Non-contrast CT markers of intracerebral hematoma expansion: a reliability study

Manuscript type : Original Research

Abbreviations:
CI : Confidence Interval
CTA : CT-Angiography
EM : Expansion Marker
HE : Hematoma Expansion
ICH: Intracerebral Hemorrhage
IQR : Interquartile Range
NCCT : Non-Contrast Computed Tomography
SD : Standard Deviation

Abstract

Objectives: We evaluated whether clinicians agree in the detection of non-contrast CT markers of intracerebral hemorrhage (ICH) expansion.

Methods: From our local dataset, we randomly sampled 60 patients diagnosed with spontaneous ICH. Fifteen physicians and trainees (Stroke Neurology, Interventional and Diagnostic Neuroradiology) were trained to identify six density (Barras density, black hole, blend, hypodensity, fluid level, swirl) and three shape (Barras shape, island, satellite) expansion markers, using standardized definitions. Thirteen raters performed a second assessment. Inter and intra-rater agreement were measured using Gwet's AC_1, with a coefficient > 0.60 indicating substantial to almost perfect agreement.

Results: Almost perfect inter-rater agreement was observed for the swirl (0.85, 95% CI: 0.78-0.90) and fluid level (0.84, 95% CI: 0.76-0.90) markers, while the hypodensity (0.67, 95% CI: 0.56-0.76) and blend (0.62, 95% CI: 0.51-0.71) markers showed substantial agreement. Inter-rater agreement was otherwise moderate, and comparable between density and shape markers. Inter-rater agreement was lower for the three markers that require the rater to identify one specific axial slice (Barras density, Barras shape, island: 0.46, 95% CI: 0.40-0.52 versus others: 0.60, 95% CI: 0.56-0.63). Inter-observer agreement did not differ when stratified for raters' experience, hematoma location, volume or anticoagulation status. Intra-rater agreement was substantial to almost perfect for all but the black hole marker.

Conclusion: In a large sample of raters with different backgrounds and expertise levels, only four of nine non-contrast CT markers of ICH expansion showed substantial to almost perfect inter-rater agreement.

MeSH Keywords

Cerebral Hemorrhage

Humans

Observer Variation

Tomography, X-Ray Computed

Key points

- In a sample of 15 raters and 60 patients, only four of nine non-contrast CT markers of ICH expansion showed substantial to almost perfect inter-rater agreement (Gwet's $AC_1 > 0.60$).
- Intra-rater agreement was substantial to almost perfect for eight of nine hematoma expansion markers.
- Only the blend, fluid level and swirl markers achieved substantial to almost perfect agreement across all three measures of reliability (inter-rater agreement, intra-rater agreement, agreement with the results of a reference reading).

Introduction

Intracerebral hematoma expansion (HE) occurs in approximately one-third of patients with spontaneous intracerebral hemorrhage (ICH) and is independently associated with early neurological deterioration, poor functional outcome, and mortality. [1-6] HE is most often defined as a 6 mL or 33% increase in ICH volume. [5] Prevention of HE by blood-pressure lowering or hemostatic therapy may improve the prognosis of ICH. Reliable predictors are necessary to stratify the risk of HE. The CT-angiography (CTA) spot sign is the most studied radiological predictor of HE. [7, 8] However, CTA is not systematically performed in hyperacute ICH. [9] To circumvent this issue, investigators have developed multiple non-contrast CT (NCCT) hematoma expansion markers (EMs), which evaluate either the density or the shape of the hematoma. [10-17] These EMs are thought to be related to a cascade phenomenon that occurs during HE, in which secondary hemorrhagic foci lead to irregular margins and heterogeneous density as the hematoma expands. [18, 19] NCCT hematoma EMs are not currently used in routine clinical practice but could eventually help select patients at high risk of HE in future trials. Prior to such utilization, these markers require independent external validation of their predictive accuracy. [20] In addition, they must show high reliability in the targeted rater population.

A recent publication standardized the definitions of nine NCCT hematoma EMs. [21] However, many EMs have complex definitions that may impact their reliability. Three markers require that the rater identifies a specific axial slice prior to evaluating hematoma density (Barras Density, Island) or shape (Barras Shape). Two markers (Black Hole, Blend) include a Hounsfield Units criterion. To date, two published agreement studies demonstrated high reliability but included at most five of nine markers, enrolled few raters [22], and/or were conducted by the ICH research experts that participated in the development of EMs. [23] These studies might overestimate the reliability of NCCT hematoma EMs in comparison with most clinical settings.

In this study, we evaluated the reliability of all nine NCCT hematoma EMs in a diverse sample of clinicians involved in the care of patients with ICH. The primary hypothesis was that, due to their complex definitions, most NCCT hematoma EMs would show limited reliability. Therefore, we sought to determine (1) inter-rater agreement, (2) the influence of EM definitions, rater experience, hematoma characteristics and anticoagulation status on inter-rater agreement, (3) intra-rater agreement and (4) agreement with the results of a reference reading.

Materials and Methods

This study was performed in accordance with the Guidelines for Reporting Reliability and Agreement Studies. [24] The study was approved by our institutional review board, including a waiver of consent for the use of deidentified patient data. All raters participating in the study provided written informed consent as required by our institutional review board.

Patients and reference standard

We retrospectively identified consecutive adult patients with acute ICH presenting at our academic comprehensive stroke center, from April 2016 to April 2020. Patients were identified through discharge codes by medical archives or by query of our institution's stroke repository. We excluded patients with a secondary cause of ICH (e.g., trauma, vascular malformation or tumor), if there was no NCCT at initial presentation or if the first NCCT available was > 6 hours from last time seen well or onset of symptoms. Anticoagulated patients were included.

Baseline NCCT images were reviewed by two of three investigators (X (Stroke Neurology, in-training, all images), XX (Stroke Neurology, 6 years of experience, half of the images) or XXX

(Diagnostic Neuroradiology, 9 years of experience, half of the images), who independently evaluated density (Barras density, black hole, blend, fluid level, hypodensity, swirl) and shape markers (Barras shape, island, satellite). When in agreement, the conclusion reached by the two investigators defined the result of the reference reading. In cases of disagreement, the third investigator (XX or XXX) independently reviewed the images, and the majority opinion was retained. These three investigators did not participate in the subsequent reliability evaluation. All investigators adhered to the EM criteria described by Morotti et al. and were blinded to clinical information and follow-up images. [21] Clinical variables were recorded by chart review.

Images

Head CTs were performed on different scanners using standard clinical parameters. (Supplementary Tables 1 and 2) Intraparenchymal and intraventricular hematomas were manually segmented to extract volumes using 3D Slicer (version 4.13.0). [25] Hematoma location was classified as lobar or deep (basal ganglia, brainstem, internal capsule, thalamus). Cerebellar hematomas were classified as either superficial or deep according to previously proposed definitions. [26]

Sample size

The sample size was determined with the method of Donner and Rotondi [27] and the *kappaSize* package in R (version 4.1.0). [28] Considering previous publications, we assumed an agreement estimate of 0.80 and a prevalence of at least 15% for each marker. [22, 23] Under these assumptions, the number of cases sampled was adequate to keep the lower bound of the 95% confidence interval (CI) higher than 0.60. [22, 23] This threshold represents the lower bound of the substantial agreement category according to Landis and Koch. [29] Cases were sampled at random from the eligible patients.

Raters

We assembled a convenience sample of 15 raters involved in the management of patients with ICH. Raters were of diverse background (Interventional Neuroradiology, Diagnostic Neuroradiology, Stroke Neurology) and level of experience (attending physicians, fellows, residents). We stratified the raters by level of experience as attending physicians or trainees (i.e., fellows and residents).

Evaluations

A training video was used for basic instructions. The video described each EM as defined by Morotti et al. and provided examples as well as common errors in interpretation. [21] Subsequently, the raters completed five practice cases, sampled from a previous publication. [21] During the readings, participants were provided with a one-page summary table of EM definitions as well the publication containing these standardized definitions. [21]

All images were de-identified and uploaded to a local secure server. Participants were blinded to all clinical information, follow-up images and other rater's assessments. They interpreted each NCCT and documented all nine EMs, without time restraint. Study data were collected and managed using the REDCap electronic data capture tools hosted at our institution. [30, 31]

The readings were performed in one or multiple sessions, at the discretion of the participants. Raters re-evaluated the cases at least two weeks after completing their initial readings. To minimize recall bias, the case order was randomly permuted, and raters were blinded to the results of their initial assessments.

Statistics

All analyses were performed using R version 4.1.0 (*irr* and *irrCAC* packages). [32-34] The Barras density and shape markers were dichotomized as positive (3-5) or negative (1-2), as previously defined. [10] Means and standard deviations (SD) are provided for normally distributed data. Medians and interquartile ranges (IQR) are provided for non-normally distributed data. Categorical variables were compared by chi-squared tests. We calculated inter-rater agreement and agreement with the reference reading using participants' first assessment.

