Universität Zürich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

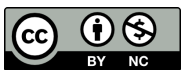
Year: 2026

Validation of the AO Spine Osteoporotic Fracture Classification-effect of geographical region on reliability and reproducibility

Bigdon, Sebastian Frederick; Scherer, Julian; Camino-Willhuber, Gaston; Spiegl, Ulrich; Joaquim, Andrei Fernandes; Chhabra, Harvinder Singh; Dvorak, Marcel; Schroeder, Gregory; El-Sharkawi, Mohammad; Bransford, Richard; Benneker, Lorin Michael; Schnake, Klaus John; AO Spine-DGOU international validation Group

DOI: https://doi.org/10.1007/s00198-026-08112-4

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://www.zora.uzh.ch/handle/20.500.14742/249157
Journal Article
Published Version



The following work is licensed under a Creative Commons: Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

Originally published at:
Bigdon, S. F., Scherer, J., Camino-Willhuber, G., Spiegl, U., Joaquim, A. F., Chhabra, H. S., Dvorak, M., Schroeder, G., El-Sharkawi, M., Bransford, R., Benneker, L. M., Schnake, K. J., & AO Spine-DGOU international validation Group. (2026). Validation of the AO Spine Osteoporotic Fracture Classification-effect of geographical region on reliability and reproducibility. Osteoporosis International, 37, 2717–2727. https://doi.org/10.1007/s00198-026-08112-4
DOI: https://doi.org/10.1007/s00198-026-08112-4

Osteoporosis International
https://doi.org/10.1007/s00198-026-08112-4

ORIGINAL ARTICLE

Validation of the AO Spine Osteoporotic Fracture Classification—effect of geographical region on reliability and reproducibility

Sebastian Frederick Bigdon[1,2] · Julian Scherer[3,4] · Gaston Camino-Willhuber[5] · Ulrich Spiegl[6] · Andrei Fernandes Joaquim[7] · Chhabra Harvinder Singh[8] · Marcel Dvorak[9] · Gregory Schroeder[10] · Mohammad El-Sharkawi[11] · Richard Bransford[12] · Lorin Michael Benneker[2] · Klaus John Schnake[13,14] · the AO Spine-DGOU international validation Group

Received: 18 July 2024 / Accepted: 11 June 2026
© The Author(s) 2026

Abstract

Summary Regional differences affect the reliability of the AO Spine-DGOU Osteoporotic Fracture Classification. European participants showed the highest agreement. This study underscores the need for addressing regional factors to improve the system's global clinical utility.

Purpose To evaluate the influence of geographic region on the reliability and reproducibility of the AO Spine-DGOU Osteoporotic Fracture Classification System.

Methods This study included 320 participants from various global regions who classified 27 cases of osteoporotic vertebral fractures using the AO Spine-DGOU system which categorizes the fractures to 5 subtypes (OF 1–OF 5). Participants underwent training via an online webinar. Interobserver reliability and intraobserver reproducibility were assessed using Fleiss' kappa coefficient, and agreement with a gold standard committee was evaluated.

Results The classification system showed moderate to substantial agreement with the gold standard globally (initial kappa 0.58, improving to 0.61). European participants had the highest agreement (kappa 0.64 and 0.66). OF4 fractures were most accurately classified, while OF3 fractures showed the least agreement. Intraobserver reliability was highest among European participants. Post hoc analysis indicated significantly better reliability among German-speaking participants compared to other Europeans (kappa 0.79 vs. 0.70, $p = 0.0026$).

Conclusion The AO Spine-DGOU Osteoporotic Fracture Classification System demonstrates moderate to substantial reliability and reproducibility, with regional differences influenced by factors such as training and clinical experience. This underlines the necessity of proper education adapted to the regional particularities.

Keywords fracture classification · osteoporosis · spinal fracture · validation study

Introduction

Osteoporosis is a widespread disease with an increasing incidence. In Switzerland, the lifetime risk of osteoporotic fractures for 50-year-old women is 51%, signifying a substantial disease burden [1, 2]. Similar occurrence and prevalence are expected in all industrialized countries. With the rising incidence of osteoporotic fractures, it is crucial to implement evidence-based treatment as widely as possible [2, 3]. For a proposed or executed therapy to be compared, a "common language" is needed. This common language starts with the description of the fracture—the classification. Until 2018, there were a variety of different fracture classifications that were used differently. The most widely used classification was probably the Genant classification [4]. In this classification system, an attempt was made to categorize fracture morphology into 12 distinct groups. However, due to limited diagnostic utility with regard to stability and therapeutic consequences, the classification fails to effectively guide treating physicians in formulating therapeutic decisions.

