Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.264496

MUSCULOSKELETAL IMAGING

ORIGINAL ARTICLE

©Copyright 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Interreader and intrareader agreement of qualitative cine four-dimensional computed tomography interpretation for dynamic carpal instability

Özgür Genç[1]
Babak Rafiee[2]
Ömer Naci Tabakçı[3]
İsmail Bülent Özçelik[4]

[1]İstanbul Aydın University Faculty of Medicine, Department of Radiology, İstanbul, Türkiye

[2]İstinye University, Medical Park Gaziosmanpaşa Hospital, Department of Radiology, İstanbul, Türkiye

[3]University of Health Sciences Türkiye, City Hospital, Clinic of Radiology, Kocaeli, Türkiye

[4]İstinye University, Medical Park Gaziosmanpaşa Hospital, Department of Orthopedics and Traumatology, Hand Surgery, İstanbul, Türkiye

PURPOSE

To determine the interreader and intrareader agreement of qualitative cine four-dimensional computed tomography (4D CT) interpretation for dynamic carpal instability separately for the binary judgment of whether instability is present and for the assignment of an instability phenotype.

METHODS

In this retrospective reader-agreement study, 50 wrists from 50 patients who were consecutively screened and imaged for clinically suspected carpal instability were read independently by three blinded musculoskeletal radiologists in 2 rounds separated by a 4- to 6-week washout with re-randomized case order and identifiers, giving 300 evaluations. The readers reviewed anonymized flexion–extension and radioulnar-deviation cine sequences using a common written guide and evaluation form that standardized the observations and reporting structure; quantitative thresholds for abnormal motion were not specified. The principal analysis focused on round-1 three-reader agreement for the binary overall judgment, estimated with Fleiss kappa (κ); 95% confidence intervals (CIs) were obtained by wrist-level bootstrap resampling (100,000 replicates, fixed seed). Supporting analyses included the Gwet first-order agreement coefficient (AC1), pairwise and intrareader Cohen κ, and replication in round 2. Phenotype, severity and sequence-specific agreement, reader descriptors, and clinical correlation were exploratory.

RESULTS

Round-1 Fleiss κ for the binary judgment was 0.679 (95% CI, 0.509–0.830), with identical judgments by all three readers in 38 of 50 wrists (76.0%); Gwet AC1 was 0.681 (95% CI, 0.521–0.840). Intrareader Cohen κ ranged from 0.755 to 0.801. Agreement for the six-category phenotype was 0.529 (95% CI, 0.382–0.664). Among the 21 wrists judged positive by all three readers, all three assigned the same phenotype category in 11 and the same named carpal compartment in 8; Fleiss κ restricted to these wrists was 0.473 (95% CI, 0.204–0.690). Diagnostic accuracy was not evaluated.

CONCLUSION

Among three experienced musculoskeletal radiologists, qualitative cine 4D CT yielded an interreader κ of 0.679 for the binary judgment of dynamic carpal instability. Assignment of a specific phenotype was less consistent, with interpretation limited by overlapping qualitative categories. These findings describe agreement under the study reading conditions; they do not establish diagnostic accuracy or clinical validity.

CLINICAL SIGNIFICANCE

The three experienced readers showed agreement in binary judgments under controlled reading conditions. Agreement does not establish that those judgments were correct, and the study does not validate qualitative cine 4D CT for clinical decision-making. Phenotype assignments should be interpreted with the clinical findings and other investigations.

KEYWORDS

Four-dimensional computed tomography, carpal instability, observer variation, wrist joint, reproducibility of results

Note on affiliations at the time of the study: Özgür Genç and Babak Rafiee were affiliated with İstanbul Yeni Yüzyıl University, Private Gaziosmanpaşa Hospital, İstanbul, Türkiye, when the study was conducted. The study was performed at that institution.

Handling editor: Adalet Elçin Yıldız

Corresponding author: Özgür Genç

E-mail: ozgurgenc@aydin.edu.tr

Received 23 September 2026; revision requested 22 October 2026; accepted 04 November 2026.

Epub: 17.11.2026

DOI: 10.4274/dir.2026.264496

You may cite this article as: Genç Ö, Rafiee B, Tabakçı ÖN, Özçelik İB. Interreader and intrareader agreement of qualitative cine four-dimensional computed tomography interpretation for dynamic carpal instability. *Diagn Interv Radiol*. 17 November 2026 DOI: 10.4274/dir.2026.264496 [Epub Ahead of Print].

Carpal instability causes chronic wrist pain and functional impairment; scapholunate dissociation is its most frequent form.[1,2] Midcarpal and lunotriquetral instability can also cause pain, clunking, and reduced grip strength.[3-5] Their motion dependence complicates assessment with static imaging. Magnetic resonance (MR) imaging and MR arthrography have imperfect sensitivity for scapholunate interosseous ligament tears, including at 3 T.[6] Arthroscopy evaluates ligament integrity and dynamic instability,[7,8] but its invasiveness prevents uniform verification in symptomatic cohorts.

Four-dimensional computed tomography (4D CT) acquires volumetric data during active motion.[9] Early studies described abnormal scapholunate interval behavior,[10] divergent scaphoid–lunate translation during dart-throwing motion,[11] and altered coupling after scaphoid non-union.[12] Later work evaluated reproducible angle measurements,[13] semi-automated detection of ligament rupture,[14] and dorsal intercalated segment instability.[15] Quantitative findings have been compared with arthroscopic testing,[16] assessed in patients with inconclusive conventional imaging,[17] and tested prospectively for diagnostic accuracy.[18] Simplified post-processing can reduce analysis time.[19]

Quantitative analysis provides continuous measurements but requires dedicated post-processing, and diagnostic thresholds remain study-specific.[15,17,18] Qualitative cine review assesses the temporal, multiarticular motion pattern. Reports describe capitate translation, catch-up motion, transient subluxation, and internal lunate instability, sometimes influencing clinical management.[20-22] Visual interpretation has also been compared with arthroscopy.[23] The reproducibility of qualitative categorical judgments requires separate evaluation from measurement reliability.

Prior evidence on the agreement of this qualitative task is limited: the one directly comparable estimate we identified comes from a nine-patient feasibility study in calcium pyrophosphate deposition disease, in which two readers agreed almost entirely on the presence of instability.[24] What has not been characterized is the combination of a common qualitative cine assessment and reporting structure, several independent blinded readers, repeated independent readings, and separate assessment of binary instability recognition and of phenotype assignment. We therefore determine the inter-reader and intrareader agreement of three blinded radiologists interpreting 4D CT cine sequences qualitatively. We expect binary recognition to be reproduced more consistently than phenotype assignment; this expectation is not prospectively registered.