Cohen and Fleiss κ coefficients are dependent on the prevalence of each marker and the distribution of marginals. Variations in these elements can lead to under or overestimation of agreement coefficients (the κ paradoxes). [35, 36] To minimise this possibility, we calculated chance-corrected agreement using Gwet's AC_1, a comparatively more paradox-resistant coefficient. [37] Percent agreement, multi-rater Fleiss κ (inter-rater) and Cohen's κ (intra-rater and agreement with the reference reading) coefficients were also calculated in a secondary analysis, to evaluate whether the agreement results varied with different coefficients. Global intra-observer kappa coefficients were calculated by comparing the pair of first and second readings from all raters at once. The same approach was used for agreement with the reference reading. A bias-corrected and accelerated bootstrap (1000 samples) was used to obtain 95% CIs. Coefficients were compared by examining if 95% CIs overlapped. Agreement was benchmarked using the Landis and Koch scale (≤ 0 : poor, 0.01-0.20 : slight, 0.21-0.40: fair, 0.41-0.60: moderate, 0.61-0.80, substantial, 0.81-1.00: almost perfect). [29, 38, 39]

We performed subgroup analyses to assess the potential impact of EM definitions, raters and case characteristics on overall inter-observer agreement. Markers were separately stratified by (1) their category (density versus shape), by (2) the presence of a Hounsfield Units criterion in their definition and by (3) the requirement for raters to identify the marker on one specific axial slice (Barras density, Barras shape, island). Raters were grouped by experience level (attending physicians vs trainees). Cases were

stratified by hematoma volume (< 30 mL versus ≥ 30 mL), location (deep versus lobar), and by anticoagulation status. [40]

Results

Patients

During the study period, 270 patients with spontaneous ICH were identified, of which 164 met the inclusion criteria. Of these, 60 were randomly selected for the reliability study. Patients had a mean age of 75 years (SD: 14) and 31 (52%) were women. A total of 12 (20%) patients had anticoagulation-associated ICH. The median time interval between last seen normal and initial NCCT was 92 minutes (interquartile range (IQR) : 65-238). Intraparenchymal hematoma was deep in 35 (58%) and lobar in 25 (42%). Median intraparenchymal hematoma volume was 31 mL (IQR: 11-57). Intraventricular extension was present in 30 (50%) of patients, with a median intraventricular hematoma volume of 1 mL (IQR: 0-7). (Table 1).

For the 60 selected patients, the two investigators achieved substantial (Gwet's AC_1 0.68, 95% CI: 0.61-0.74) inter-rater agreement for the detection of EMs. Based on the results of the reference reading, the proportion of positive EMs varied between 13% (fluid level) and 82% (satellite and swirl). (Table 2).

Raters

Raters included seven attending physicians (four Stroke Neurologists, two Interventional Neuroradiologists and one Diagnostic Neuroradiologist), five fellows (two Interventional Neuroradiology, two Diagnostic Neuroradiology and one Stroke Neurology) and three Neurology residents. Attending

physicians had a median of 16 years of practice (range: 2-34). Trainees (i.e., fellows and residents) had a median of 6 years of postgraduate training (range 3-7).

Inter-rater agreement

Almost perfect inter-rater agreement was observed for the swirl (Gwet's AC_1 0.85, 95% CI: 0.78-0.90) and fluid level (0.84, 95% CI: 0.76-0.90) markers, while the hypodensity (0.67, 95% CI: 0.56-0.76) and blend (0.62, 95% CI: 0.51-0.71) markers showed substantial agreement. Inter-rater agreement for the other markers was moderate (point estimate range: 0.41-0.60). (Figure 1).

Inter-rater agreement was significantly lower for the three markers that require raters to identify one specific axial slice (Barras density, Barras shape, island : Gwet's AC_1 0.46, 95% CI 0.40-0.52) versus those that do not (0.60, 95% CI 0.56-0.63). The inclusion of a Hounsfield unit criterion in the EM definition did not influence the inter-rater agreement (with : 0.57, 95% CI 0.49-0.64 versus without: 0.58, 95% CI 0.54-0.61). Inter-rater agreement was also comparable between markers of density (0.57, 95% CI: 0.53-0.61) and shape (0.52, 95% CI: 0.45-0.57). Attending physicians and trainees achieved the same inter-rater agreement (0.55, 95% CI 0.51-0.58 in both groups). Inter-rater agreement did not vary with intraparenchymal hematoma volume (< 30 mL : 0.55, 95% CI 0.50-0.60 versus ≥ 30 mL : 0.52, 95% CI : 0.47-0.57), location (deep : 0.55, 95% CI 0.51-0.59 versus lobar: 0.57, 95% CI 0.51-0.62) or anticoagulation status (with : 0.51, 95% CI 0.43-0.60 versus without : 0.56, 95% CI : 0.53-0.60).

Intra-rater agreement

Thirteen raters (86.7%) completed the second evaluation a median of 42 days (IQR: 27-50) after their initial assessment. They achieved almost perfect intra-rater agreement for the swirl (Gwet's AC_1

0.90, 95% CI 0.87-0.92) and fluid level (0.89, 95% CI 0.85-0.91) markers. The black hole marker showed moderate intra-rater agreement (0.60, 95% CI 0.54-0.66). Intra-rater agreement for the other markers was substantial (point estimate range: 0.61-0.80). (Figure 2). Pairwise intra-rater results are presented in Supplementary Figure 1.

Agreement with the reference reading

In comparison with the reference reading, individual raters tended to obtain higher proportions of positive EMs. (Table 2). Agreement with the reference reading varied from fair (hypodensity: Gwet's AC_1 0.37, 95% CI 0.31-0.44) to almost perfect (fluid level: 0.82, 95% CI 0.79-0.85). (Figure 3). Pairwise measurements of agreement with the reference reading are presented in Supplementary Figure 2.

Comparison of agreement measures

Figure 4 summarizes the three measures of agreement for each EM. Substantial to almost perfect agreement was observed across all measurements for the blend, fluid level and swirl markers. Figure 5 shows examples of cases with maximal discordance between raters. Fleiss κ (inter-rater) and Cohen's κ (intra-rater and agreement with the reference reading) results were equal to or lower than Gwet's AC_1 coefficients. (Supplementary Tables 3-6).

Discussion

In this study, the reliability of NCCT hematoma EMs was highly variable. Only four of nine markers (swirl, fluid level, hypodensity, blend) achieved substantial to almost perfect inter-rater agreement.

Non-contrast CT hematoma EMs are all associated with increased odds of HE. [17, 41, 42] They are included in HE prediction scores, and their presence may justify increased clinical and radiological monitoring. [43, 44] They may allow for HE risk stratification as a complement to the CTA spot sign, or as an alternative in centers where CTA is not available in the hyperacute setting. [45] If large multicenter datasets validate their predictive accuracy, these EMs could serve as inclusion criteria for prospective randomized control trials in ICH management. Although reliable assessments of NCCT hematoma EMs do not guarantee their predictive accuracy, poor reliability is more likely to lead to invalid hematoma expansion prediction and strongly limit their use in clinical practice.

Two previously published studies were designed to evaluate the reliability of NCCT hematoma EMs. [22, 23] Both found substantial to almost perfect agreement for all the included EMs. The generalizability of these studies is however limited. One study assessed five EMs in 40 cases [23], while the second study only included two raters and four EMs, in a much larger sample of 473 patients. [22] In one of the two studies, multiple raters were leaders in ICH research and contributed to the standardization of EM definitions. [23] Despite being limited to a single center, the measures of reliability obtained in our study are probably more representative of the targeted rater population. In a secondary analysis of the TICH-2 (Tranexamic for IntraCerebral Hemorrhage-2) trial, three raters evaluated four EMs and similarly obtained moderate to substantial inter-rater reliability. [46]

Our raters tended to obtain higher proportions of positive EMs when compared to previous studies. [21] This may be in part due to inclusion of anticoagulated and neurosurgical patients in our study combined with the lack of upper limit on intraparenchymal hematoma volume. An additional selection bias may have existed as patients with severe neurological deficits detected by Emergency Medical Services are preferentially transported to our institution. Altogether these factors may explain the higher median hematoma volumes in our sample, which may have increased the prevalence of EMs. [47, 48]

The presence of stringent criteria that define EMs may impact their reliability. For example, the Barras density and shape markers require that raters identify the axial slice with the largest cross-sectional area. The island marker requires raters to find scattered hematomas all separate from the main hematoma on the same axial slice. These definitions may increase the variability between observations and explain the moderate inter-rater agreement for the Barras density/shape and island markers. Future development of EMs should avoid this requirement to ensure greater reliability.