In 2013, AO Spine introduced a new classification system for thoracolumbar fractures [5]. The AO Spine thoracolumbar injury classification system was quickly accepted and showed high reliability in an international validation [6]. The classification also gives a certain indication of the stability of the fracture through the description of the morphology

Extended author information available on the last page of the article

and thus an indication of the possible treatment. However, this classification of fractures has been developed primarily for spinal fractures of skeletally healthy humans with a high exposure of energy to the axial skeleton. The discriminatory power for osteoporotic fractures is limited. Moreover, the osteoporotic spine may behave biomechanically different from a healthy axial skeleton; thus, fractures show different clinical and radiological outcomes.

Consequently, a distinct evaluation and corresponding fracture classification are warranted for these patients. In 2018, the "Osteoporotic Fracture working group" affiliated with the German Society for Orthopaedic and Trauma Surgery (DGOU) introduced a novel classification system specifically for osteoporotic thoracolumbar vertebral body fractures [7]. This classification organizes fractures into five categories, aiming to provide insights into fracture stability and inform therapeutic decisions. While smaller validation studies have demonstrated reliability and reproducibility, a larger-scale validation is necessary to further substantiate these findings.

It is essential to consider that in other large international validation studies, such as within the AO spine upper cervical classification and the AO spine sacral injury classification system, the reliability of the classifications was found to vary significantly based on factors such as demographics, location/region, surgical training, subspecialty, experience, and trauma center affiliation [8–12]. This underscores the importance of addressing these variables when validating the osteoporotic thoracolumbar vertebral body fracture classification system.

Previously, we published the effect of surgical training, subspecializing and clinical experience [13]. The main motivation for the paper is to provide baseline information about geographical variability in reliability and/or accuracy of ratings which can be used to guide further steps (e.g., training development). The present study aims to investigate the influence of geographic region on the reliability and reproducibility of the AO Spine-DGOU Osteoporotic Fracture Classification System in a large international validation group.

Material and methods

AO Spine-DGOU Osteoporotic Fracture Classification System

This classification was conceived as a straightforward, user-friendly framework (Fig. 1). The AO Spine-DGOU Osteoporotic Fracture Classification System organizes osteoporotic vertebral fractures into five different categories. The severity of the deformity of the vertebral body essentially determines the classification. A non-deformed vertebral body with fracture edema visible on MRI describes an OF 1 fracture. An OF 2 fracture is characterized by a deformation of one endplate only with or without minor involvement of the posterior wall (< 1/5 of the posterior wall). If one endplate of the vertebral body is deformed, but there is a distinct involvement of the posterior wall (> 1/5 of posterior wall), this fracture is classified as OF 3. An OF 4 fracture describes a deformation of both endplates with or without involvement of the posterior wall leading to either a complete structural loss of the vertebral body, a pincer-type deformation, or fish-bone deformation. An OF 5 injury also describes a deformation with additional failure of either the posterior or anterior tension band. The classification also shows an algorithm as an aid to correct classification (Fig. 1).

Reproducibility and reliability evaluation

For the validation, the present work strictly follows the previously published methodology and recommendation of Lambrechts et al. [14].

The full validation reached 320 participants. The majority of participants were practicing spine surgeons; we published an additional analysis on the professional background of the participants [13]. All participants were invited in advance to watch a video on the correct application and classification. In addition to the video, they also had the official poster (Fig. 1) to illustrate the classification as an aid. Before the start of the validation, all participants were given the opportunity to ask questions to the instructor. The instructor was one of the original developers of the classification. The webinar was conducted in English.

As previously published, the validation was conducted via an online webinar. After a short introduction by the instructor, the participants were given the opportunity to practice the handling in the context of training cases. Subsequently, the 27 cases were presented by means of a key-image and subsequently by means of CT and (if available or relevant) MRI in sagittal, coronal, and axial slices. MRI was only required for OF1 fractures; in other cases, CT imaging was sufficient. Only imaging was provided, without additional clinical information, to avoid bias in the classification. The participants were then able to give their rating for the respective case in REDCap (REDCap Software – Version 6.5.2 - © 2023 Vanderbilt University). During the session, participants were able to submit comprehension questions or procedural questions via a chat function.

