Hidradenitis suppurativa

Rondags, Angelique

DOI:
10.33612/diss.119123035

5

THE REFINED HURLEY CLASSIFICATION: THE INTERRATER AND INTRARATER RELIABILITY AND FACE VALIDITY

Published in adapated form in
British Journal of Dermatology. 2019 Dec; 181(6): 1335–1337. doi: 10.1111/bjd.18235

Angelique Rondags[1,*]
Lisette M. Prens[1,*]
Rob J. Volkering[1]
Ineke C. Janse[2]
Klaziena Politiek[1]
Yolinde S. Zuidema[1]
Iana Turcan[1]
Jelmer R. van Hasselt[1]
Hessel H. van der Zee[3]
Barbara Horváth[1]

**both authors contributed equally*

1. Department of Dermatology, University of Groningen University Medical Center Groningen Groningen, the Netherlands.
2. Department of Dermatology, Meander Medical Center, Amersfoort, the Netherlands.
3. Department of Dermatology, Erasmus Medical Center, Rotterdam, the Netherlands.

ABSTRACT

Background*:* Hidradenitis suppurativa (HS) is a heterogeneous, chronic, inflammatory skin disease, usually staged according to the Hurley classification. This classification has some limitations; it can only describe one body region and not classify the whole patient. Therefore, a modification was proposed: the "refined Hurley classification".

Objectives*:* To determine the interrater and intrarater reliability and face validity of the refined Hurley classification.

Methods*:* In this observational study two sub-studies were conducted at the dermatology department of the University Medical Center Groningen. Adult patients with active HS were either clinically assessed in real-life or photographed systematically for digital assessment. The real-life assessment included two groups, each with two independent raters. The digital assessment included one group with ten independent raters.

Results*:* Real-life assessment: 25 patients were assessed: 13 in group 1 and 12 in group 2. The interrater agreement varied from 46.2 to 83.3%, and the interrater reliability ranged from Krippendorff's $\alpha = 0.68$ (95% CI 0.32-0.95) to $\alpha = 0.92$ (95% CI 0.78-1.00).
Digital assessment: 15 digital cases were assessed. The interrater reliability demonstrated $\alpha = 0.74$ (95% CI 0.71-0.78) for the first round and $\alpha = 0.80$ (95% CI 0.77-0.82) for the second round. The intrarater reliability demonstrated a mean α of 0.83 (95% CI 0.78-0.89). The face validity showed scores of 78.7 ± 10.3 and 76.5 ± 9.7, on a scale of 0–100.

Conclusions*:* The refined Hurley classification might be a reliable, useful tool for classifying HS patients.

INTRODUCTION

Hidradenitis suppurativa (HS) is a common, debilitating, chronic inflammatory skin disease with recurrent painful abscesses and nodules in the intertriginous regions of the body such as the axilla and groin.[1,2] In a later stage, formation of sinus tracts and hypertrophic scars can occur.[2] The three stage Hurley classification is used in daily practice to assess severity of HS patients. Although easy to use, it was only intended to describe the symptoms in one anatomical region and guide surgical treatment options, and therefore it is not a valid instrument to classify HS in a whole patient.[3] As HS is a heterogeneous disease, the Dutch HS expert group recently proposed a modification of the Hurley classification aiming to classify the whole HS patient and guide holistic treatment options).[4] In this simple three-step algorithm the presence of sinus tracts, inflammation and number affected body areas are assessed (Figure 1). The refined Hurley consists of seven stages, subdividing each of the first two stages into A (mild), B (moderate) and C (severe) based on the degree of inflammation and extend of the disease. This classification aims for more adequately staging of HS patients in daily clinical practice as well as for scientific purposes, ultimately in order to improve treatment outcomes. Evaluation of the *reliability* of this modified classification is essential. Recently, we showed that the patient reported quality of life (Dermatology Life Quality Index, DLQI) and physician-assessed disease severity (International HS Severity Score System, IHS4) correlated with the mild, moderate and severe categories of the refined Hurley classification indicating the construct validity.[5]

The aim of this study is to assess the interrater, intrarater reliability and face validity of the refined Hurley classification in daily practice.