Anticoagulation has a significant impact on baseline hematoma volume, expansion rate and prognosis. [1, 49] Accordingly, anticoagulation may result in larger and more heterogeneous hematomas on initial NCCT, and thus be associated with a greater prevalence of EMs (e.g., fluid level). [18] A study by Zimmer et al. suggests that EMs can be used in the setting of oral anticoagulation. [50] In our study, inter-rater reliability was not impacted by anticoagulation status. Similarly, hematoma volume was previously shown to modify the prevalence of EMs [48] but did not impact inter-rater agreement in our study.

One major limitation of our study is our experimental design. It is possible that the reliability of EMs in a research protocol may differ from live expedited interpretations in the acute setting. The simultaneous evaluation of nine EMs may have fatigued raters and decreased the reliability coefficients. On the other hand, the dedicated training session and the documentation provided to raters may have increased their performance, in comparison with the average clinician. Attending physicians and trainees obtained similar measures of inter-rater agreement, which indicates that their different levels of experience with the interpretation of NCCT images do not explain our results. Gwet's AC_1 coefficients were higher than Fleiss κ measurements, which demonstrates that our choice of primary agreement measure did not penalize EM's agreement coefficients. Gwet's AC_1 is preferable to Fleiss κ in situations where the distribution of positive versus negative markers is unbalanced, as was the case for the blend, fluid level, hypodensity and swirl markers. [51] However, the use of Gwet's AC_1 coefficient limits direct comparison with previous studies, which used Fleiss κ to measure reliability. [22, 23]

Using a larger sample of raters with different background and expertise, only four of nine markers showed substantial to almost perfect inter-rater agreement. Future studies that use NCCT hematoma EMs should consider the impact of the reliability of EMs in their study design. Potential strategies to improve EM reliability could include simplifying their definitions or using automated processing and interpretation of NCCT images. The potential shift in HE predictive accuracy introduced by such approaches would also require further assessment.

References

1. Al-Shahi Salman R, Frantzias J, Lee RJ, et al (2018) Absolute risk and predictors of the growth of acute spontaneous intracerebral haemorrhage: a systematic review and meta-analysis of individual patient data. Lancet Neurol 17:885-894
2. Wang X, Arima H, Al-Shahi Salman R, et al (2015) Clinical prediction algorithm (BRAIN) to determine risk of hematoma growth in acute intracerebral hemorrhage. Stroke 46:376-381
3. Qureshi AI, Palesch YY, Barsan WG, et al (2016) Intensive Blood-Pressure Lowering in Patients with Acute Cerebral Hemorrhage. N Engl J Med 375:1033-1043
4. Brott T, Broderick J, Kothari R, et al (1997) Early hemorrhage growth in patients with intracerebral hemorrhage. Stroke 28:1-5
5. Dowlatshahi D, Demchuk AM, Flaherty ML, Ali M, Lyden PL, Smith EE (2011) Defining hematoma expansion in intracerebral hemorrhage: relationship with patient outcomes. Neurology 76:1238-1244
6. Davis SM, Broderick J, Hennerici M, et al (2006) Hematoma growth is a determinant of mortality and poor outcome after intracerebral hemorrhage. Neurology 66:1175-1181
7. Wada R, Aviv RI, Fox AJ, et al (2007) CT angiography "spot sign" predicts hematoma expansion in acute intracerebral hemorrhage. Stroke 38:1257-1262
8. Demchuk AM, Dowlatshahi D, Rodriguez-Luna D, et al (2012) Prediction of haematoma growth and outcome in patients with intracerebral haemorrhage using the CT-angiography spot sign (PREDICT): a prospective observational study. Lancet Neurol 11:307-314
9. Sprigg N, Flaherty K, Appleton JP, et al (2018) Tranexamic acid for hyperacute primary IntraCerebral Haemorrhage (TICH-2): an international randomised, placebo-controlled, phase 3 superiority trial. Lancet 391:2107-2115
10. Barras CD, Tress BM, Christensen S, et al (2009) Density and shape as CT predictors of intracerebral hemorrhage growth. Stroke 40:1325-1331
11. Li Q, Liu QJ, Yang WS, et al (2017) Island Sign: An Imaging Predictor for Early Hematoma Expansion and Poor Outcome in Patients With Intracerebral Hemorrhage. Stroke 48:3019-3025
12. Li Q, Zhang G, Huang YJ, et al (2015) Blend Sign on Computed Tomography: Novel and Reliable Predictor for Early Hematoma Growth in Patients With Intracerebral Hemorrhage. Stroke 46:2119-2123
13. Li Q, Zhang G, Xiong X, et al (2016) Black Hole Sign: Novel Imaging Marker That Predicts Hematoma Growth in Patients With Intracerebral Hemorrhage. Stroke 47:1777-1781
14. Selariu E, Zia E, Brizzi M, Abul-Kasim K (2012) Swirl sign in intracerebral haemorrhage: definition, prevalence, reliability and prognostic value. BMC Neurol 12:109
15. Yu Z, Zheng J, Ali H, et al (2017) Significance of satellite sign and spot sign in predicting hematoma expansion in spontaneous intracerebral hemorrhage. Clin Neurol Neurosurg 162:67-71
16. Boulouis G, Morotti A, Brouwers HB, et al (2016) Association Between Hypodensities Detected by Computed Tomography and Hematoma Expansion in Patients With Intracerebral Hemorrhage. JAMA Neurol 73:961-968
17. Blacquiere D, Demchuk AM, Al-Hazzaa M, et al (2015) Intracerebral Hematoma Morphologic Appearance on Noncontrast Computed Tomography Predicts Significant Hematoma Expansion. Stroke 46:3111-3116
18. Boulouis G, Morotti A, Charidimou A, Dowlatshahi D, Goldstein JN (2017) Noncontrast Computed Tomography Markers of Intracerebral Hemorrhage Expansion. Stroke 48:1120-1125
19. Fisher CM (1971) Pathological observations in hypertensive cerebral hemorrhage. J Neuropathol Exp Neurol 30:536-550

20. Arba F, Rinaldi C, Boulouis G, Fainardi E, Charidimou A, Morotti A (2021) Noncontrast Computed Tomography Markers of Cerebral Hemorrhage Expansion: Diagnostic Accuracy Meta-Analysis. Int J Stroke. DOI: 10.1177/17474930211061639
21. Morotti A, Boulouis G, Dowlatshahi D, et al (2019) Standards for Detecting, Interpreting, and Reporting Noncontrast Computed Tomographic Markers of Intracerebral Hemorrhage Expansion. Ann Neurol 86:480-492
22. Nawabi J, Elsayed S, Kniep H, et al (2020) Inter- and Intrarater Agreement of Spot Sign and Noncontrast CT Markers for Early Intracerebral Hemorrhage Expansion. J Clin Med. DOI: 10.3390/jcm9041020
23. Dowlatshahi D, Morotti A, Al-Ajlan FS, et al (2019) Interrater and Intrarater Measurement Reliability of Noncontrast Computed Tomography Predictors of Intracerebral Hemorrhage Expansion. Stroke 50:1260-1262
24. Kottner J, Audigé L, Brorson S, et al (2011) Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 64:96-106
25. Fedorov A, Beichel R, Kalpathy-Cramer J, et al (2012) 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 30:1323-1341
26. Pasi M, Marini S, Morotti A, et al (2018) Cerebellar Hematoma Location: Implications for the Underlying Microangiopathy. Stroke 49:207-210
27. Donner A, Rotondi MA (2010) Sample size requirements for interval estimation of the kappa statistic for interobserver agreement studies with a binary outcome and multiple raters. Int J Biostat. DOI: 10.2202/1557-4679.1275
28. Rotondi MA (2016) kappaSize: Sample Size Estimation Functions for Studies of Interobserver Agreement, R package version 1.2. Available via: https://CRAN.R-project.org/package=kappaSize. Accessed 1 April 2021
29. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159-174
30. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG (2009) Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 42:377-381
31. Harris PA, Taylor R, Minor BL, et al (2019) The REDCap consortium: Building an international community of software platform partners. J Biomed Inform 95:103208
32. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.
33. Gamer M, Lemon J, Fellows I, Singh P (2012) irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1 2012. Available via: https://CRAN.R-project.org/package=irr. Accessed 1 April 2021
34. Gwet KL (2019) irrCAC: Computing Chance-Corrected Agreement Coefficients (CAC). R package version 1.0 2019. Available via: https://CRAN.R-project.org/package=irrCAC. Accessed 1 April 2021
35. Feinstein AR, Cicchetti DV (1990) High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 43:543-549
36. Cicchetti DV, Feinstein AR (1990) High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 43:551-558
37. Gwet KL (2008) Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 61:29-48
38. Gwet KL (2014) Handbook of Inter-Rater Reliability, 4th edition. Advanced Analytics, LLC, United States of America.
39. Varmdal T, Ellekjær H, Fjærtoft H, Indredavik B, Lydersen S, Bonaa KH (2015) Inter-rater reliability of a national acute stroke register. BMC Res Notes 8:584