Main points

- Three blinded musculoskeletal radiologists interpreting anonymized four-dimensional computed tomography cine sequences agreed on whether dynamic carpal instability was present in 38 of 50 wrists; Fleiss kappa (κ) was 0.679 with a 95% confidence interval (CI) of 0.509–0.830.
- Each reader reproduced his or her own binary judgment across a 4- to 6-week washout (intrareader Cohen κ, 0.755–0.801), and the analysis was closely replicated in the second reading round.
- Phenotype assignment showed a lower observed agreement. In the subgroup of 21 wrists judged positive by all three readers, all three assigned the same phenotype in 11 and the same named carpal compartment in 8 (Fleiss κ, 0.473; 95% CI, 0.204–0.690). This conditional estimate does not represent phenotype agreement across the entire referral population.
- The interpretation guide used no explicit magnitude threshold, no reference description of normal carpal motion, and no operational rule for indeterminate or multi-compartment findings; refining these definitions is a testable target for prospective evaluation.
- These results describe reproducibility only. No uniform reference standard was available, and diagnostic accuracy was neither assessed nor inferable from these data.

Methods

Study design

This retrospective reader-agreement study was conducted at a single tertiary referral hospital (İstanbul Yeni Yüzyıl University, Private Gaziosmanpaşa Hospital, İstanbul, Türkiye) between January 2023 and May 2025. The İstanbul Yeni Yüzyıl University Ethics Committee for Health Sciences Research Not Requiring Physical and Medical Intervention, approved the study (approval number 2025/06-1591; approval date 17 June 2025) and waived the requirement for informed consent due to the retrospective design. All clinical and imaging data were anonymized before evaluation.

The study was not prospectively registered, and no dated analysis plan exists; the terms principal, supporting, and exploratory used below describe the analytic hierarchy applied by the investigators and do not denote prospective specification. Because arthroscopy was not performed in every patient, no uniform reference standard existed, and the study measures agreement rather than diagnostic accuracy.

Throughout this article, "compartment" denotes one of the three articular relationships to which the phenotype categories refer—scapholunate, lunotriquetral, and midcarpal—although only the midcarpal joint is a compartment in the strict anatomical sense.

Participants

All kinematic 4D CT wrist examinations performed for clinically suspected carpal instability during the study period were screened consecutively. No additional case selection based on diagnostic difficulty or clinical outcome was performed beyond the stated eligibility criteria. Adults with a technically adequate examination containing both dynamic sequences were eligible; exclusions were previous wrist surgery, an acute fracture preventing the assessment of the intended instability phenotype, advanced degenerative carpal collapse, severe metallic or motion artifact, and incomplete dynamic acquisition. Technical adequacy was therefore an entry criterion, and inadequate examinations left the cohort before reading began. One wrist per patient was included; bilateral examinations were an exclusion criterion, but none occurred during the screening period.

Of 54 consecutive examinations screened, 4 were excluded (previous wrist surgery, n = 2; severe motion artifact, n = 1; incomplete dynamic acquisition, n = 1), leaving 50 wrists from 50 patients (Figure 1). All consecutive eligible examinations in the study period were included; no sample-size calculation for an agreement design[25] was performed, and precision is indicated by the reported 95% confidence intervals (CIs).

Four-dimensional computed tomography acquisition and cine preparation

Examinations were acquired on a Revolution CT scanner (GE Healthcare, Waukesha, WI, USA) at 120 kV with automatic tube current modulation (SmartmA, 100–300 mA) and 0.35-s gantry rotation; each motion arc was a continuous 6.3-s acquisition reconstructed as 18 volumetric phases (0.35-s interval) at 0.625-mm section thickness with a bone kernel. Mean total dose-length product for the two arcs was 4,630 mGy·cm.

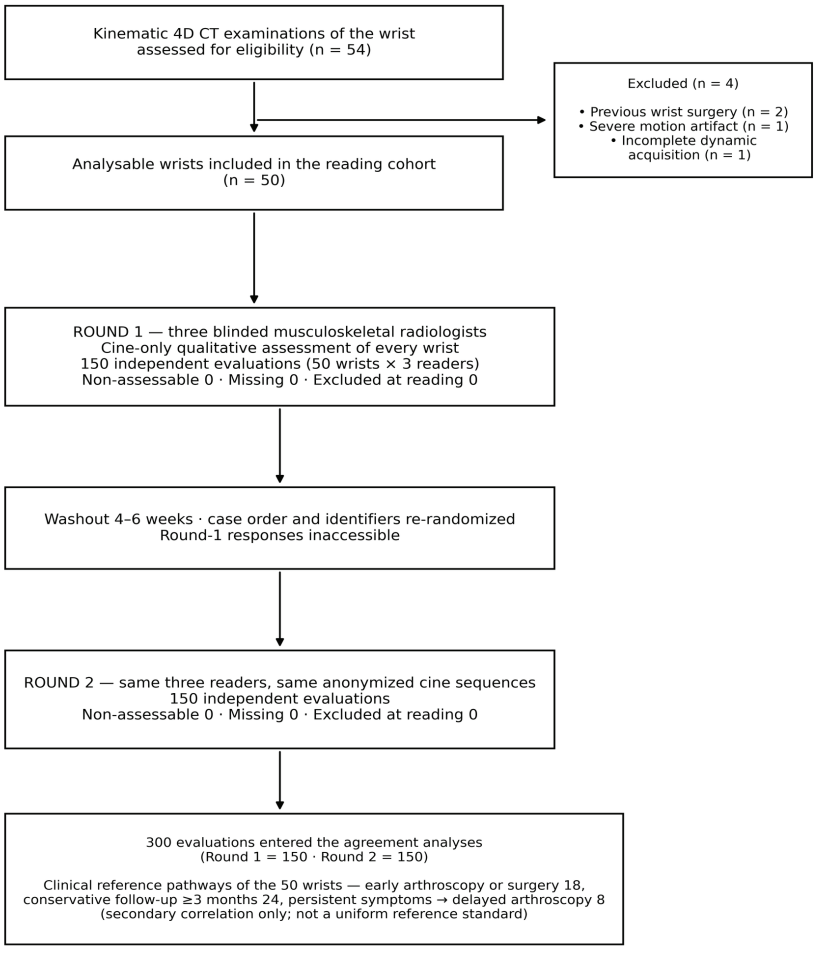


Figure 1. Study flow. Of 54 screened 4D CT examinations, 4 were excluded: previous surgery (n = 2), severe motion artifact (n = 1), and incomplete acquisition (n = 1). No bilateral examinations occurred. The remaining 50 wrists were assessed by three blinded readers in two rounds separated by 4–6 weeks, yielding 300 evaluations. Clinical management pathways provide exploratory context, not a uniform reference standard. 4D CT, four-dimensional computed tomography.