METHODS

Study design

In order to assess the reliability and validity of the refined Hurley classification two sub-studies were designed, both carried out at the dermatology department of the University Medical Center Groningen (UMCG), a tertiary referral centre for HS. For this study, no medical ethical committee approval is required under Dutch law. The study design follows the proposed "Guidelines for Reporting Reliability an Agreement Studies" (GRASS) guidelines were followed.[6]

Real-life assessment for interrater reliability

Consecutive patients with active HS visiting the UMCG dermatology clinic, who gave written informed consent, were included. Two groups of two raters were formed. Additionally, another rater (BH, a dermatologist), who was involved in developing the refined Hurley classification, assessed all patients as well. This assessment served as the control (i.e. reference standard). The other independent raters were residents in dermatology who regularly see HS patients, and are trained to assess them. First, each rater received a brief training on how to use the refined Hurley classification. For every assessment, the raters filled out a standardized form. Raters were allowed to use the refined Hurley classification flowchart, as is possible in daily practice. The raters were not allowed to discuss their findings with one another during the study. Assessments were performed between May and November 2017.

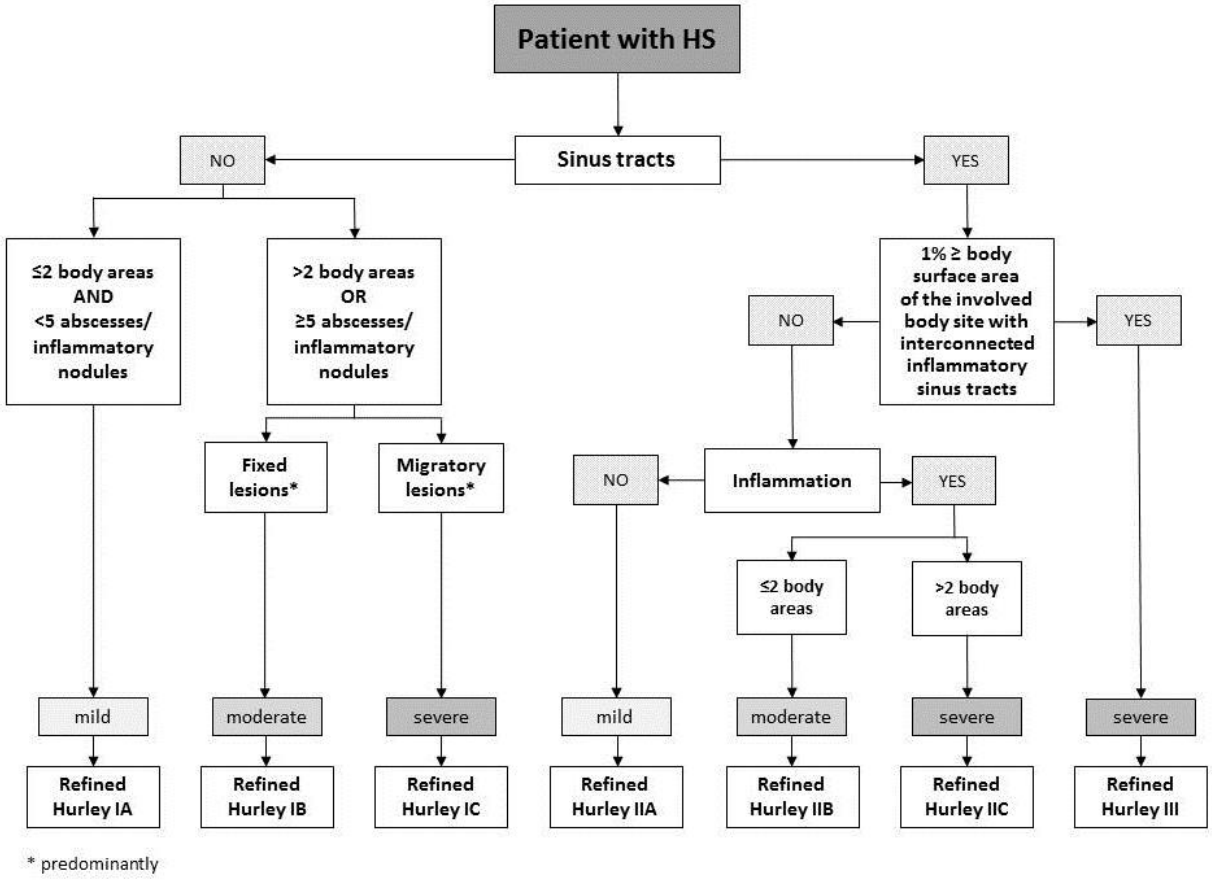

Figure 1. Flowchart refined Hurley classification

Adapted with permission of the Dutch hidradenitis suppurativa expert group.[4]

Available as an app in the Apple App Store and Android play store, search term 'hidradenitis suppurativa app'.

Digital photo assessment for interrater and intrarater analysis

In addition, a digital photo assessment study was designed since HS is a dynamic disease, which makes a real-life assessment for intrarater analysis unsuitable.

All adult HS patients were eligible to participate. After obtaining informed consent, patients were photographed according to a standardized protocol. All photographs were assessed by two independent researchers (LP and AR) for eligibility. At least two patients per refined Hurley stadium were included. A web-based survey was created using Qualtrics (Qualtrics 2018, Provo, Utah). All raters received a brief training, similar to the training for the real-life assessment and