40. Hemphill JC, 3rd, Bonovich DC, Besmertis L, Manley GT, Johnston SC (2001) The ICH score: a simple, reliable grading scale for intracerebral hemorrhage. Stroke 32:891-897
41. Morotti A, Arba F, Boulouis G, Charidimou A (2020) Noncontrast CT markers of intracerebral hemorrhage expansion and poor outcome: A meta-analysis. Neurology 95:632-643
42. Yang H, Luo Y, Chen S, et al (2020) The predictive accuracy of satellite sign for hematoma expansion in intracerebral hemorrhage: A meta-analysis. Clin Neurol Neurosurg 197:106139
43. Morotti A, Dowlatshahi D, Boulouis G, et al (2018) Predicting Intracerebral Hemorrhage Expansion With Noncontrast Computed Tomography: The BAT Score. Stroke 49:1163-1169
44. Yogendrakumar V, Moores M, Sikora L, et al (2020) Evaluating Hematoma Expansion Scores in Acute Spontaneous Intracerebral Hemorrhage: A Systematic Scoping Review. Stroke 51:1305-1308
45. Morotti A, Boulouis G, Charidimou A, et al (2018) Integration of Computed Tomographic Angiography Spot Sign and Noncontrast Computed Tomographic Hypodensities to Predict Hematoma Expansion. Stroke 49:2067-2073
46. Law ZK, Ali A, Krishnan K, et al (2020) Noncontrast Computed Tomography Signs as Predictors of Hematoma Expansion, Clinical Outcome, and Response to Tranexamic Acid in Acute Intracerebral Hemorrhage. Stroke 51:121-128
47. Falcone GJ, Biffi A, Brouwers HB, et al (2013) Predictors of hematoma volume in deep and lobar supratentorial intracerebral hemorrhage. JAMA Neurol 70:988-994
48. Kim YS, Chae HY, Jeong HG, et al (2021) Size-Related Differences in Computed Tomography Markers of Hematoma Expansion in Acute Intracerebral Hemorrhage. Neurocrit Care. DOI: 10.1007/s12028-021-01347-5
49. Seiffge DJ, Goeldlin MB, Tatlisumak T, et al (2019) Meta-analysis of haematoma volume, haematoma expansion and mortality in intracerebral haemorrhage associated with oral anticoagulant use. J Neurol 266:3126-3135
50. Zimmer S, Meier J, Minnerup J, et al (2020) Prognostic Value of Non-Contrast CT Markers and Spot Sign for Outcome Prediction in Patients with Intracerebral Hemorrhage under Oral Anticoagulation. J Clin Med. DOI: 10.3390/jcm9041077
51. Quarfoot D, Levine RA (2016) How Robust Are Multirater Interrater Reliability Indices to Changes in Frequency Distribution? The American Statistician 70:373-384

A) Seven out of 15 raters rated this hematoma as heterogenous (Barras 3-5). B) Seven out of 15 raters identified a hypodensity marker in this hematoma. C) Eight out of 15 raters identified a black hole marker corresponding to the dominant encapsulated hypoattenuation. A secondary region-of-interest (ROI) evaluation performed by one author revealed that differences between mean and median ROI attenuations of the dominant hypodense focus versus the surrounding hyperdense hematoma regions on the same axial slice were approximately 24 and 23 Hounsfield units respectively. D) Seven out of 15 raters identified a blend marker. A secondary ROI analysis revealed attenuation differences between the hypodense and hyperdense components of 19 and 20 using mean and median ROI attenuation respectively. E) Seven out

of 15 raters identified a fluid level marker. F) Seven out of 15 raters rated this hematoma as irregular (Barras 3-5). G) Eight out of 15 raters classified this hematoma as containing an island marker. H) Eight out of 15 raters classified this hematoma as containing a satellite sign.

Supplementary Figure 1 - Pairwise intra-rater agreement (Gwet's AC_1)

Supplementary Figure 2 - Pairwise agreement with the reference reading (Gwet's AC_1)

Table 1 – Baseline characteristics of 60 patients

Age, mean (SD)	75 (14)
Female sex, n (%)	31 (52)
Hypertension, n (%)	46 (77)
Diabetes, n (%)	12 (20)
Dyslipidemia, n (%)	26 (43)
Coronary artery disease, n (%)	8 (13)
Atrial fibrillation, n (%)	14 (23)
Previous ischemic stroke, n (%)	11 (18)
Previous intracerebral hemorrhage, n (%)	7 (12)
Active smoking, n (%)	6 (10)
Antiplatelet, n (%)	18 (30)
Anticoagulation, n (%)	12 (20)
Glasgow coma scale, median (IQR)	13 (11-15)
NIHSS score, median (IQR)	19 (14-24)
Last time seen normal to initial CT, minutes, median (IQR)	92 (65-238)
Platelet count, 10^9/L, mean (SD)	219 (76)
International normalized ratio, mean (SD)	1.2 (0.7)
Activated partial thromboplastin time, seconds, mean (SD)	25 (4)
Intraparenchymal hematoma volume, mL, median (IQR)	31 (11-57)
Intraventricular extension, n (%)	30 (50)
Intraventricular hematoma volume, mL, median (IQR)	1 (0-7)
Hematoma location	
Deep (n, %)	35 (58)
Lobar (n, %)	28 (42)
Hematoma etiology	
Hypertensive microangiopathy (n, %)	37 (62)
Cerebral amyloid angiopathy (n, %)	13 (22)
Undetermined (n, %)	10 (16)

IQR : interquartile range; NIHSS : National Institute of Health Stroke Scale; SD : standard deviation

Table 2 – Prevalence of positive non-contrast CT markers of intracerebral hematoma expansion

	Investigator reference reading (60 patients)	**Rater results (mean of 15 ratings in 60 patients)**	**p-value**
Barras Density, n (%)	37 (62)	38 (64)	0.85
Black Hole, n (%)	11 (18)	22 (37)	0.02
Blend, n (%)	12 (20)	15 (25)	0.51
Fluid Level, n (%)	8 (13)	7 (11)	0.78
Hypodensity, n (%)	30 (50)	48 (80)	< 0.01
Swirl, n (%)	49 (82)	55 (92)	0.10
Barras Shape, n (%)	37 (62)	38 (64)	0.85
Island, n (%)	11 (18)	21 (35)	0.04
Satellite, n (%)	49 (82)	43 (71)	0.20

Supplementary Table 1 - CT scanners used for the study (total of 60 head CT exams)

Manufacturer	Model - Slice number	Count
Siemens	Definition Flash 128	51
Siemens	Sensation 64	5
Philips	Brilliance 64	2
Philips	Philips iCT 256	1
Siemens	Sensation 16	1

Supplementary Table 2 - Descriptive statistics for X-ray tube voltage (kV) and slice thickness

	kV	**Slice thickness (mm)**
Mean	115	3.52
Standard deviation	13.08	1.03
Minimum	80	2
First quartile	120	3
Median	120	3
Second Quartile	120	4
Maximum	140	7

Since automatic exposure control was used in most studies, mAs are not reported.

Supplementary Table 3 - Inter-rater agreement in 15 raters

Marker	Percent agreement	Gwet's AC_1 (95% CI)	Fleiss κ (95% CI)
Barras Density	0.68	0.41 (0.30-0.52)	0.31 (0.22-0.42)
Black Hole	0.75	0.52 (0.41-0.63)	0.46 (0.35-0.58)
Blend	0.76	0.62 (0.51-0.71)	0.37 (0.25-0.52)
Fluid Level	0.87	0.84 (0.76-0.90)	0.34 (0.17-0.61)
Hypodensity	0.78	0.67 (0.56-0.76)	0.31 (0.22-0.42)
Swirl	0.87	0.85 (0.78-0.90)	0.17 (0.05-0.36)
Barras Shape	0.77	0.58 (0.47-0.68)	0.51 (0.41-0.62)
Island	0.73	0.51 (0.40-0.63)	0.42 (0.32-0.53)
Satellite	0.75	0.58 (0.46-0.68)	0.41 (0.29-0.54)

