

Protocol

Reliability and reproducibility of computed tomography-based facet joint measurements in patients with AO spine type A2, A3 and A4 thoracolumbar fractures: protocol for the REFACT study

Andrey Grin¹, Aleksandr Talypov¹, Anton Kordonskiy¹, Vasiliy Karanadze¹, Ivan Lvov^{1*}, Mikhail Kucherov¹, Rinat Abdrafiev¹, Aleksandr Tupikin¹, Layla Khamidova², Roman Elokhovskiy², Ksenia Dmitrenko², Maria Nanoyan², Olga Minkina²

¹Department of Neurosurgery, Sklifosovsky Research Institute for Emergency Medicine, Moscow, Russia

²Department of Radiology, Sklifosovsky Research Institute for Emergency Medicine, Moscow, Russia

Received: 21 July 2026

Revised: 16 October 2026

Accepted: 22 October 2026

*Correspondence:

Dr. Ivan Lvov,

E-mail: dr.speleolog@gmail.com

Copyright: © the author(s), publisher and licensee Medip Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

Background: Computed tomography (CT) is widely used to assess facet joint (FJ) degeneration. However, previous studies have reported inconsistent interobserver agreement for the Pathria classification, while the reproducibility of quantitative FJ measurements remains insufficiently investigated. The REFACT study aims to evaluate the interobserver and intraobserver reliability of CT-based FJ assessment in patients with AO spine type A2, A3, and A4 thoracolumbar fractures.

Methods: REFACT is a single center, multiobserver reliability study using retrospectively collected, anonymized CT datasets. Six independent raters (three neurosurgeons and three radiologists) will assess at least 142 evaluable FJs using a standardized protocol. Assessments will include Pathria grading, mean FJ-space width, and FJ angle measurements. Each rater will perform two assessment rounds with randomized case order and a one-month washout interval. Interobserver and intraobserver reliability will be evaluated using intraclass correlation coefficients (ICC) for continuous measurements and Fleiss' kappa or weighted kappa coefficients for ordinal grading, with 95% confidence intervals. Exploratory analyses will assess agreement by specialty, vertebral level, and side. Patient-level clustered bootstrap analyses will account for within-patient correlation.

Conclusions: The REFACT study will establish the reproducibility of CT-based FJ grading and morphometric measurements, identify potential limitations of the standardized assessment protocol, and provide a methodological basis for subsequent longitudinal studies of FJ degeneration following thoracolumbar fractures.

Trial Registration: ClinicalTrials.gov, NCT07670039.

Keywords: Facet joint, Pathria classification, Computed tomography, Interobserver agreement, Intraobserver agreement, Thoracolumbar spine, Joint-space width, Facet joint angle

INTRODUCTION

Adjacent segment disease is a common long-term complication of spinal surgery. After cervical spine procedures, its incidence reaches 25% over 10 years of follow-up; after lumbar spine surgery, it is reported in up

to 14% of patients.^{1,2} CT of the spine is among the most practical modalities for detecting these changes.

The large number of acquired sections and the ability to generate multiplanar reconstructions enable sufficiently accurate assessment of the severity of degenerative changes involving FJs.

The Pathria grading system is the most widely used classification for degenerative changes of the FJs.³ Later grading systems generally reproduce this scale, introducing only minor modifications in interpretation. In addition, quantitative assessment of specific parameters may also be important when evaluating the FJs. Studies in small patient cohorts have demonstrated an association between the development of degenerative changes and the FJ angle.⁴ The study by Tiwari et al further highlighted the need for careful assessment of joint-space width when grading the severity of degenerative changes.⁵

To date, interobserver agreement for the Pathria classification based on CT data has been assessed in only four studies, the most recent of which was published 7 years ago.^{3,6-8} These studies reported highly inconsistent findings, with interobserver agreement ranging from poor to slight ($\kappa=-0.01$ to 0.12) to moderate (weighted $\kappa=0.60$). This variability precludes an objective assessment of how consistently the scale is interpreted across multiple specialists. Furthermore, we found no studies in the available literature evaluating interobserver or intraobserver agreement in the measurement of FJ-space width.

The Reliability Evaluation of FJ Assessment by CT in thoracolumbar fractures (REFACT) study represents the first methodological stage of a broader research program aimed at evaluating the longitudinal dynamics of degenerative changes in the FJs of patients with thoracolumbar junction fractures treated surgically or conservatively. Before longitudinal changes in FJ morphology can be interpreted, the reproducibility of CT-based grading and morphometric measurements must be established in the same target population. Therefore, patients with thoracolumbar junction fractures were selected as the study cohort.