Two active arcs, flexion–extension and radioulnar deviation, were acquired for every wrist as a single slow continuous excursion through the maximal tolerated range, performed by the patient under the examiner's verbal guidance and timed to the acquisition, without axial load or clenched-fist stress; patients rehearsed the movements before acquisition, and the examiner visually checked movement adequacy during acquisition. Patients with pain moved through their maximal tolerated range. Angular range was not measured as part of the acquisition adequacy assessment. Restricted motion may therefore have reduced the visibility of a dynamic abnormality. Volumes were exported as anonymized cine sequences on an AW Server workstation, version 3.2 (GE Hungary Kft., Budapest, Hungary), with all identifiers, burned-in annotation, and accession data removed. The exported sequences were native reconstructions; no segmentation, bone-locking, or model-based post-processing was used. A representative pathological example of the dynamic findings assessed by the readers is shown in Figure 2 and Supplementary Video 1.

Reading conditions

For each wrist, readers reviewed the two motion arcs, each presented as coronal and sagittal multiplanar reformats together with three dimensional (3D) volume-rendered cine sequences (six cine loops per wrist). Sequences were reviewed on the same AW Server workstation on a diagnostic radiology monitor, with playback, looping, frame-by-frame stepping, stack scrolling, and window/level adjustment available and no time limit.

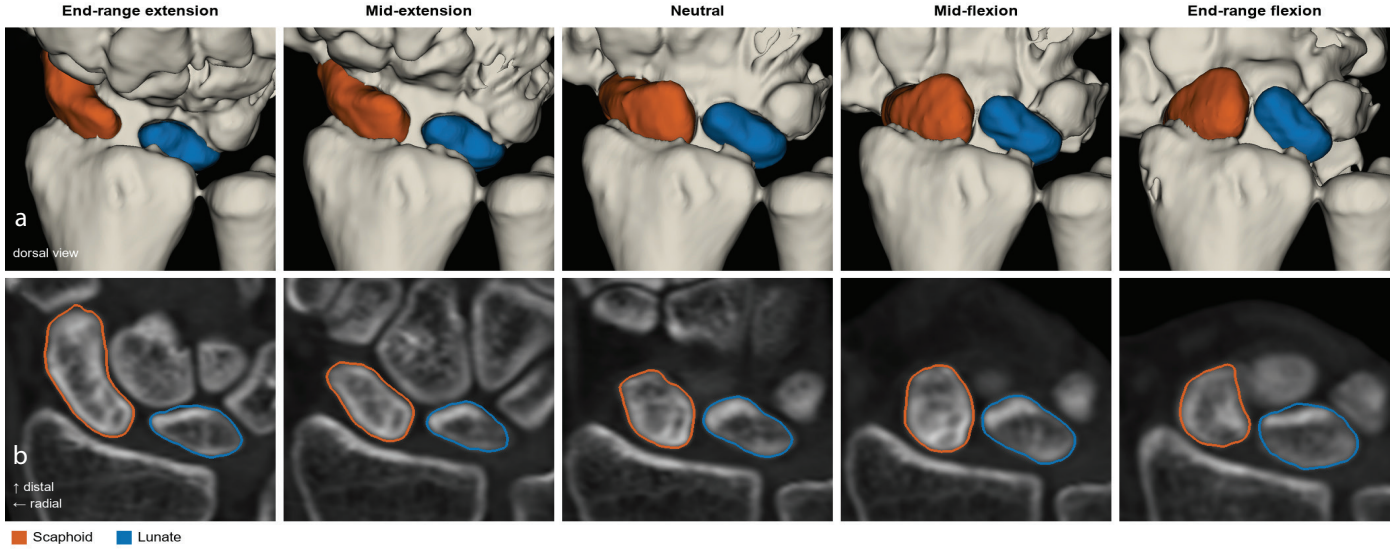


Figure 2. Representative scapholunate instability during flexion–extension. Five of 18 phases are shown. **(a)** Dorsal surface renderings. **(b)** Corresponding coronal reformats. The scaphoid is orange, and the lunate is blue. Persistent scapholunate separation and differing motion of the two bones are visible. Segmentation and a radius-fixed display were used for illustration only; readers assessed native cine reconstructions without segmentation or bone-locking.

No standardized distance or angle measurement protocol was prescribed, and no model-based quantitative post-processing formed part of the assessment; ad hoc reader measurements were not prohibited but were neither recorded nor analyzed, so the study output was the categorical judgment from cine review using the evaluation form. Separate quantitative measurements were not supplied as inputs to these qualitative reader assessments.

The supplied planes and motion arcs address different aspects of coupled 3D carpal motion. Coronal images depict scapholunate opening, whereas sagittal images depict lunate rotation and intercarpal translation.[22] A sign's display plane is distinct from the provoking arc: sagittal-plane angles have been measured during radioulnar deviation.[13,15]

Readers

Three radiologists served as readers, each with at least 10 years of general radiology and at least 5 years of dedicated musculoskeletal radiology experience. Each had previously interpreted at least 50 4D wrist CT examinations. At the institution, 4D wrist CT had been performed in selected patients rather than as a routine high-volume examination. The readers had ordinary professional experience in dynamic carpal assessment but no specific formal training. All received the same written interpretation guide. No calibration cases, consensus training, or joint review of example cases was performed before the formal readings. During the formal readings, no consensus discussion or feedback occurred, and the readers worked independently. Readers were blinded to detailed clinical information, prior imaging and reports, operative or arthroscopic findings, subsequent treatment course, and the assessments of the other readers. During round 2, readers were additionally blinded to their own round-1 assessments.

Reading rounds

Every wrist was read twice by every reader; round 2 followed a 4- to 6-week washout with re-randomized case order, re-coded identifiers, and the same guide and cine sequences, giving 150 evaluations per round and 300 in total.

Interpretation framework

For each evaluation, the readers recorded technical adequacy; dynamic instability on flexion–extension and on radioulnar deviation separately (0 = absent, 1 = present); a forced binary overall judgment of dynamic carpal instability, with no indeterminate option; and their confidence in that overall judgment on a five-point scale (1 = very low, 2 = low, 3 = moderate, 4 = high, 5 = very high). For positive judgments, they also recorded a single dominant phenotype, any associated second pattern, the motion sequence and phase in which the abnormality was most evident, the dominant anatomical level, checklists of abnormal anatomical relationships and dynamic signs, the dominant kinematic movement quality, the temporal behavior of the abnormality, and a qualitative severity grade on a three-point ordinal scale (1 = mild, 2 = moderate, 3 = marked).