The second rating took place with rearranged cases 4 weeks after the initial assessment to avoid recall bias.

Gold standard committee

Members of the Osteoporotic Fracture working group of the DGOU provided the gold standard classification for each

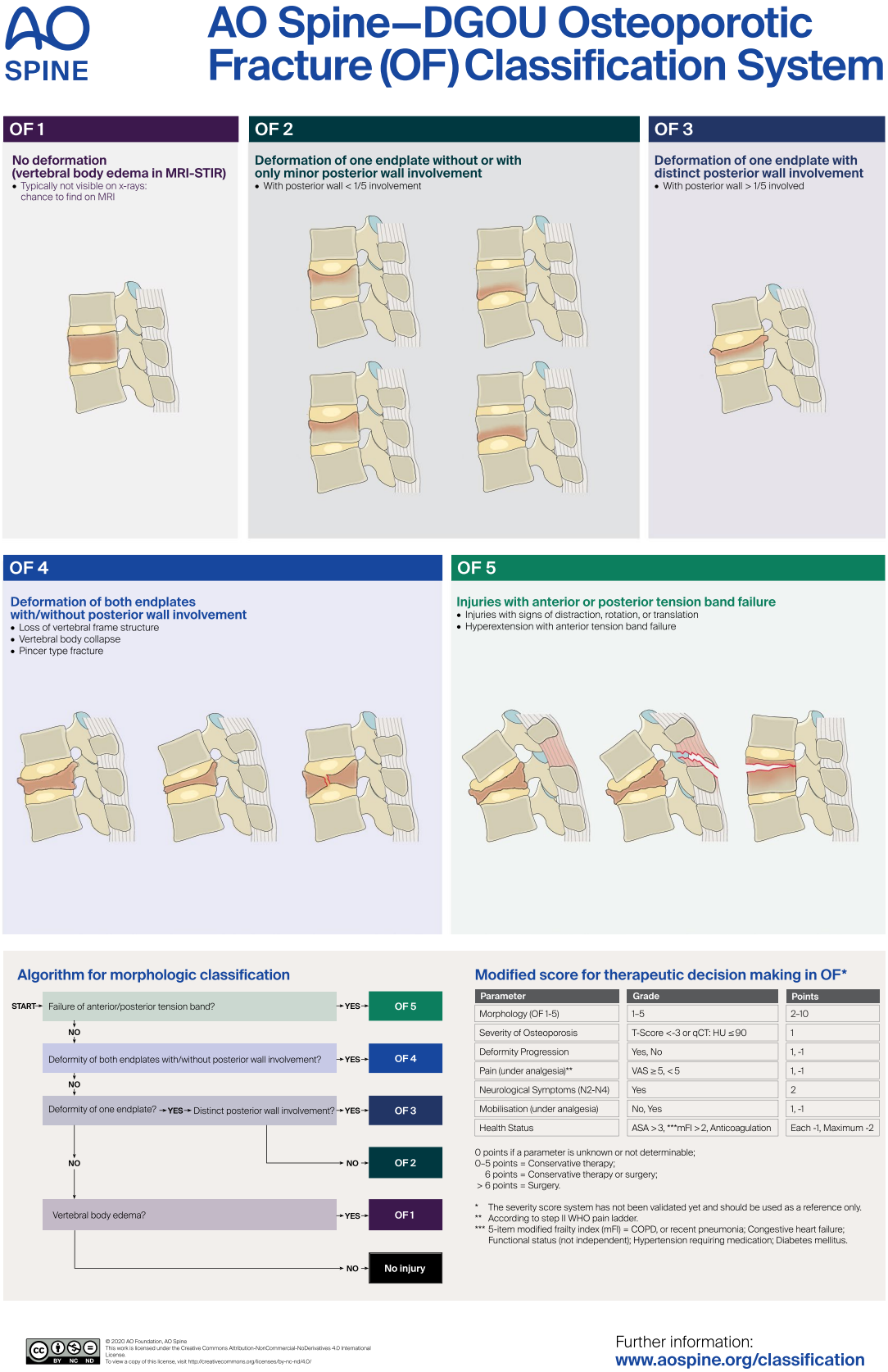


Fig. 1 Classification poster

injury on radiological imaging. Prior to evaluation of the injuries by the study participants, the gold standard committee reached unanimous agreement on the classification of each injury film.

