td>
<td>9.40</td>
<td>33.6</td>
</tr>
<tr>
<td>Non-inflammatory nodules</td>
<td>24</td>
<td>12</td>
<td>-8.59 to 8.59*</td>
<td>0.20 (0.10 to 0.36)</td>
<td>12.58</td>
<td>25.7</td>
</tr>
</tbody>
</table>

This article is protected by copyright. All rights reserved
| Abscesses | 24 | 12 | -3.01 to 3.01* | 0.07 (0.02 to 0.17) | 4.38 | 39.8  | 0.3 |

*Limits of agreement, because the variable has >4 categories

†Hurley was rated separately for left and right region, but the observations were collected in the same analysis.

‡ The final version of IHS4 was agreed upon after the ratings took place, scores were therefore calculated from the lesion counts.

§ A Hurley staging refined scoring sheet was lost for one rater.

¶ Since no maximum were defined for these instruments, the calculations were based on the observed maximum values from the data.

**Bold** emphasises what is most important for staging systems (reliability) vs. OMI (agreement) as described in the main text.

OMIs, Outcome Measurement Instruments; HSSI, HS severity index; HS4, Hidradenitis Suppurativa Severity Scoring System; VAS, Visual Analog Scale; MSS, Modified Sartorius Score; HS-PGA, The Physician’s Global Assessment; HiSCR baseline AN, The Hidradenitis Suppurativa Clinical Response baseline sum of abscesses and inflammatory nodules; AISI, Acne Inversa Severity Index; ICC, Intraclass Correlation Coefficient; MDC, Minimal Detectable Change.