A positive judgment required an abnormality of motion observed across the arc. The guide stated that static widening on a single frame, a single-frame angular abnormality, chronic deformity, or a fixed anatomical variant does not by itself establish dynamic instability. The core visual criteria were dynamic opening and closing of the scapholunate interval, loss of synchronous scaphoid–lunate motion, abnormal rotation of the scaphoid or lunate relative to adjacent carpal bones, abrupt catch-up motion, transient subluxation or abnormal translation, abnormal motion between the proximal and distal carpal rows, and delayed or abnormal lunate rotation.

The five positive phenotypes were defined as follows: scapholunate—loss of coordinated scaphoid–lunate motion, dynamic opening or closing, or abnormal relative rotation; midcarpal—abnormal translation, rotation, shift, catching, or subluxation between the proximal and distal carpal rows; lunotriquetral—dynamic motion discordance or dissociation between the lunate and triquetrum; combined/complex—definite dynamic instability involving more than one carpal compartment; and non-specific/indeterminate—dynamic instability present but no dominant anatomical phenotype assignable with confidence. With "none" for negative judgments, these form the six nominal categories analyzed.

The instrument is reproduced in Supplementary Material 1 (including Supplementary Table 1); its response options and operational definitions were not modified retrospectively. Four of its properties are relevant to interpreting the phenotype results: the absence of an explicit magnitude threshold or reference description of normal carpal motion (each reader applied an internal standard), the absence of an operational trigger for the non-specific/indeterminate category, the absence of a rule for when the involvement of more than one compartment should be recorded as combined/complex rather than as a dominant phenotype with an associated second pattern, and a scapholunate definition that repeats three of the seven core visual criteria (an observation meeting any of those three could satisfy both the positive judgment and the scapholunate label). The common guide standardized what readers observed and recorded but did not standardize the magnitude of motion each reader considered abnormal. These limitations are reported as properties of the instrument used; no new thresholds or classification rules were applied retrospectively.

The overall binary judgment was a separate form item, not derived from the two sequence-specific items; in all 300 evaluations, it nevertheless coincided with their logical union. The sequence-specific analyses are therefore not statistically independent of the principal binary analysis. Anatomical level was not analyzed separately, because it was largely determined by the assigned phenotype. Fields applying only to positive judgments are structurally not applicable to negative judgments and were never analyzed as observed negative findings. Separate anatomical-region agreement was not estimated from the descriptor checklists: Those fields were completed only after a positive overall judgment, and their zeros in negative evaluations denote structural non-applicability rather than an independently observed absence of regional instability. In addition, the combined/complex category did not specify a uniform set of involved regions. Converting these fields into regional binary diagnoses would require assumptions not defined in the original reading task.

Clinical and operative data

Clinical and operative data were used only for an exploratory correlation analysis; because verification was selective and no uniform reference standard was available, they were not used to assess diagnostic accuracy.

Statistical analysis

Statistics were computed from the reader-level and patient-level source data by a single frozen analysis script in Python 3.11 (Python Software Foundation, Beaverton, OR, USA). Continuous variables were assessed for normality with the Shapiro–Wilk

test; because neither age nor symptom duration was normally distributed, both were summarized as median with interquartile range (IQR) and range. Categorical variables are summarized as n (%) with explicit denominators.

The wrist was the unit of analysis and resampling. The principal statistic was Fleiss kappa (κ) for the three readers,[26] applied to the round-1 binary overall judgment, with a 95% CI from wrist-level bootstrap resampling (100,000 replicates, seed 20260812, percentile intervals). The same procedure was used for every other agreement coefficient; all intervals reflect the sampling of wrists, treat the three readers as fixed, and do not estimate between-reader variability in a broader reader population.

Supporting analyses were the Gwet first-order agreement coefficient (AC1) for the same judgment, pairwise round-1 Cohen κ, intrareader Cohen κ between rounds, replication of the three-reader analysis in round 2, and the stability of the three-reader majority judgment across rounds. Gwet AC1 is an alternative chance-corrected agreement coefficient[27] that may be less sensitive than κ to some marginal imbalances,[28] and it is reported alongside κ rather than as independent corroboration.

Exploratory analyses included phenotype agreement, severity agreement, sequence-specific agreement, reader descriptors, reader confidence, and the clinical correlation. Phenotype agreement used unweighted Fleiss κ across the six nominal categories, with negative judgments contributing "none." Weighting was not applied, because the categories are nominal. Because "none" occurs in exactly the negative evaluations, that coefficient reflects both the binary decision and the choice of compartment, so the analysis was repeated among the wrists judged positive by all three readers. That subgroup is selected by the readers' own binary judgments, and the resulting estimate is conditional. It is not an estimate of phenotype reproducibility in the full population referred for suspected instability. Pairwise phenotype cross-tabulations describe where disagreement occurred; each wrist contributes up to three reader pairs, so proportions carry wrist-level cluster bootstrap intervals, and the effective sample size is 50 wrists. Severity agreement was estimated with linearly weighted Fleiss κ, because the three-point grade is ordinal, and was restricted to the wrists judged positive by all three readers, since the field is recorded only for positive judgments. Sequence-specific agreement was estimated separately for flexion–extension and radioulnar deviation; no formal comparison between the two coefficients was performed. Reader confidence is reported descriptively. The exploratory clinical correlation used Fisher's exact test and a conditional maximum-likelihood odds ratio with an exact 95% CI; *P* values are given to three decimal places. The Strengthening the Reporting of Observational Studies in Epidemiology recommendations were taken into account in the reporting of this study.[29]

Results

Cohort

Fifty wrists from 50 patients entered the analysis (Figure 1, Table 1). Median age was 37 years (IQR, 27–48; range, 18–72), 31 patients (62%) were men, and the imaged wrist was the dominant hand in 24 (48%). Antecedent trauma, symptom duration, and management pathways are given in Table 1; 26 wrists (52%) underwent arthroscopy or surgery (Supplementary Table 2).

All 300 evaluations were complete for every mandatory item; none was rated non-assessable, and none was excluded at the reading stage. Because technical adequacy was an entry criterion, this does not estimate the technical success rate in unselected practice.

Binary interreader and intrareader agreement

In round 1, readers judged instability to be present in 24, 29, and 27 of 50 wrists. All three gave identical binary judgments in 38 of 50 wrists (76.0%)—17 unanimously negative and 21 unanimously positive—with two of three agreeing in the remaining 12. Fleiss κ was 0.679 (95% CI, 0.509–0.830) (Table 2; Figure 3).