CI : Confidence Interval

Supplementary Table 4 – Inter-rater subgroup analyses

	Percent agreement	**Gwet's AC_1 (95% CI)**	**Fleiss κ (95% CI)**
Attending Physicians (7 raters)	0.77	0.55 (0.51-0.58)	0.54 (0.50-0.58)
Trainees (8 raters)	0.77	0.55 (0.51-0.58)	0.55 (0.50-0.58)
All Six Density Markers	0.79	0.57 (0.53-0.61)	0.57 (0.53-0.61)
All Three Shape Markers	0.75	0.52 (0.45-0.57)	0.50 (0.44-0.55)
Hounsfield Unit Criterion (2 markers)	0.75	0.57 (0.49-0.64)	0.43 (0.34-0.53)
No Hounsfield Unit Criterion (7 markers)	0.78	0.58 (0.54-0.61)	0.54 (0.51-0.58)
Find One Axial Slice Criterion (3 markers)	0.73	0.46 (0.40-0.52)	0.45 (0.40-0.51)
No Find One Axial Slice Criterion (6 markers)	0.80	0.60 (0.56-0.63)	0.59 (0.55-0.63)
Intraparenchymal Hematoma Volume < 30 mL (n=29)	0.77	0.55 (0.50-0.60)	0.52 (0.47-0.57)
Intraparenchymal Hematoma Volume ≥ 30 mL (n=31)	0.78	0.60 (0.56-0.66)	0.51 (0.46-0.56)
Deep hematoma (n=35)	0.78	0.55 (0.51-0.59)	0.55 (0.51-0.59)
Lobar hematoma (n = 25)	0.77	0.57 (0.51-0.62)	0.52 (0.47-0.57)
Anticoagulation (n= 12)	0.75	0.51 (0.43-0.60)	0.47 (0.40-0.57)
No Anticoagulation (n=48)	0.78	0.56 (0.53-0.60)	0.56 (0.52-0.60)

CI : Confidence Interval

Supplementary Table 5 - Intra-rater agreement in 13 raters

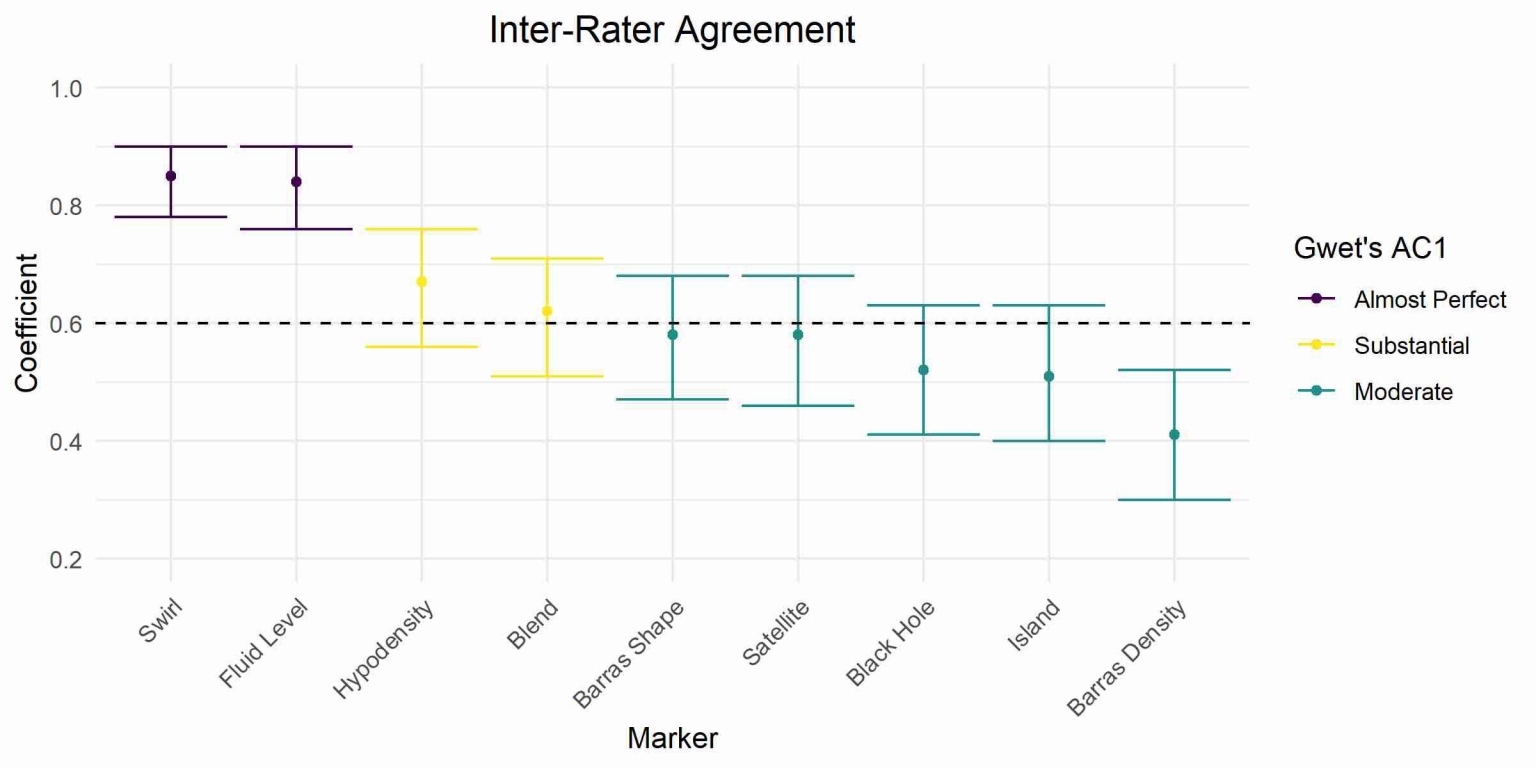
Marker	**Percent Agreement**	**Gwet's AC_1 (95% CI)**	**Cohen's κ (95% CI)**
Barras Density	0.81	0.66 (0.61-0.71)	0.57 (0.51-0.63)
Black Hole	0.79	0.60 (0.54-0.66)	0.55 (0.50-0.61)
Blend	0.88	0.80 (0.75-0.84)	0.69 (0.62-0.75)
Fluid Level	0.91	0.89 (0.85-0.91)	0.61 (0.52-0.69)
Hypodensity	0.86	0.79 (0.75-0.83)	0.56 (0.49-0.63)
Swirl	0.92	0.90 (0.87-0.92)	0.44 (0.32-0.55)
Barras Shape	0.86	0.76 (0.71-0.80)	0.69 (0.63-0.74)
Island	0.81	0.64 (0.58-0.69)	0.59 (0.53-0.65)
Satellite	0.82	0.69 (0.64-0.74)	0.57 (0.51-0.63)