use of the flowchart was again permitted (Figure 1). Raters were requested to fill out the same survey twice, with four weeks in between. Each case in the survey started with pictures of the axilla/chest and inguinal area, followed by three questions, according to the refined Hurley flowchart: 'are there sinus tracts present?', 'is active inflammation present?' and 'what refined Hurley stage would you give this patient?'. Twenty dermatologists and dermatology residents at the UMCG Dermatology department were invited (non-committal) to participate. All participating raters filled out the assessment independently. Assessments were performed between April and July 2018.

In order to assess face validity, which is part of the content validity, raters were asked to indicate the usefulness of the refined Hurley classification. The question 'How would you rate the usefulness of the refined Hurley classification on a scale between 0 and 100?', in which a score of 0 (not useful at all) and 100 (very useful) was asked prior to filling in the first digital assessment and after completing the assessment for the second time. Raters were also requested to clarify their choice of scoring by a written response. This was intentionally an open question to avoid bias via prefabricated answers by the authors. This subjective information indicates whether clinicians find this classification a practical tool for staging patients with HS.

Statistical analyses

Continuous variables are presented as mean ± standard deviation (SD). The interrater agreement and reliability were analysed. Interrater agreement determines the percent agreement between raters. For the digital assessment, the interrater agreement was calculated manually for questions with binary outcomes. The Krippendorff's alpha (α) test was used to determine the interrater and intrarater reliability. According to Zapf *et al.*, this test is suitable for interrater reliability analysis in case of nominal data.[7] In this statistical test, both the number of raters and number of refined Hurley stages are taken into account. Furthermore, it is able to deal with missing data.[8] Statistics were performed with IBM SPSS Statistics version 23.0 for Windows (SPSS Inc., Chicago, IL, USA). A specifically designed syntax for the Krippendorff's α test in SPSS was used, as referred to by Zapf *et al.*.[7] Bootstrap level for the confidence intervals was set at 10.000.