Statistical analysis

The percentage of agreement between study participants and the gold standard committee was calculated for classification of images. Comparisons between groups were based on the AO region, which includes Africa, Asia, Central/South America, Europe, the Middle East, and North America.

Kappa coefficients (κ) were calculated to evaluate the agreement among different study participants (interobserver reliability) and within the same participant after a 4-week interval (intraobserver reproducibility). All reported kappa values utilized Fleiss' kappa coefficient, which also accommodates missing assessments and comparisons involving more than two validation members [15]. The reliability and reproducibility were interpreted using Landis and Koch's convention, which classifies kappa values as "slight" (< 0.2), "fair" (0.2–0.4), "moderate" (0.41–0.60), "substantial" (0.61–0.8), and "excellent" (0.81–1.0) [16]. To identify potentially relevant relationships, Pearson's chi-square test or Fisher's exact test was employed.

We conducted a post-hoc analysis to compare the German-speaking participants with the other participants. Analysis entails a two-sided unpaired two-sample Wilcoxon test (normal approximation with continuity correction) to determine:

- if intra-rater reliability in Europe is significantly higher than in all other regions.
- if intra-rater reliability in German-speaking countries in Europe is significantly higher than in non-German-speaking countries in Europe.

Both tests are significant at $p < 0.05$.

Results

Participants and ratings

A total of 320 participants rated the fracture type (OF 1/OF 2/OF 3/OF 4/OF 5) for at least one of the 27 cases presented on at least one occasion. Two rounds of scoring were conducted with identical cases in different order. If an evaluator evaluated a case more than once within the same evaluation round (e.g., due to an interrupted and resumed questionnaire session), the first evaluation was used.

The fractures were classified after gold standard evaluation according to Table 1.

Table 1 Fracture type according to gold standard classification

Type	Frequency
OF 1	7.4% (2/27)
OF 2	25.9% (7/27)
OF 3	14.8% (4/27)
OF 4	33.3% (9/27)
OF 5	18.5% (5/27)

The largest proportion of participants came from Europe (40.6%), a quarter were from Asia (25.6%), and the third largest group was from Central and South America (12.5%), followed by Africa (7.8%), North America (6.9%), and the Middle East (6.6%). A detailed overview is shown in Table 2.

The pooled classification results by region are shown in Table 3 and Fig. 2.

Accuracy of the classification

Globally, the evaluation showed moderate to substantial agreement with the gold standard. The results are shown in detail in Tables 4 and 5. In the first assessment, the global agreement was still moderate at kappa 0.58, whereas in the second assessment, an improvement to kappa 0.61 and thus to a substantial level was observed. There was a considerable and significant difference between the individual regions in both assessments. Participants from Europe had the highest level of agreement and substantial agreement in both assessments ($k = 0.64$; $k = 0.66$). Substantial agreement was also achieved by participants from North America in the second assessment ($k = 0.61$). The other regions achieved moderate agreement in both assessments, with a range of $k = 0.46$ (Africa during the first assessment) to $k = 0.58$ (Central and South Africa during the first assessment). Although the distribution of the number of raters for different regions is imbalanced, we could not find evidence for a relevant correlation between the number of raters and interrater agreement Fleiss Kappa statistics (data not shown).

There were also considerable differences in the assessment within the subgroups of the classification. Within the second assessment, all subgroups showed substantial agreement except for the OF3 fractures. OF 3 fractures even showed only "fair" agreement in some cases. The kappa in the second assessment for OF 3 fractures was less than 0.4 in all regions except Europe and Central and South America. OF 4 injuries were best recognized in all regions with a $k = 0.69$ (range $k = 0.76$–$k = 0.57$).