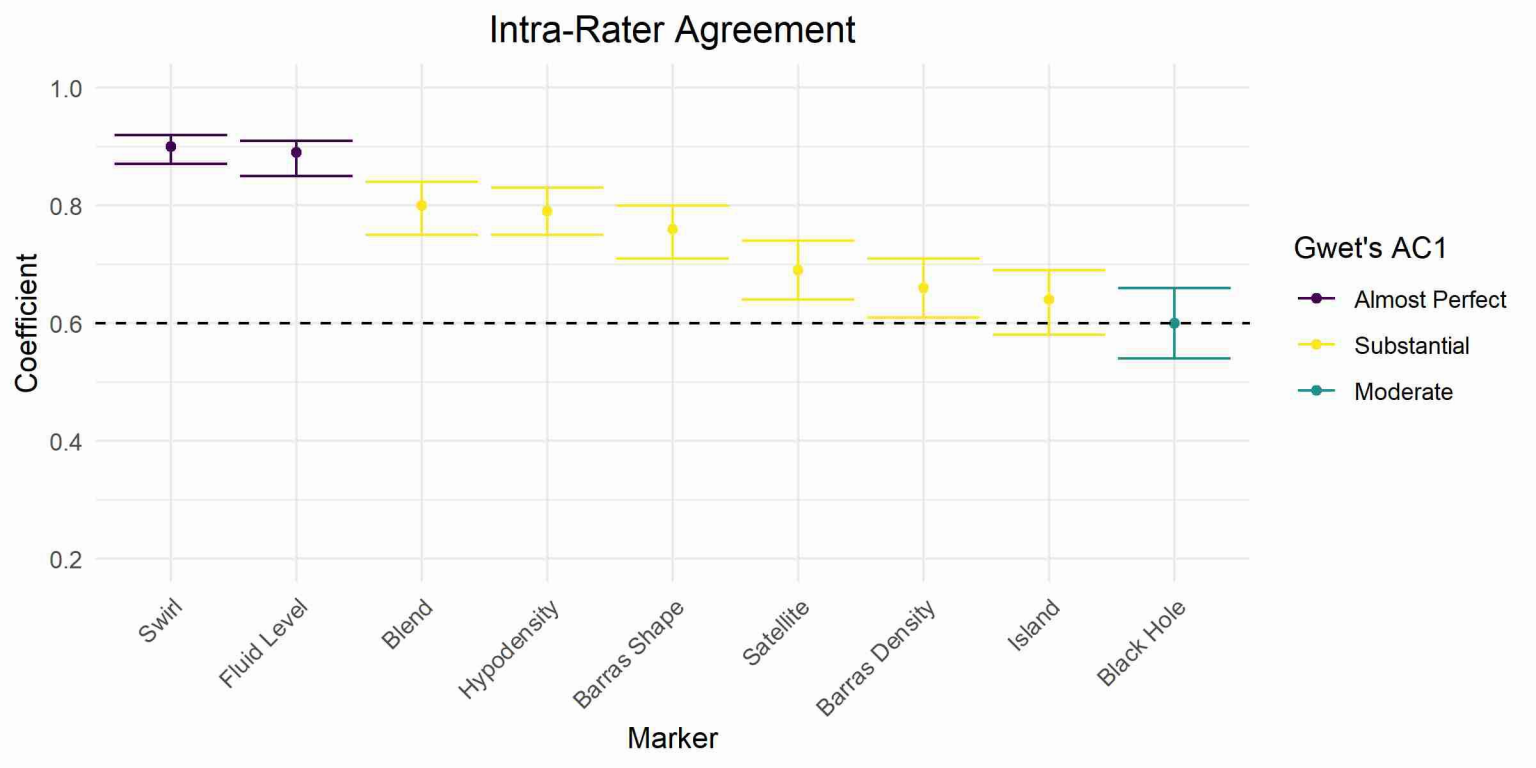
CI : Confidence Interval

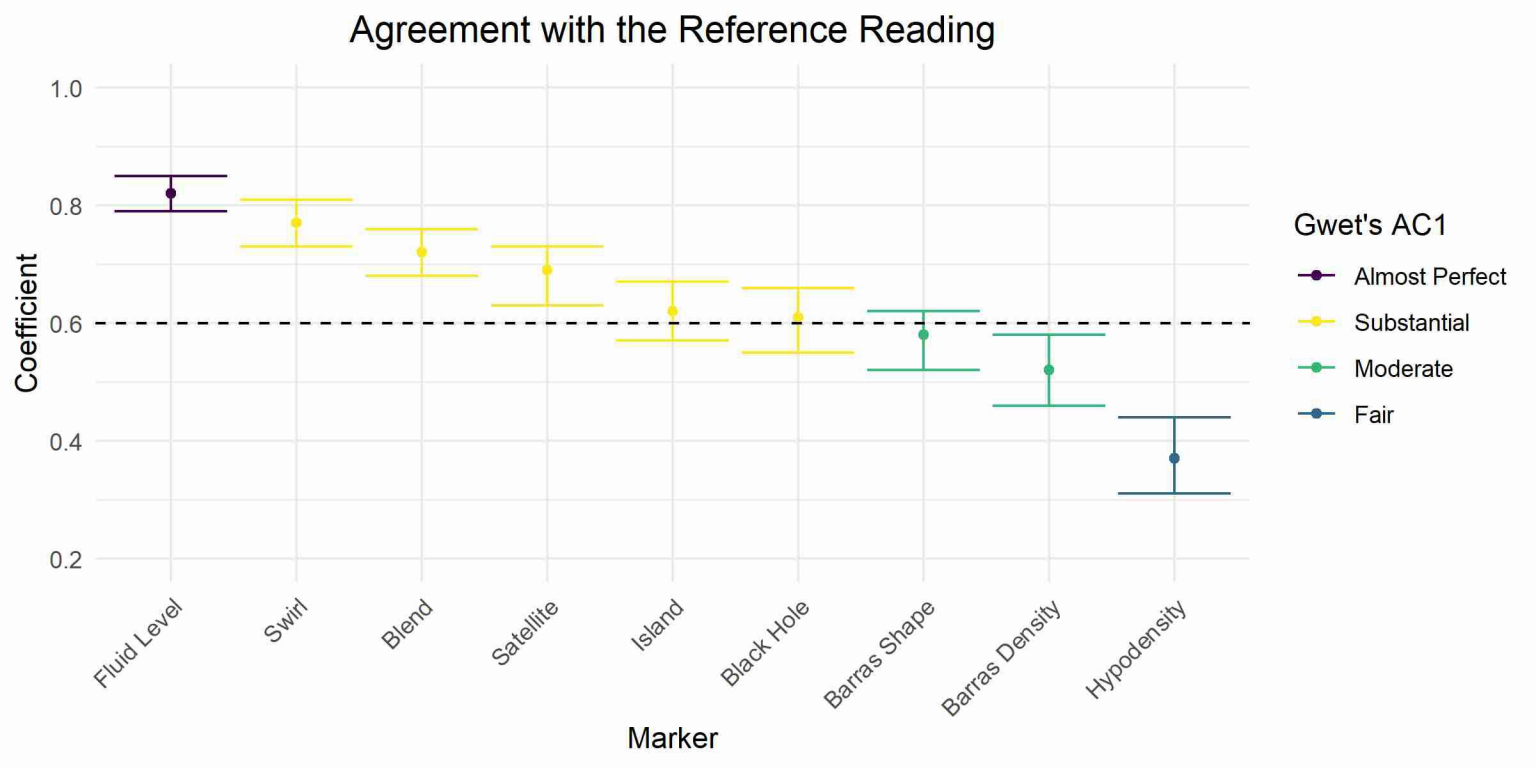
Supplementary Table 6 - Agreement with the reference reading in 15 raters

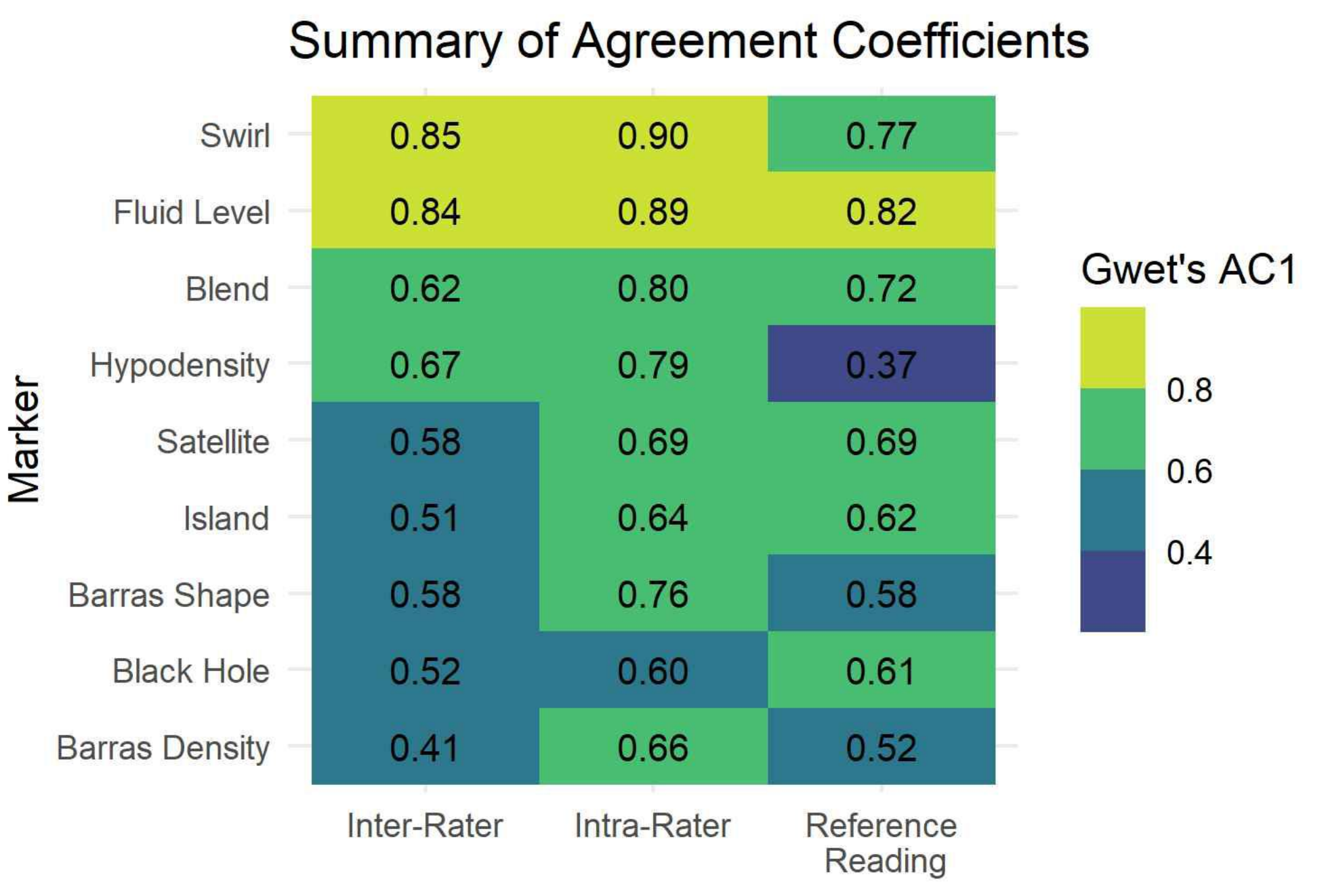
Marker	**Percent Agreement**	**Gwet's AC_1 (95% CI)**	**Cohen's κ (95% CI)**
Barras Density	0.74	0.52 (0.46-0.58)	0.45 (0.38-0.51)
Black Hole	0.76	0.61 (0.55-0.66)	0.41 (0.35-0.48)
Blend	0.82	0.72 (0.68-0.76)	0.48 (0.41-0.55)
Fluid Level	0.86	0.82 (0.79-0.85)	0.34 (0.26-0.44)
Hypodensity	0.66	0.37 (0.31-0.44)	0.25 (0.19-0.32)
Swirl	0.82	0.77 (0.73-0.81)	0.23 (0.15-0.31)
Barras Shape	0.77	0.58 (0.52-0.62)	0.52 (0.46-0.58)
Island	0.77	0.62 (0.57-0.67)	0.42 (0.35-0.48)
Satellite	0.80	0.69 (0.63-0.73)	0.45 (0.38-0.52)