The results are reported as α with 95% confidence interval (CI). As reported by Krippendorff, an α >0.8 indicates high reliability and 0.67< α <0.8 suggests moderate reliability.[8]

RESULTS

Real-life assessment

In total, 13 patients were included in the first group of raters and 12 in the second group. An overview of patient characteristics per group is shown in Table 1. Group 1 (n = 13) consisted of 7 males and 6 females, with an average age of 36.3 ± 13.4 years. Results of the interrater agreement and reliability assessment showed an agreement of 46.2% and α = 0.68 (95% CI 0.32-0.95). Group 2 (n = 12) consisted of 10 females and 2 males, with an average age of 43.1 ± 11.7 years. For this group, the assessments were performed on six different time points in the same month. Results showed an interrater agreement of 83.3% and α = 0.92 (95% CI 0.78-1.00). Next, every individual rater (n=4) was compared to the "reference standard". One rater from the first group displayed low interrater reliability (α = 0.60; 95% CI 0.25-0.90), while the other three raters showed high interrater reliability, with α-values ranging from 0.88 (95% CI 0.65-1.00) to 0.98 (95% CI 0.93-1.00).

Table 1. Patient characteristics real-life assessment

	Group 1 n = 13	**Group 2 n = 12**
Sex (n)		
Female	6	10
Male	7	2
Age (years, mean ± SD)	36.3 ± 13.4	43.1 ± 11.7
Ethnicity (n)		
Caucasian	13	11
Hindustani	0	1
Refined Hurley classification stage (n)		
Refined Hurley IA	2	3
Refined Hurley IB	1	0
Refined Hurley IC	3	3
Refined Hurley IIA	0	1
Refined Hurley IIB	3	0
Refined Hurley IIC	4	5
Refined Hurley III	0	0

Digital photo assessment

In total, 23 patients were screened and photographed for the digital assessment, of which 15 cases were selected for inclusion. The majority was Caucasian with skin type I or II in

86.7% of the patients. Ten raters filled out the survey (residents n=8; dermatologists n=2) and, nine raters completed the survey for both time points. The average interval between both time points was 37.5 ± 17.9 days.

At the first time point, an interrater reliability for the refined Hurley stage of α = 0.74 (95% CI 0.71-0.78) was calculated. The second round showed an interrater reliability of α = 0.80 (95% CI 0.77-0.82). The interrater agreement for the refined Hurley stage for both time points is graphically shown in Figure 2. Interrater agreement and reliability analysis was also performed on the sub questions regarding determination of presence of sinus tracts and presence of inflammation, and showed a mean agreement of 92.0% and α = 0.71 (95% CI 0.56-0.86) and a mean agreement of 86% and α = 0.07 (95% CI -0.28-0.38), respectively, for the first round. For the second round, there were similar findings: for assessing presence of sinus tracts the mean agreement was 92.7% and α = 0.73 (95% CI 0.52-0.88) and for presence of inflammation a mean agreement of 83.3% and α = 0.04 (95% CI -0.29-0.34).