Objectives

Objectives were to determine the degree of interobserver and intraobserver agreement in grading FJs according to the Pathria classification on CT images, as well as in measuring the FJ angle and mean joint-space width in patients with AOSpine type A2, A3 and A4 thoracolumbar junction fractures.

METHODS

Study design

The REFACT study is a multirater interobserver and intraobserver agreement study evaluating CT-based FJ measurements and Pathria grading.

Study setting

The study will be conducted at a level I trauma center. CT examinations of patients with thoracolumbar junction

fractures will be selected according to predefined eligibility criteria. All datasets will be anonymized before assessment. Six independent raters will participate in the evaluation: three neurosurgeons and three radiology specialists. The raters will assess the same DICOM datasets using a standardized measurement protocol and VIDAR DICOM Viewer.

The assessment will be performed in two rounds. During the first round, all raters will evaluate the randomized CT datasets within a four-week period. After a one-month washout interval, the same datasets will be re-randomized and reassessed by the same raters during a second four-week period. The setting is therefore an institutional imaging-based reliability study without direct patient contact or prospective treatment allocation.

Pre-study training procedure

Before the start of the formal assessment rounds, all raters will participate in a pre-study session. During this session, one representative anonymized CT DICOM dataset will be assessed jointly using the standardized measurement protocol and VIDAR DICOM Viewer. This case will be used to demonstrate the measurement procedure, clarify the interpretation of the Pathria classification, discuss potential sources of measurement error, and resolve possible ambiguities in the assessment of FJ parameters.

The results of the assessment will be discussed among the raters and study coordinator to identify common discrepancies, technical difficulties, and misunderstandings related to image orientation, vertebral level identification, joint-space measurement, FJ angle measurement, and Pathria grading. If necessary, the measurement instructions will be clarified before the start of the formal assessment rounds.

The calibration CT dataset will not be included in the main study sample and will not be used for inter-rater or intra-rater reliability analyses. After completion, raters will perform all formal assessments independently and will not discuss individual study cases, measurements, or grading decisions with each other until completion of both assessment rounds.

Patient selection

Anonymized CT DICOM datasets will be selected from the institutional imaging archive. No direct patient recruitment, study-specific imaging, or treatment allocation will be performed at this study.

Inclusion and exclusion criteria are presented in Table 1.

Selection procedure

Eligible CT datasets will be anonymized before distribution to the raters. The order of patients will be

individually randomized for each rater during the first assessment round. After a one-month interval, the same CT datasets will be redistributed in a changed order for the second assessment round. This procedure is intended to minimize recall bias and allow evaluation of both inter-rater and intra-rater reliability.

Interventions

No patient-level intervention will be performed. The study is based on retrospective analysis of anonymized CT DICOM datasets. The only study-related procedure is standardized image assessment by six independent raters.

Inclusion criteria

Patient age from 18 to 50 years; thoracolumbar fracture classified as AO Spine type A2, A3, or A4, availability of CT images covering vertebral levels from Th9 to L3 and absence of imaging artifacts at the levels planned for assessment were included in study.

Exclusion criteria

Bone density of the thoracolumbar junction vertebrae below 100 HU, systemic bone or joint disorders associated with changes in the FJs or intervertebral discs, including diffuse idiopathic skeletal hyperostosis, ankylosing spondylitis, rheumatoid arthritis, or other comparable conditions, previous spinal surgery, including procedures performed without implants and fractures with distraction or translational injury of the posterior supporting column were excluded from study.

Primary outcome

The primary outcome will be the inter-rater reliability of standardized CT-based FJ assessment performed by six independent raters. Reliability will be evaluated separately for categorical grading of FJ degeneration according to the Pathria classification and for quantitative morphometric measurements of the FJs, including mean joint space width and FJ angle.

Secondary outcomes

Secondary outcomes will include the intra-rater reliability of repeated CT-based assessments and exploratory subgroup analyses of agreement between different groups of raters.

Intra-rater reliability will be assessed by comparing the first and second assessment rounds for each rater after the one-month interval.

This analysis will be performed for both categorical assessment of FJ degeneration according to the Pathria classification and continuous morphometric measurements, including mean joint space width, and FJ angle.

Additional secondary analyses will evaluate inter-rater reliability separately within neurosurgeons and CT/radiology specialists, as well as agreement between these professional groups. Reliability may also be analyzed separately by vertebral level and side to assess whether measurement reproducibility differs across anatomical locations.

Subgroup analyses are exploratory and hypothesis-generating, without formal inferential claims

Sample size calculation

The sample size was calculated separately for continuous measurements and categorical Pathria grading. For continuous CT-based measurements, the calculation was performed in R using the ICC. Sample size package and the calculate ICC sample size function. With a minimally acceptable ICC of 0.60, an expected ICC of 0.75, $\alpha=0.05$, 80% power, and six raters, the minimum required sample size was 47 FJs. After adding 20% for potentially non-evaluable data, the required sample size was 58 FJs.