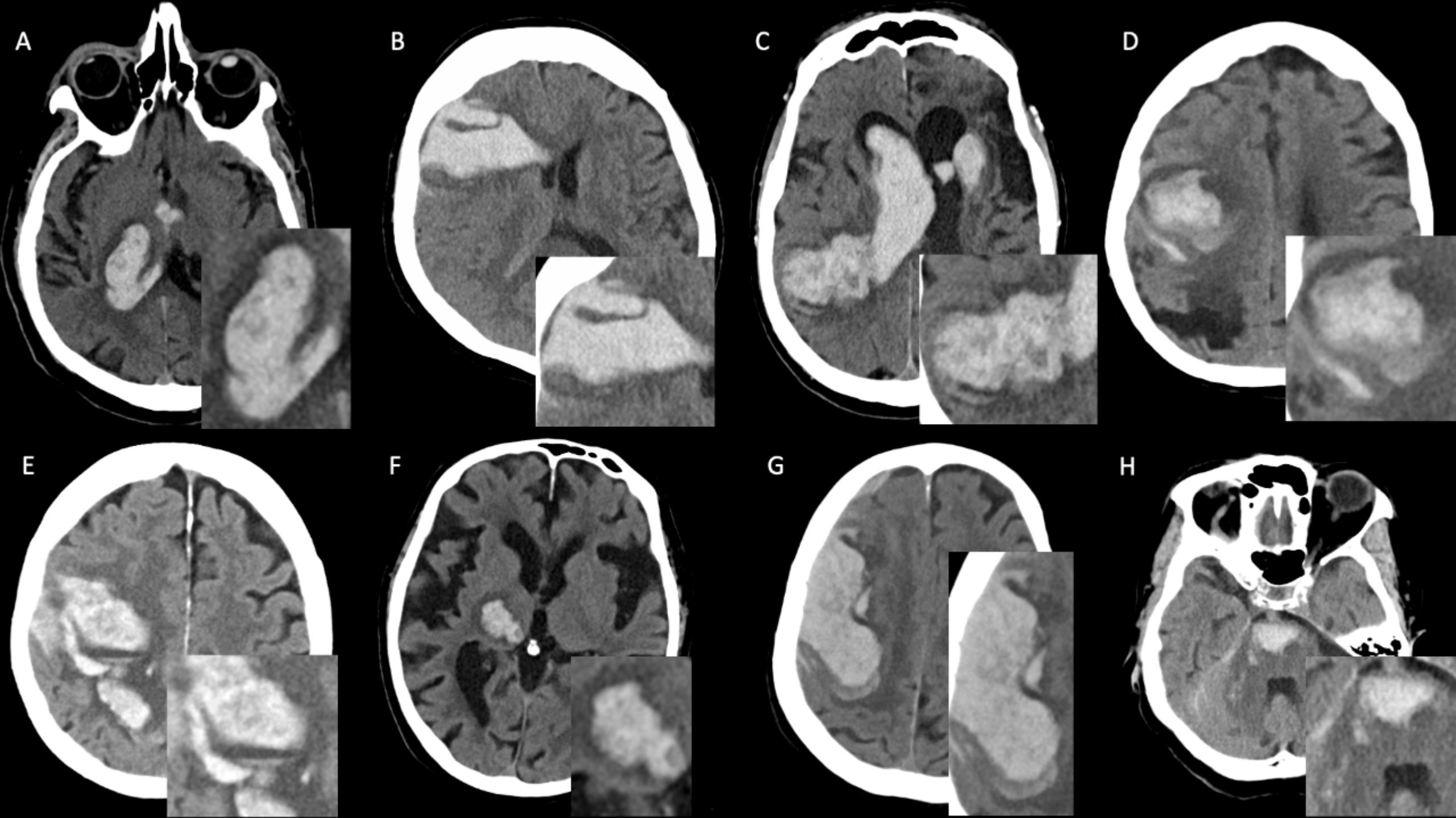
CI : Confidence Interval

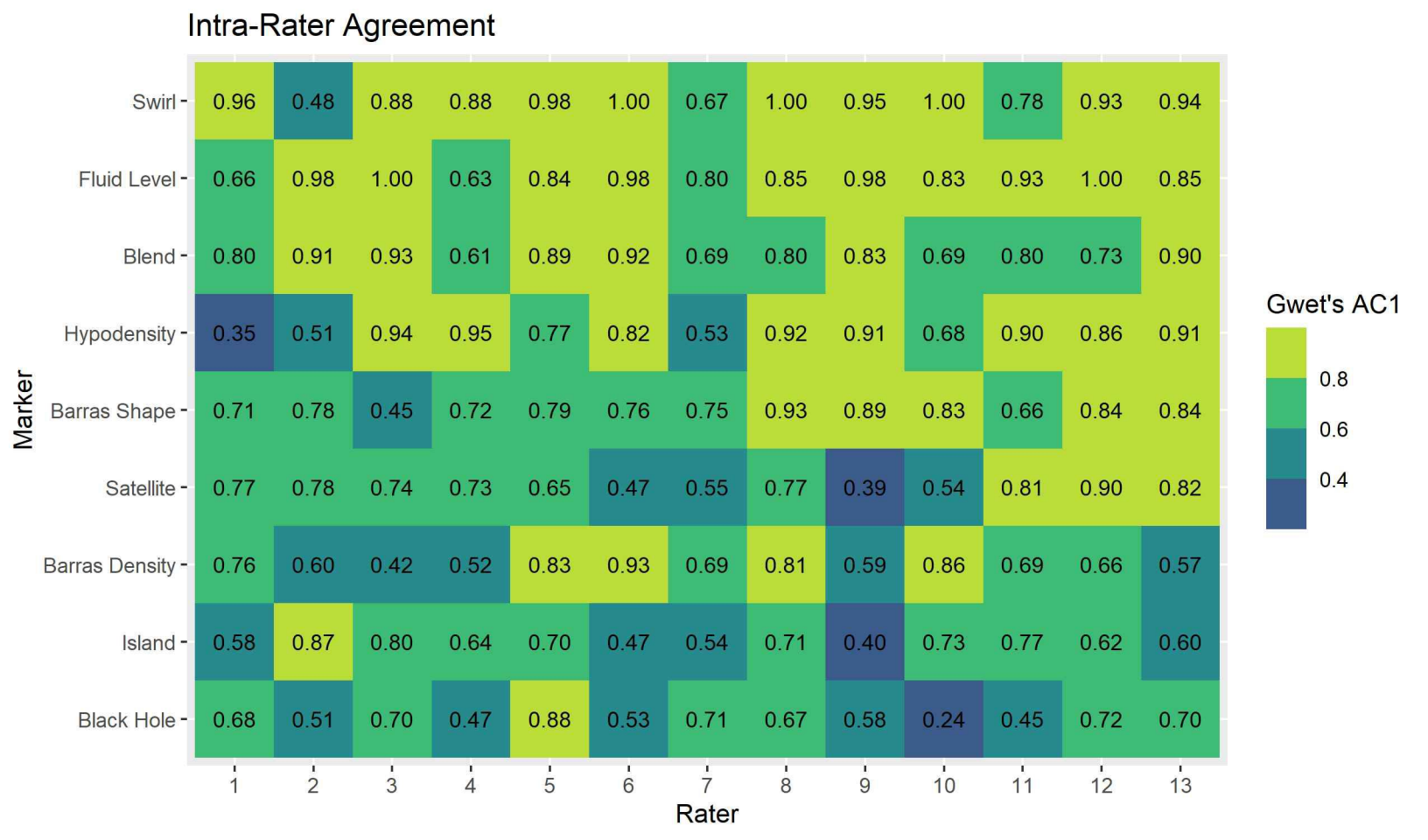

Inter-Rater Agreement
1.0
0.8
0.6
0.4
0.2
Coefficient
Swirl
Fluid Level
Hypodensity
Blend
Barras Shape
Satellite
Black Hole
Island
Barras Density
Marker
Gwet's AC1
Almost Perfect
Substantial
Moderate