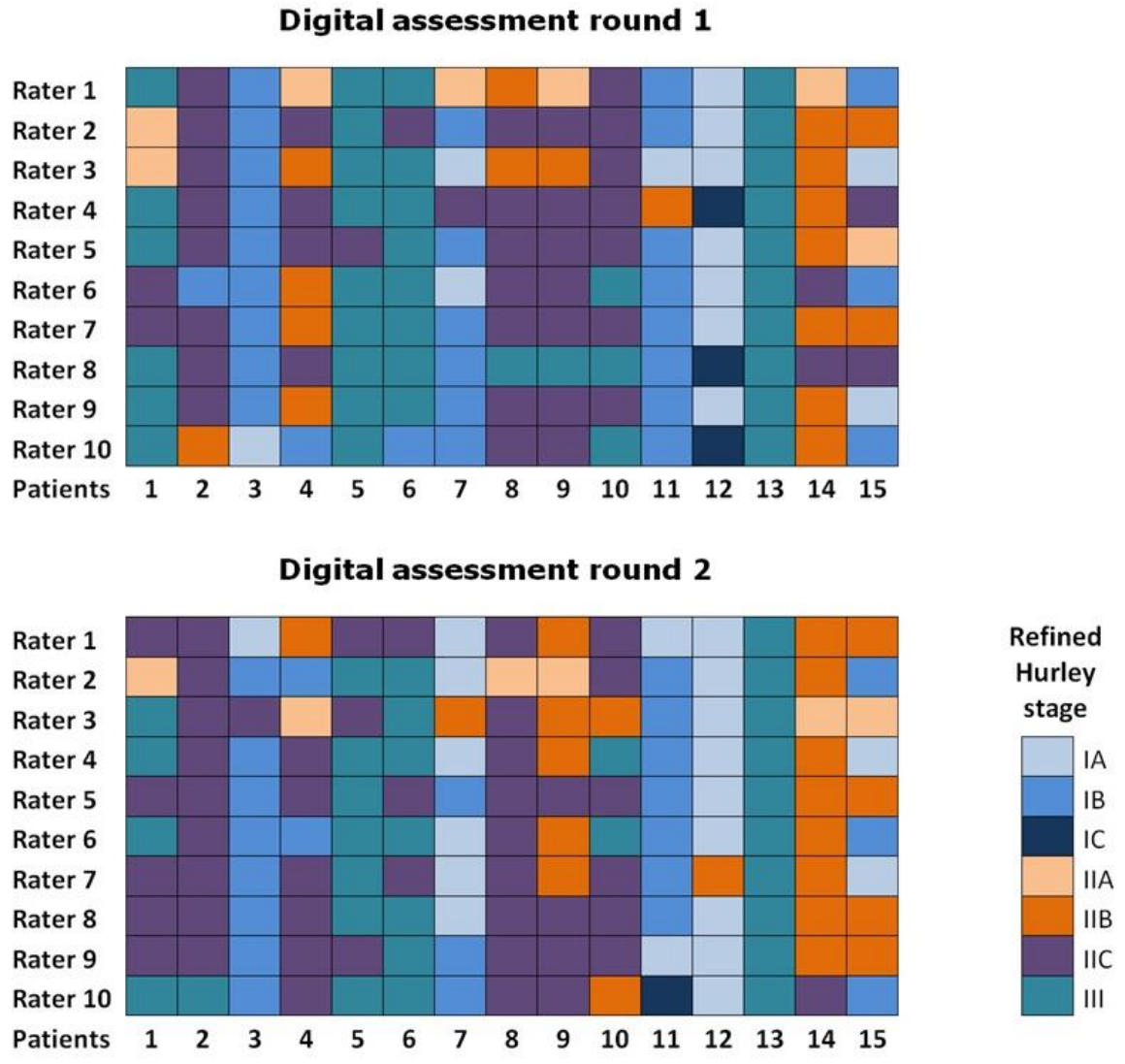

Figure 2. Interrater agreement results of the digital assessment

Intrarater agreement and reliability outcomes showed a mean agreement of 64.4% (range 40-93.3%) and a mean α of 0.82 (95% CI 0.76-0.88) (range 0.68 (95% CI 0.44-0.88) – 0.96 (95% CI 0.90-1.00)), respectively.
The analyses of intrarater agreement and reliability for assessing sinus tracts demonstrated a mean agreement of 91.9% (range 86.7-100%) and a mean α = 0.81 (95% CI 0.72-0.90) and for presence of inflammation a mean agreement of 85.9% (range 73.3-100%) and mean α = 0.14 (95% CI -0.01-0.32).
Regarding face validity, raters scored the usefulness of the refined Hurley classification with a mean of 78.7 ± 10.3 after completing the survey of the first round and a mean score of 76.5 ± 9.7 after the second round. Nine out of 10 raters commented on their choice of score. Before the first round, seven raters commented the classification to be 'clear' or 'useful', two raters found the classification helpful for choosing a treatment strategy. Two raters reported the use of the classification to be time consuming. After the second round, two raters additionally commented on the difficulty to judge presence of erythema on photographs.