Gwet AC1 for the same judgment was 0.681 (95% CI, 0.521–0.840); positive judgments accounted for 53.3% of round-1 evaluations. Pairwise round-1 Cohen κ was 0.722 (95% CI, 0.520–0.880) for readers 1 and 2 (43/50 concordant), 0.641 (95% CI, 0.410–0.839) for readers 1 and 3 (41/50) and 0.676 (95% CI, 0.448–0.874) for readers 2 and 3 (42/50).

Between rounds, intrareader agreement was 0.801 (95% CI, 0.615–0.959) for reader 1 (45/50), 0.757 (95% CI, 0.556–0.919) for reader 2 (44/50), and 0.755 (95% CI, 0.558–0.919) for reader 3 (44/50). Round 2 reproduced the principal analysis closely (Fleiss κ, 0.674; 95% CI, 0.495–0.828; again 38 of 50 wrists unanimous), and the majority judgment was stable across rounds (Cohen κ, 0.799; 95% CI, 0.610–0.959).

Phenotype agreement

Across the six categories, round-1 Fleiss κ was 0.529 (95% CI, 0.382–0.664), with all three readers concordant in 28 of 50 wrists (56.0%): 17 unanimously "none" and 11 unanimously assigned the same positive phenotype. Round 2 gave a similar estimate (0.512; 95% CI, 0.361–0.650; 28/50 concordant) (Table 3).

Table 1. Cohort characteristics (n = 50 wrists in 50 patients)

Characteristic	Value
Age, y, median (IQR); range	37 (27–48); 18–72
Male sex	31 (62)
Left wrist imaged	26 (52)
Imaged wrist is the dominant hand	24 (48)
Right-hand dominant	44 (88)
Documented antecedent trauma	30 (60)
Symptom duration, mo, median (IQR); range	8 (5–14); 1–39
Clinical management pathway	
Early arthroscopy or instability-directed surgery	18 (36)
Conservative follow-up ≥ 3 mo with symptom regression	24 (48)
Persistent symptoms followed by delayed arthroscopy	8 (16)
Underwent arthroscopy and/or surgery	26 (52)

Data are n (%) unless otherwise stated. IQR, interquartile range. Percentages use a denominator of 50. Age and symptom duration were not normally distributed (Shapiro–Wilk test) and are given as median (IQR) with range; age quartiles are 26.8 and 47.8, rounded to integers. One wrist per patient was included, and patients imaged bilaterally were excluded. The exploratory composite clinical classification, the clinical phenotype distribution, and the operative or arthroscopic findings are given in Supplementary Table 2.

Table 2. Interreader and intrareader agreement for the binary overall judgment of dynamic carpal instability

Analysis	Role	Coefficient	Estimate (95% CI)	Concordant wrists, n/N (%)
Three-reader agreement, round 1	Principal	Fleiss κ	0.679 (0.509–0.830)	38/50 (76.0)
Three-reader agreement, round 1	Supporting	Gwet AC1	0.681 (0.521–0.840)	38/50 (76.0)
Reader 1 vs. reader 2, round 1	Supporting	Cohen κ	0.722 (0.520–0.880)	43/50 (86.0)
Reader 1 vs. reader 3, round 1	Supporting	Cohen κ	0.641 (0.410–0.839)	41/50 (82.0)
Reader 2 vs. reader 3, round 1	Supporting	Cohen κ	0.676 (0.448–0.874)	42/50 (84.0)
Reader 1, round 1 vs. round 2	Supporting	Cohen κ	0.801 (0.615–0.959)	45/50 (90.0)
Reader 2, round 1 vs. round 2	Supporting	Cohen κ	0.757 (0.556–0.919)	44/50 (88.0)
Reader 3, round 1 vs. round 2	Supporting	Cohen κ	0.755 (0.558–0.919)	44/50 (88.0)
Three-reader agreement, round 2	Supporting	Fleiss κ	0.674 (0.495–0.828)	38/50 (76.0)
Three-reader agreement, round 2	Supporting	Gwet AC1	0.686 (0.524–0.841)	38/50 (76.0)
Three-reader majority, round 1 vs. round 2	Supporting	Cohen κ	0.799 (0.610–0.959)	45/50 (90.0)

AC1, first-order agreement coefficient; CI, confidence interval; κ, kappa. All 95% CIs are percentile intervals from wrist-level bootstrap resampling (100,000 replicates, seed 20260812); they reflect sampling of wrists and treat the three readers as fixed. Wrists judged positive were as follows: round 1—reader 1, 24; reader 2, 29; and reader 3, 27; and round 2—reader 1, 27; reader 2, 27; and reader 3, 31. Majority-positive wrists were as follows: round 1, 26; and round 2, 29. No formal comparison between coefficients was performed.

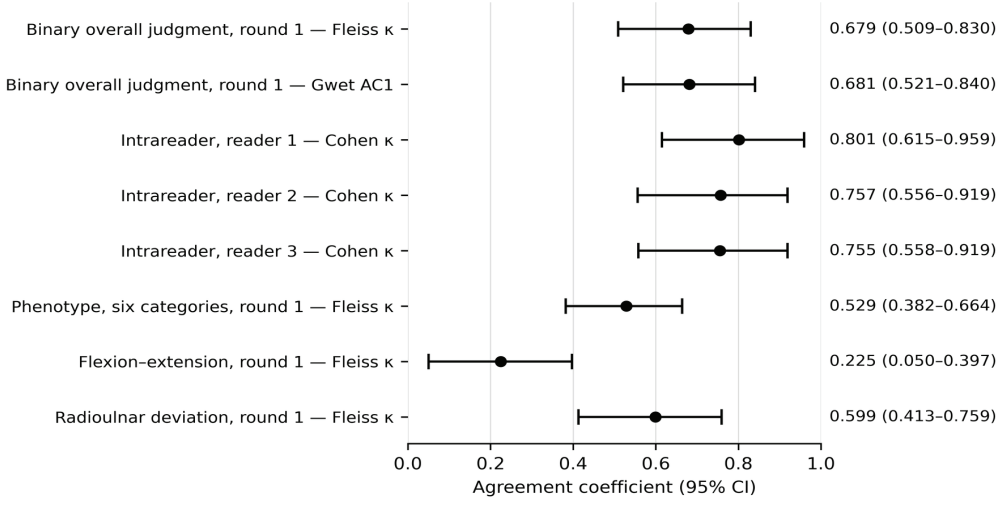


Figure 3. Agreement estimates and 95% confidence intervals (CIs) for binary interreader and intrareader judgments, phenotype assignment, and the two motion sequences. Intervals were obtained by wrist-level bootstrap resampling (100,000 replicates) and treat the readers as fixed. No formal comparison between coefficients was performed.