For the four-grade Pathria classification, the sample size was calculated in R using the kappa size package and the Power4Cats function. With six raters, a minimally acceptable kappa of 0.50, an expected kappa of 0.60, $\alpha=0.05$, 80% power, and an assumed equal distribution across four categories, the minimum required sample size was 118 FJs. After adding 20% for potentially non-evaluable data, the required sample size was 142 FJs.

Because multiple FJs may be assessed within the same patient, FJs are not fully independent observations. The primary unit of analysis will be the FJ; however, the clustered structure of the data at the patient level will be acknowledged. The target sample size refers to the number of evaluable FJs. To ensure this target is reached, CT datasets will be selected until at least 142 evaluable FJs are available.

Randomization procedure

All anonymized CT DICOM datasets will be assigned unique study identification codes. For the first assessment round, the order of datasets will be randomized individually for each rater using a computer-generated random sequence. Thus, all raters will assess the same datasets, but in different orders.

After a one-month washout interval, the same datasets will be re-randomized and redistributed to the raters in a changed order for the second assessment round. Raters will remain blinded to patient identifiers, previous measurements, and the assessments of other raters.

Blinding

Raters will be blinded to patient identifiers, clinical information, previous measurements, and the results of

other raters. All CT datasets will be anonymized and coded before assessment. During the second assessment round, datasets will be re-randomized and evaluated after a one-month washout interval to minimize recall bias.

Data collection

The study coordinator will be responsible for data collection and data administration during both assessment rounds. All study data will be entered into a standardized data collection form and stored at the data center of the participating institution. Access to the study database will be restricted to authorized study personnel until completion of the final analysis.

Statistical methods

The primary reliability analyses will be performed using AgreeStat 2015.6.2. Patient-level clustered bootstrap analyses will be performed in R.

Descriptive statistics will be used to summarize all collected variables. Continuous CT-based measurements will be presented as mean±standard deviation or median with interquartile range, depending on data distribution. Categorical variables, including the Pathria grade of FJ degeneration, will be presented as absolute numbers and percentages.

Inter-rater and intra-rater reliability for continuous variables will be assessed using the ICC with 95% confidence intervals. The ICC will be calculated using a two-way random-effects model with absolute agreement. Single-measure ICC (model ICC 2.1) will reflect the reliability of an individual rater, whereas average-measure ICC (ICC 2.k) will reflect the reliability of the mean rating across raters.

Interrater agreement for the categorical ordinal Pathria classification will be assessed using Fleiss' kappa with 95% confidence intervals. Because the Pathria classification is an ordinal four-grade scale, weighted Fleiss' kappa with quadratic weights will also be calculated. Unweighted Fleiss' kappa will be used to estimate overall categorical agreement without accounting for the magnitude of disagreement, whereas weighted Fleiss' kappa will account for the ordered nature of the Pathria grades. Intra-rater agreement for Pathria grading will be assessed separately for each rater using weighted Cohen's kappa with quadratic weights.

For descriptive interpretation, kappa coefficients will be categorized according to the Landis et al criteria: <0.00, poor agreement; 0.00-0.20, slight agreement; 0.21-0.40, fair agreement; 0.41-0.60, moderate agreement; 0.61-0.80, substantial agreement; and 0.81-1.00, almost perfect agreement.⁹ ICC values will be interpreted according to the criteria proposed by Koo et al: <0.50, poor reliability; 0.50-0.75, moderate reliability; 0.75-0.90, good reliability; and >0.90, excellent reliability.¹⁰

Additional analyses may be performed separately by vertebral level, side, and rater specialty, including neurosurgeons and radiology specialists.

To account for clustering of FJs within patients, additional patient-level clustered bootstrap analyses will be performed for the confidence intervals of the main reliability estimates. In each bootstrap iteration, patients rather than individual FJs will be resampled with replacement, and all evaluable FJs belonging to the selected patients will be retained.

Missing, technically inadequate, or non-evaluable measurements will not be imputed and will be excluded from the corresponding analysis. The number and proportion of missing or non-evaluable values will be reported descriptively.

Interim analysis and stopping rules

No interim analysis is planned. Statistical analysis will be performed only after completion of both assessment rounds and finalization of the study database.

Formal stopping rules are not applicable. No therapeutic or diagnostic intervention will be administered to patients, and no patient-level safety monitoring or efficacy assessment is required.

The study may be suspended or terminated prematurely only in case of major protocol violations, inability to complete data collection, loss or corruption of study data, or other administrative or ethical reasons that prevent valid completion of the reliability analysis.