DISCUSSION

In this study, we assessed the interrater and intrarater reliability and face validity of the refined Hurley classification. In the real-life assessment, the first group demonstrated low agreement and moderate interrater reliability, while the second group showed high agreement and interrater reliability. The lower outcomes in the first group might be due to the fact that one of the raters showed a low agreement and reliability when compared to the reference standard. The other rater from group one and both raters from group two demonstrated to be in high agreement with the reference standard. In the digital assessment, we found a moderate interrater reliability for the first round. For the second round, the ten raters showed a high agreement for assigning the refined Hurley stadium, possibly indicating the positive learning curve for using the refined Hurley classification.
Additionally, high intrarater reliability for the refined Hurley classification was found. The intrarater reliability cannot be determined easily in real life due to the dynamic nature of HS. Hence, the choice for a digital assessment in which the exact same cases were presented. The intrarater reliability demonstrated high percentages for these assessments. However, scoring HS disease activity on photographs is challenging, which reflects in the low to moderate reliability between the raters for assessing inflammation and presence of sinus tracts. This difference was particularly evident for the assessment of inflammation. A possible explanation for this is, that the variation in answers was very low, over 80% of the answers in both rounds was 'yes'. This lack in variation of answers makes the outcome of

the Krippendorff's alpha test less informative, because the test cannot adequately cope with this.[8] The percent agreement results are therefore said to be more meaningful.[6,7]

The original Hurley classification was first introduced in 1989 and created for surgical purposes. The inflammatory component of the disease and the number of involved anatomic areas are not taken into account. Even though this static classification serves as a standard for severity assessment of HS, only one recent publication studied its interrater and intrarater reliability. A moderate interrater reliability and substantial intrarater reliability was found.[9] This study only used digital photo assessments (n=30), therefore extrapolation of the results to real-life assessments should perhaps be done with care. The refined Hurley classification was very recently tested in a study in which 9 instruments for HS were assessed in 24 live patients by 12 HS experts in one session. A fair interrater reliability was found. One may question if the used study design reliably reflects real-life situations as well, for which the refined Hurley classification is predominantly intended.[10]

Two other classification systems for HS, based on phenotypes, were previously proposed. In 2013 Canoui-Poitrine identified three HS phenotypes by latent class analysis: axillary-mammary type, follicular type and gluteal type.[11] Recently, the interrater reliability of these three phenotypes have been assessed in a digital setting (n=30). The classification demonstrated low interrater reliability with a Fleiss' kappa of only 0.37 (95% CI 0.32-0.42), and therefore may only be of limited use in daily practice.[12] The other phenotype classification was outlined by Van Der Zee et al. based on expert opinion, describing six possible subtypes of HS: the regular type; frictional furuncle type; scarring folliculitis type; conglobata type; syndromic type and ectopic type.[13] This classification has not been validated yet.

Multiple instruments assessing the severity of HS have also been developed over the last decades. These include the Modified Sartorius Score (MSS), the Hidradenitis Suppurativa Clinical Response (HiSCR), and the Hidradenitis Suppurativa Severity Score System (IHS4). However, validation of these instruments is sometimes incomplete or of mainly low methodological quality and perhaps more valuable in research, than for use in daily practice.[10,14–19]

For assessing face validity of the refined Hurley classification, a specific part of the validation process, there are no standards on how it should be assessed. However, 'lack of face validity' is a very strong argument for not using an instrument.[20] In our study, we found a high face validity outcome for the refined Hurley classification, which most likely indicates that the assessors found it useful. However, this finding might be biased for only half of the invited raters participated. We asked our raters to comment on the usefulness of this classification. A few raters found it somewhat time consuming. The overall opinion was that the refined Hurley classification was clear and easy to use. The different treatment strategies linked to each stage, were also appreciated as helpful in daily practice.

Digital assistance, for example through an application on a mobile device, could help to assess the patient.

Possible limitations of our study could be the relatively small number of raters and patients. However, standardized protocols or guidelines on methodology about reliability assessments of classification systems are lacking.[6] Therefore, our sample size was based on studies in the same field and topic and expert advice. A recent study investigating several measurement and classification systems in HS also used similar number of raters and subjects.[10]

In conclusion, our results show an overall moderate to high interrater and intrarater agreement and reliability of the refined Hurley classification in real life as well as in digital assessments. Face validity results were also positively high. Therefore, the refined Hurley classification might be a useful and practical tool to stage HS patients. We recommend further research investigating the validity of the refined Hurley classification.