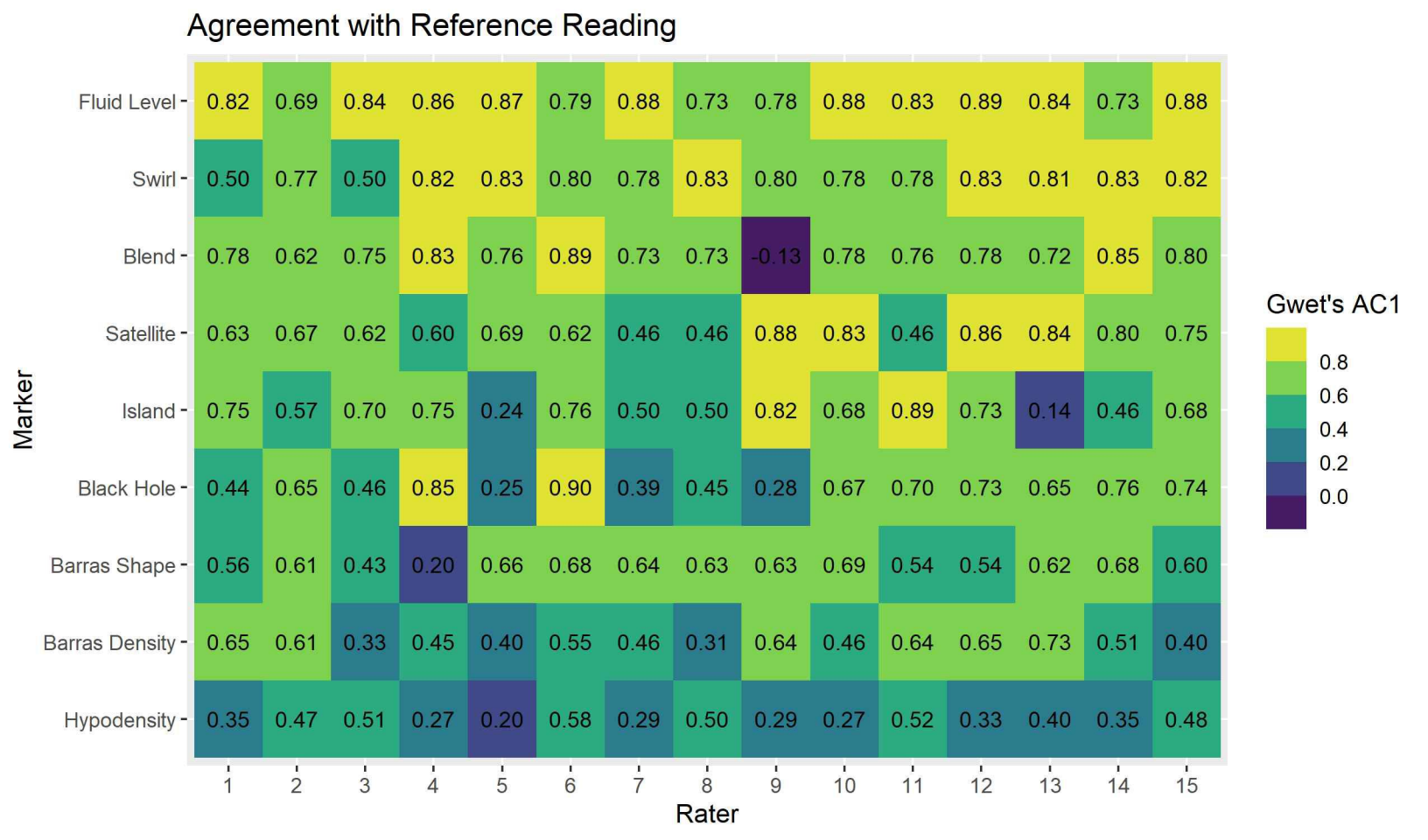

Intra-Rater Agreement
1.0
0.8
0.6
0.4
0.2
Coefficient
Swirl
Fluid Level
Blend
Hypodensity
Barras Shape
Satellite
Barras Density
Island
Black Hole
Marker
Gwet's AC1
Almost Perfect
Substantial
Moderate


Agreement with the Reference Reading
1.0
0.8
0.6
0.4
0.2
Coefficient
Fluid Level
Swirl
Blend
Satellite
Island
Black Hole
Barras Shape
Barras Density
Hypodensity
Marker
Gwet's AC1
Almost Perfect
Substantial
Moderate
Fair


Summary of Agreement Coefficients
Marker
Swirl
Fluid Level
Blend
Hypodensity
Satellite
Island
Barras Shape
Black Hole
Barras Density
0.85
0.90
0.77
0.84
0.89
0.82
0.62
0.80
0.72
0.67
0.79
0.37
0.58
0.69
0.69
0.51
0.64
0.62
0.58
0.76
0.58
0.52
0.60
0.61
0.41
0.66
0.52
Inter-Rater
Intra-Rater
Reference Reading
Gwet's AC1
0.8
0.6
0.4


A
B
C
D
E
F
G
H


Intra-Rater Agreement
Swirl 0.96 0.48 0.88 0.88 0.98 1.00 0.67 1.00 0.95 1.00 0.78 0.93 0.94
Fluid Level 0.66 0.98 1.00 0.63 0.84 0.98 0.80 0.85 0.98 0.83 0.93 1.00 0.85
Blend 0.80 0.91 0.93 0.61 0.89 0.92 0.69 0.80 0.83 0.69 0.80 0.73 0.90
Hypodensity 0.35 0.51 0.94 0.95 0.77 0.82 0.53 0.92 0.91 0.68 0.90 0.86 0.91
Barras Shape 0.71 0.78 0.45 0.72 0.79 0.76 0.75 0.93 0.89 0.83 0.66 0.84 0.84
Satellite 0.77 0.78 0.74 0.73 0.65 0.47 0.55 0.77 0.39 0.54 0.81 0.90 0.82
Barras Density 0.76 0.60 0.42 0.52 0.83 0.93 0.69 0.81 0.59 0.86 0.69 0.66 0.57
Island 0.58 0.87 0.80 0.64 0.70 0.47 0.54 0.71 0.40 0.73 0.77 0.62 0.60
Black Hole 0.68 0.51 0.70 0.47 0.88 0.53 0.71 0.67 0.58 0.24 0.45 0.72 0.70
Marker
1 2 3 4 5 6 7 8 9 10 11 12 13
Rater
Gwet's AC1
0.8
0.6
0.4


Agreement with Reference Reading
Marker
Fluid Level 0.82 0.69 0.84 0.86 0.87 0.79 0.88 0.73 0.78 0.88 0.83 0.89 0.84 0.73 0.88
Swirl 0.50 0.77 0.50 0.82 0.83 0.80 0.78 0.83 0.80 0.78 0.78 0.83 0.81 0.83 0.82
Blend 0.78 0.62 0.75 0.83 0.76 0.89 0.73 0.73 -0.13 0.78 0.76 0.78 0.72 0.85 0.80
Satellite 0.63 0.67 0.62 0.60 0.69 0.62 0.46 0.46 0.88 0.83 0.46 0.86 0.84 0.80 0.75
Island 0.75 0.57 0.70 0.75 0.24 0.76 0.50 0.50 0.82 0.68 0.89 0.73 0.14 0.46 0.68
Black Hole 0.44 0.65 0.46 0.85 0.25 0.90 0.39 0.45 0.28 0.67 0.70 0.73 0.65 0.76 0.74
Barras Shape 0.56 0.61 0.43 0.20 0.66 0.68 0.64 0.63 0.63 0.69 0.54 0.54 0.62 0.68 0.60
Barras Density 0.65 0.61 0.33 0.45 0.40 0.55 0.46 0.31 0.64 0.46 0.64 0.65 0.73 0.51 0.40
Hypodensity 0.35 0.47 0.51 0.27 0.20 0.58 0.29 0.50 0.29 0.27 0.52 0.33 0.40 0.35 0.48
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Rater
Gwet's AC1
0.8
0.6
0.4
0.2
0.0