Among the 21 wrists judged positive by all three readers in round 1, the same dominant phenotype was chosen by all three in 11 (52.4%) and by at least two in 19 (90.5%); Fleiss κ restricted to these wrists was 0.473 (95% CI, 0.204–0.690). Of the 11 unanimous assignments, 8 named a single compartment (scapholunate 7, lunotriquetral 1, midcarpal none), 2 were unanimous non-specific/indeterminate, and 1 unanimous combined/complex. Among the 68 pairwise comparisons in which both readers judged the wrist positive, the dominant phenotype differed in 24 (35.3%; 95% CI, 20.0–50.9%). Scapholunate was the most frequently assigned phenotype; the distribution of assignments, discordant category pairs, and pairwise cross-tabulations are given in Supplementary Tables 3 and 4. This subgroup was selected by unanimous positive binary judgments; the restricted κ therefore describes phenotype assignment conditional on that agreement and should not be generalized to all 50 wrists or the wider referral population.

Severity agreement

Severity was recorded only for positive judgments and analyzed on the same conditional subgroup of 21 wrists, giving 63 grades (mild 12, moderate 25, marked 26). All three readers assigned the same grade in 11 wrists and grades within one category of each other in 20; only one wrist spanned the full range. The linearly weighted Fleiss κ was 0.560 (95% CI, 0.286–0.762) (Table 3); the unweighted estimate is given in Supplementary Table 4.

Sequence-specific agreement

For flexion–extension, Fleiss κ was 0.225 (95% CI, 0.050–0.397; unanimous in 25 of 50 wrists, 50.0%; positives 12, 18, and 17 of 50), and Gwet AC1 was 0.415 (95% CI, 0.198–0.614). For radioulnar deviation, Fleiss κ was 0.599 (95% CI, 0.413–0.759; unanimous in 35 of 50, 70.0%; positives 21, 29, and 21 of 50), and Gwet AC1 was 0.601 (95% CI, 0.440–0.766) (Table 3). The marginal positive proportions differed between the two sequences (0.313 and 0.473), and no formal comparison was made.

Table 3. Exploratory phenotype, severity, and sequence-specific agreement

Analysis	Coefficient	Estimate (95% CI)	Complete three-reader concordance, n/N (%)
Dominant phenotype, 6 categories, round 1	Fleiss κ (unweighted)	0.529 (0.382–0.664)	28/50 (56.0)
Dominant phenotype, 6 categories, round 2	Fleiss κ (unweighted)	0.512 (0.361–0.650)	28/50 (56.0)
Dominant phenotype among wrists judged positive by all three readers, round 1 (n = 21)	Fleiss κ (unweighted)	0.473 (0.204–0.690)	11/21 (52.4)
Severity grade among wrists judged positive by all three readers, round 1 (n = 21)	Fleiss κ (linearly weighted)	0.560 (0.286–0.762)	11/21 (52.4)
Flexion–extension instability, round 1	Fleiss κ	0.225 (0.050–0.397)	25/50 (50.0)
Flexion–extension instability, round 1	Gwet AC1	0.415 (0.198–0.614)	25/50 (50.0)
Radioulnar-deviation instability, round 1	Fleiss κ	0.599 (0.413–0.759)	35/50 (70.0)
Radioulnar-deviation instability, round 1	Gwet AC1	0.601 (0.440–0.766)	35/50 (70.0)

AC1, first-order agreement coefficient; CI, confidence interval; κ, kappa. All analyses in this table are exploratory. Weighted κ was not used for the phenotype, because those categories are nominal; linear weights were used for severity because that grade is ordinal. The estimates restricted to the 21 wrists judged positive by all three readers are conditional on that agreement. The 21 wrists contributed 63 severity grades (mild 12, moderate 25, marked 26); all three readers agreed exactly in 11 wrists and within one grade in 20. The distribution of dominant phenotype assignments is given in Supplementary Table 4. Wrists judged positive for flexion–extension, round 1, were as follows: 12, 18, and 17 of 50 by readers 1, 2, and 3, respectively. Wrists judged positive for radioulnar deviation, round 1, were as follows: 21, 29, and 21 of 50, by readers 1, 2, and 3, respectively, and marginal positive proportions were 0.313 and 0.473. The sequence-specific analyses are not statistically independent of the principal binary analysis, and no superiority of either sequence is implied.

Reader descriptors and confidence

Among the 80 positive round-1 evaluations, the abnormality was reported as most evident on radioulnar deviation in 38 (48%), equally on both sequences in 29 (36%) and on flexion–extension in 13 (16%); the dominant anatomical level was the proximal carpal row in 59 (74%); and the signs selected most often were loss of scaphoid–lunate synchrony (50; 63%) and dynamic scapholunate opening or closing (42; 53%). The full descriptor distributions are given in Supplementary Table 5.

Confidence was high overall: 119 of 150 round-1 evaluations were rated 4 or 5, with a median of 4 for each reader. Confidence was higher on wrists with complete three-reader agreement (median of the per-wrist mean, 4.33; IQR, 4.00–4.58) than on discordant wrists (3.33; IQR, 2.92–3.75). Because agreement-group membership derives from the same ratings, this observation is descriptive and was not tested formally.

An exploratory comparison of the round-1 majority judgment with the composite clinical classification is given in Supplementary Table 6. Reader-by-wrist judgments across both rounds are shown in Figure 4.

Discussion

Three blinded radiologists interpreting anonymized 4D CT cine sequences with a standardized evaluation form agreed on whether dynamic carpal instability was present in 38 of 50 wrists, with a Fleiss κ of 0.679 (95% CI, 0.509–0.830), and each reproduced his or her own judgment across a 4- to 6-week washout with a κ between 0.755 and 0.801. Observed agreement in assigning predefined phenotype categories was lower: All three readers judged 21 wrists as abnormal and chose the same phenotype category in 11 and the same named compartment in 8, and the restricted κ was 0.473 (95% CI, 0.204–0.690). Compartment assignment is potentially more directly relevant to management because scapholunate reconstruction, lunotriquetral stabilization and midcarpal procedures are different operations (1-5); management impact was not examined here. Agreement on the binary question was not uniform either: The readers differed on 12 of 50 wrists. The subgroup estimate is conditional on unanimous binary positivity, and the coefficients were not formally compared. Reader agreement cannot establish that a finding is diagnostically correct, even when all three readers agree.

Phenotype discordance most often involved non-specific/indeterminate or scapholunate labels. Possible contributors include overlapping definitions, differing emphasis on display planes, reader-dependent abnormal-motion thresholds, and the difficulty of localizing coupled carpal motion. These mechanisms cannot be separated in this study. Thus, lower agreement in category assignment should not be interpreted as proof that the underlying imaging phenomenon itself is poorly reproducible. Readers may recognize similar motion abnormalities while selecting different, overlapping labels.