Acknowledgments

The authors thank all the patients who participated.

References

1. Revuz J. Hidradenitis suppurativa. *J Eur Acad Dermatology Venereol*. 2009 Sep;23(9):985–98.
2. Jemec GBE, Kimball AB. Hidradenitis suppurativa: Epidemiology and scope of the problem. *J Am Acad Dermatol*. 2015 Nov 1;73(5):S4–7.
3. Hurley HJ. Axillary hyperhidrosis, apocrine bromhidrosis, hidradenitis suppurativa, and familial benign pemphigus: surgical approach. *Dermatologic surgery.* New York: Marcel Dekker. 1989:729–739.
4. Horváth B, Janse IC, Blok JL, et al. Hurley staging refined: A proposal by the dutch hidradenitis suppurativa expert group. *Acta Derm Venereol*. 2017;97(3):412–3.
5. Rondags A, van Straalen KR, van Hasselt JR, et al. Correlation of the Refined Hurley Classification for Hidradenitis suppurativa with Patient Reported Quality of Life and Objective Disease Severity Assessment. *Br J Dermatol*. 2018 Dec 4;
6. Kottner J, Audigé L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011 Jan 1;64(1):96–106.
7. Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Med Res Methodol*. 2016 Dec 5;16(1):93.
8. Krippendorff K. Content Analysis: An Introduction to Its Methodology. Vol. 79, Education. 2004. 440 p.
9. Ovadja ZN, Schuit MM, van der Horst CMAM, Lapid O. Inter- and Intrarater Reliability of the Hurley Staging for Hidradenitis Suppurativa. *Br J Dermatol*. 2018 Dec 26;
10. Thorlacius L, Garg A, Riis PT, et al. Interrater agreement and reliability of outcome measurement instruments and staging systems used in hidradenitis suppurativa. *Br J Dermatol*. 2019;181(3):483-491
11. Canoui-Poitrine F, Le Thuaut A, Revuz JE, et al. Identification of Three Hidradenitis Suppurativa Phenotypes: Latent Class Analysis of a Cross-Sectional Study. *J Invest Dermatol*. 2013 Jun 1;133(6):1506–11.
12. van Straalen KR, Verhagen T, Horváth B, et al. Poor interrater reliability of hidradenitis suppurativa phenotypes. *J Am Acad Dermatol*. 2018 Sep 1;79(3):577–8.
13. van der Zee HH, Jemec GBE. New insights into the diagnosis of hidradenitis suppurativa: Clinical presentations and phenotypes. *J Am Acad Dermatol*. 2015 Nov 1;73(5):S23–6.
14. Sartorius K, Emtestam L, Jemec GBE, Lapins J. Objective scoring of hidradenitis suppurativa reflecting the role of tobacco smoking and obesity. *Br J Dermatol*. 2009 Oct 1;161(4):831–9.
15. Sartorius K, Killasli H, Heilborn J, Jemec GBE, Lapins J, Emtestam L. Interobserver variability of clinical scores in hidradenitis suppurativa is low. *Br J Dermatol*. 2010 Feb 22;162(6):1261–8.
16. Kimball AB, Jemec GBE, Yang M, et al. Assessing the validity, responsiveness and meaningfulness of the Hidradenitis Suppurativa Clinical Response (HiSCR) as the clinical endpoint for hidradenitis suppurativa treatment. *Br J Dermatol*. 2014 Dec 1;171(6):1434–42.
17. 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 Nov;177(5):1401–9.
18. Kimball A, Tzellos T, Calimlim B, Teixeira H, Geng Z, Okun M. Achieving Hidradenitis Suppurativa Response Score (HiSCR) Is Associated With Significant Improvement in Clinical and Patient-Reported Outcomes: Post Hoc Analysis of Pooled Data From PIONEER I and II. *Acta Derm Venereol*. 2018;0.
19. 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 Aug 1;175(2):263–72.
20. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine - A Practical Guide*. Cambridge University Press; 2011.