Ethics and dissemination

The study will be performed according to Helsinki declaration of the World Medical Association. The protocol for this study was approved by the local ethical committee.

The study has been registered at ClinicalTrials.gov under registration number NCT07670039. The results of the study will be reported in accordance with the Guidelines for Reporting Reliability and Agreement Studies (GRRAS).¹¹ Study findings are planned to be disseminated through publication in peer-reviewed journals in neurosurgery, spine surgery, radiology, or related fields, and through presentation at relevant scientific conferences.

DISCUSSION

This study has several strengths. It uses a multirater design with six independent raters from two relevant clinical specialties, repeated assessment rounds separated by a one-month washout interval, randomized dataset order, blinded assessment, and a standardized CT-based measurement protocol. The study will evaluate both inter-

rater and intra-rater reliability and will include categorical Pathria grading as well as quantitative morphometric measurements of the FJs. A pre-study session will be performed to ensure consistent understanding of the measurement procedure before the formal assessment rounds.¹¹

Several limitations should also be acknowledged. First, this is a single-center retrospective imaging-based study, which may limit generalizability. Second, the study population consists of patients with thoracolumbar junction fractures and may not fully represent the broader population of patients. However, this cohort was selected because the present reliability study represents the first methodological stage of a planned longitudinal project evaluating FJ degenerative changes after surgical and conservative treatment of thoracolumbar fractures. Third, the distribution of Pathria grades in this relatively young trauma cohort may be imbalanced, which may affect kappa estimates and the precision of categorical agreement analyses.^{3,6-8} Fourth, multiple FJs will be assessed within the same patient, creating a clustered data structure. This will be addressed in the interpretation of the results and, where applicable, by patient-level clustered bootstrap analyses. Finally, subgroup analyses by vertebral level, side, and rater specialty will be considered exploratory.

CONCLUSION

The REFRACT study is designed to determine whether CT-based Pathria grading and quantitative FJ measurements can be applied reproducibly by specialists from different clinical backgrounds. By evaluating both interobserver and intraobserver reliability, the study will identify the strengths and potential limitations of the proposed assessment protocol. The findings will help determine whether mean FJ-space width and FJ angle are sufficiently reliable for use in subsequent longitudinal analyses. They may also support the standardization of CT-based FJ assessment in patients with thoracolumbar fractures and inform the design of future multicenter studies.

Funding: No funding sources

Conflict of interest: None declared

Ethical approval: The study was approved by the Institutional Ethics Committee

REFERENCES

1. Cho SK, Riew KD. Adjacent segment disease following cervical spine surgery. *J Am Acad Orthop Surg.* 2013;21(1):3-11.
2. Saavedra-Pozo FM, Deusdara RA, Benzel EC. Adjacent segment disease perspective and review of the literature. *Ochsner J.* 2014;14(1):78-83.
3. Pathria M, Sartoris DJ, Resnick D. Osteoarthritis of the facet joints: accuracy of oblique radiographic assessment. *Radiology.* 1987;164(1):227-30.
4. Kalichman L, Suri P, Guermazi A, Ling L, David JH. Facet orientation and tropism: associations with facet joint osteoarthritis and degeneratives. *Spine (Phila Pa 1976).* 2009;34(16):E579-85.
5. Tiwari P, Kaur H, Jha V. Is thoracic facet joint arthritis over-reported? It's time to review CT grading parameters-an analytical cross-sectional study. *Indian J Radiol Imaging.* 2020;30(4):427-35.
6. Coste J, Judet O, Barre O, Siaud JR, Cohen de Lara A, Paolaggi JB. Inter- and intraobserver variability in the interpretation of computed tomography of the lumbar spine. *J Clin Epidemiol.* 1994;47(4):375-81.
7. Berg L, Thoresen H, Neckelmann G, Håvard F, Christian H, Ansgar E. Facet arthropathy evaluation: CT or MRI? *Eur Radiol.* 2019;29(9):4990-8.
8. Stieber J, Quirno M, Cunningham M, Thomas JE, John AB. The reliability of computed tomography and magnetic resonance imaging grading of lumbar facet arthropathy in total disc replacement patients. *Spine (Phila Pa 1976).* 2009;34(23):E833-40.
9. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159-74.
10. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016;15(2):155-63.
11. Kottner J, Audigé L, Brorson S, Allan D, Byron JG, Asbjørn H, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol.* 2011;64(1):96-106.

Cite this article as: Grin A, Talypov A, Kordonskiy A, Karanadze V, Lvov I, Kuchеров M, et al. Reliability and reproducibility of computed tomography-based facet joint measurements in patients with AO spine type A2, A3 and A4 thoracolumbar fractures: protocol for the REFRACT study. *Int J Clin Trials* 2026;13(4):442-6.