Four properties of the interpretation framework are relevant here and are described rather than corrected retrospectively: no explicit magnitude threshold, no reference description of normal carpal motion, no operational trigger for the non-specific/indeterminate category, and no rule distinguishing combined/complex involvement from a dominant phenotype with an associated second pattern. In addition, part of the scapholunate definition duplicates the general criteria for a positive judgment. The observed disagreement identifies refinement of the interpretive framework as a testable target. Improved phenotype-level consistency may eventually have implications for preoperative communication, but this requires prospective validation. A related gap concerns the definition of normal: Quantitative work has begun to describe reference behavior of the scapholunate interval in asymptomatic wrists,[10] but no qualitative equivalent exists for cine interpretation, and the present cohort, referred entirely for suspected instability, cannot define normal limits. Without calibration or a shared normal-motion reference, readers could apply different internal standards to the same finding. The observed disagreement may therefore reflect differences in applying the framework as well as variation in the imaging appearances.

Future prospective studies should test a simpler assessment structure with explicit definitions of normal and abnormal motion, rules for equivocal and multi-compartment findings, and calibration independent of the study cases. The present data do not establish that such changes would improve agreement.

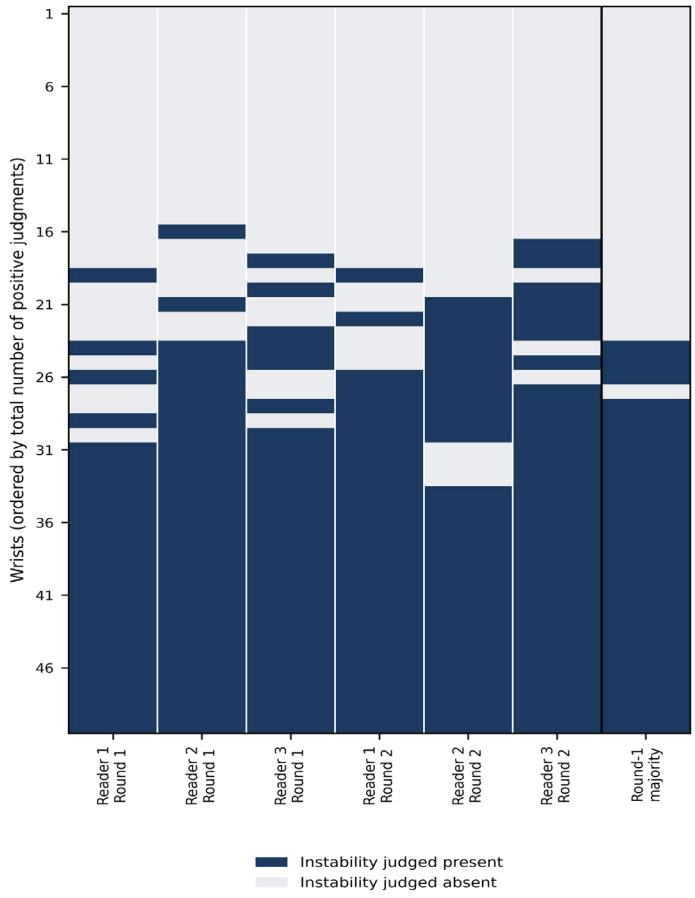


Figure 4. Binary judgments by wrist, reader, and round. Dark shading indicates instability judged as present, and pale shading indicates absence. The seventh column shows the round-1 majority judgment. Wrists are ordered by the number of positive judgments across all six readings.

Quantitative 4D CT studies generally report high measurement reproducibility, although values vary by maneuver and post-processing method.[14-16,18,19] Intraclass correlations for continuous measurements and κ values for categorical judgments are not directly comparable; these data do not establish that qualitative interpretation performs worse. Likewise, diagnostic-accuracy estimates against operative findings[18] cannot be compared with agreement coefficients. The qualitative feasibility study reporting a Cohen κ of 0.89 involved only 9 patients.[24] A study of 21 patients with suspected scapholunate tears compared interpretations with surgery rather than estimating interreader agreement.[23]

Sequence-specific coefficients differed numerically, but positive proportions also differed (0.313 vs. 0.473), and no formal comparison was performed. Because the overall decision matched the union of the sequence judgments, these are related components of the same task. Radioulnar deviation is widely used in quantitative studies,[13,14,19] but these findings do not establish superiority over flexion–extension or justify omitting either sequence.

The comparison with the composite clinical classification (Supplementary Table 6) is an exploratory clinical correlation only. Verification was selective, restricted to wrists that underwent arthroscopy or surgery, and may have been influenced by the clinical 4D CT report. The 24 wrists without operative verification were classified as negative in the composite clinical classification on the basis of their clinical pathway; this does not establish absence of instability. Consequently, this comparison cannot serve as diagnostic validation, and sensitivity, specificity, predictive values, and area under the receiver operating characteristic curve were not computed.

Several limitations should be acknowledged. This study is retrospective, single-center, and not prospectively registered, and the analytic hierarchy was applied by the investigators rather than documented in advance. Three readers evaluating 50 wrists produced wide CIs. The estimates apply to the three experienced musculoskeletal radiologists studied and cannot be assumed to apply to general radiologists or centers using different acquisition protocols. Only technically adequate examinations were included, and all readers used the same workstation and cine presentation without time restrictions. These favorable conditions do not reproduce the variation in image quality, patient motion, tolerated motion range, equipment, and reader experience encountered in routine practice. No formal calibration was performed despite the absence of quantitative thresholds and a shared normal-motion reference. Although a 4- to 6-week washout and re-randomization were used, residual recall of visually distinctive cases cannot be completely excluded. The detailed form may also have increased decision burden and exposed ambiguity between overlapping categories; without a simpler-form comparator, its effect cannot be determined. Three phenotype categories were assigned too infrequently for stable category-specific estimates. Phenotype and severity estimates among the 21 unanimously positive wrists are conditional on that selection and do not describe the full referral population. Finally, no uniform reference standard was available; agreement therefore provides no evidence that the judgments were correct.

Future studies should prospectively test clearer interpretation rules and calibration with more readers, including general radiologists, across multiple centers. Diagnostic accuracy requires an independent reference-standard strategy, using arthroscopy where feasible and a prespecified approach for patients without surgery.

For the three experienced musculoskeletal radiologists studied, qualitative cine 4D CT assessment yielded an interreader Fleiss κ of 0.679 (95% CI, 0.509–0.830) for binary judgments of dynamic carpal instability and intrareader κ values of 0.755–0.801. Agreement in assigning predefined phenotype categories was lower, with interpretation limited by category overlap and selection of the unanimously positive subgroup. These results concern reader agreement under controlled conditions and do not establish diagnostic accuracy, clinical validity, or suitability for treatment planning. Prospective evaluation with clearer operational definitions and an independent reference-standard strategy is required.

Acknowledgment

Claude (Anthropic) was used under author supervision for language editing and improving clarity and readability. All AI-assisted revisions were reviewed and verified by the authors, who take full responsibility for the final manuscript.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary Materials: https://d2v96fxpocvxx.cloudfront.net/688d2d00-d207-464d-89b6-73f393f4f50c/content-images/a251592d-13af-4167-8e3b-3ce608aa7d5c.pdf



Supplementary Video 1. https://youtu.be/ND9o5GiFOoU

References

1. Kitay A, Wolfe SW. Scapholunate instability: current concepts in diagnosis and management. *J Hand Surg Am.* 2012;37(10):2175-2196. [Crossref]
2. Wessel LE, Wolfe SW. Scapholunate instability: diagnosis and management - classification and treatment considerations - part 2. *J Hand Surg Am.* 2023;48(12):1252-1262. [Crossref]
3. Hargreaves DG. Midcarpal instability. *J Hand Surg Eur Vol*. 2016;41(1):86-93. [Crossref]
4. Rabinovich RV, Rahman OF, Nasra MH, Polatsch DB, Beldner S. Midcarpal instability. *J Am Acad Orthop Surg.* 2023;31(15):834-844. [Crossref]
5. Wilson MS. Diagnosis and management of lunotriquetral ligament injuries. *Curr Rev Musculoskelet Med.* 2023;16:55-59. [Crossref]
6. Hafezi-Nejad N, Carrino JA, Eng J, et al. Scapholunate interosseous ligament tears: diagnostic performance of 1.5T, 3T MRI, and MR arthrography-a systematic review and meta-analysis. *Acad Radiol.* 2016;23(9):1091-1103. [Crossref]
7. Lindau TR. The role of arthroscopy in carpal instability. *J Hand Surg Eur Vol.* 2016;41(1):35-47. [Crossref]
8. Geissler WB, Freeland AE, Savoie FH, McIntyre LW, Whipple TL. Intracarpal soft-tissue lesions associated with an intra-articular fracture of the distal end of the radius. *J Bone Joint Surg Am*. 1996;78(3):357-365. [Crossref]
9. Wong MT, Wiens C, Kuczynski M, Manske S, Schneider PS. Four-dimensional computed tomography: musculoskeletal applications. *Can J Surg*. 2022;65(3):E388-E393. [Crossref]
10. Demehri S, Hafezi-Nejad N, Morelli JN, et al. Scapholunate kinematics of asymptomatic wrists in comparison with symptomatic contralateral wrists using four-dimensional CT examinations: initial clinical experience. *Skeletal Radiol.* 2016;45(4):437-446. [Crossref]
11. Garcia-Elias M, Alomar Serrallach X, Monill Serra J. Dart-throwing motion in patients with scapholunate instability: a dynamic four-dimensional computed tomography study. *J Hand Surg Eur Vol.* 2014;39(4):346-352. [Crossref]
12. de Roo MGA, Dobbe JGG, van der Horst CMAM, Streekstra GJ, Strackee SD. Carpal kinematic changes after scaphoid nonunion: an in vivo study with four-dimensional CT imaging. *J Hand Surg Eur Vol.* 2019;44(10):1056-1064. [Crossref]
13. Rauch A, Abou Arab W, Dap F, Dautel G, Blum A, Gondim Teixeira PA. Four-dimensional CT analysis of wrist kinematics during radioulnar deviation. *Radiology.* 2018;289(3):750-758. [Crossref]
14. Abou Arab W, Rauch A, Chawki MB, et al. Scapholunate instability: improved detection with semi-automated kinematic CT analysis during stress maneuvers. *Eur Radiol.* 2018;28(10):4397-4406. [Crossref]
15. Athlani L, Granero J, Rouizi K, et al. Four-dimensional CT analysis of dorsal intercalated segment instability in patients with suspected scapholunate instability. *J Wrist Surg*. 2021;10(3):234-240. [Crossref]
16. Granero J, Orkut S, Rauch A, et al. Correlation between dynamic 4-dimensional computed tomography data and arthroscopic testing of scapholunate instability: a preliminary study. *J Hand Surg Am.* 2023;48(5):509.e1-509.e8. [Crossref]
17. Orkut S, Gillet R, Granero J, et al. Assessment of scapholunate instability on 4D CT scans in patients with inconclusive conventional images. *Radiology.* 2023;308(3):e230193. [Crossref]
18. Goelz L, Pinther M, Güthoff C, et al. Assessing diagnostic accuracy of four-dimensional CT for instable scapholunate dissociation: the prospective ACTION trial. *Radiology.* 2023;308(3):e230292. [Crossref]
19. Ayoub B, Sinan O, Gabriela H, et al. Post-processing of quantitative 4D-CT for initial evaluation of scapholunate instability: assessment of simplified approaches to data analysis. *Eur J Radiol.* 2024;177:111544. [Crossref]
20. Carr R, MacLean S, Slavotinek J, Bain GI. Four-dimensional computed tomography scanning for dynamic wrist disorders: prospective analysis and recommendations for clinical utility. *J Wrist Surg.* 2019;8(2):161-167. [Crossref]
21. MacLean SBM, Bain GI. Kinematics of the wrist in Kienböck's disease: a four-dimensional computed tomography study. *J Hand Surg Eur Vol.* 2021;46(5):504-509. [Crossref]
22. Repse SE, Amis B, Troupis JM. Four-dimensional computed tomography and detection of dynamic capitate subluxation. *J Med Imaging Radiat Oncol.* 2015;59(3):331-335. [Crossref]
23. Kumaş A, Teule E, Hummelink S, Buckens S, Becks M, VAN DER Heijden B. Diagnostic use of four-dimensional computed tomography in scapholunate ligament tear detection. *J Hand Surg Asian Pac Vol.* 2026;31(2):156-166. [Crossref]
24. Ulas ST, Pochandke L, Ohrndorf S, Diekhoff T, Ziegeler K. Four-dimensional computed tomography detects dynamic three-dimensional pathologies of the wrist in patients with calcium pyrophosphate deposition disease. *Front Med (Lausanne).* 2023;10:1231667. [Crossref]
25. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med.* 1998;17(1):101-110. [Crossref]
26. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull.* 1971;76(5):378-382. [Crossref]
27. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol.* 2008;61(Pt 1):29-48. [Crossref]
28. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol.* 1990;43(6):543-549. [Crossref]
29. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol.* 2008;61(4):344-349. [Crossref]