Intra-rater reliability

In the box-plot diagram (Fig. 3), it can be observed that the intra-rater agreement among European participants was

Table 2 Allocation of participants

	Participants (total number)	Ratings (total)	Ratings (%)
Africa	7.8% (25/320)	7.1% (1022/14,419)	7.1
Asia	25.6% (82/320)	26.3% (3789/14,419)	26.3
Central and South America	12.5% (40/320)	11.7% (1688/14,419)	11.7
Europe	40.6% (130/320)	42.0% (6051/14,419)	42.0
Middle East	6.6% (21/320)	5.9% (845/14,419)	5.9
North America	6.9% (22/320)	7.1% (1024/14,419)	7.1
	100% (320/320)	14,419	100

Table 3 Participants and ratings

Classification	Africa	Asia	Central and South America	Europe	Middle East	North America
OF 1	8.32% (85/1022)	8.26% (313/3789)	8.59% (145/1688)	8.05% (487/6051)	8.64% (73/845)	8.3% (85/1024)
OF 2	25.44% (260/1022)	27.18% (1030/3789)	27.55% (465/1688)	27.1% (1640/6051)	27.46% (232/845)	28.81% (295/1024)
OF 3	16.63% (170/1022)	15.84% (600/3789)	13.98% (236/1688)	15.24% (922/6051)	14.2% (120/845)	15.92% (163/1024)
OF 4	28.57% (292/1022)	28.05% (1063/3789)	28.91% (488/1688)	29.58% (1790/6051)	29.23% (247/845)	28.71% (294/1024)
OF 5	21.04% (215/1022)	20.67% (783/3789)	20.97% (354/1688)	20.03% (1212/6051)	20.47% (173/845)	18.26% (187/1024)

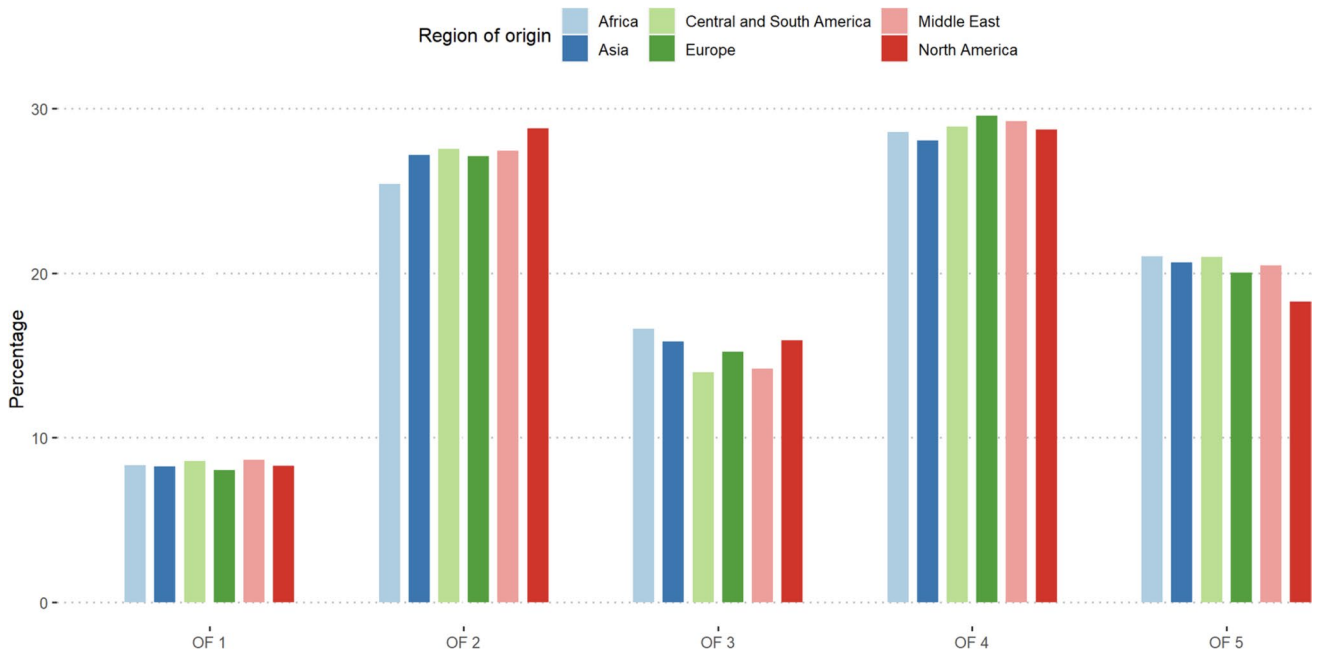


Fig. 2 Fracture type ratings according to the region—pooled across the ratings

overall very good with low dispersion. Descriptively, there is also a difference in intra-rater reliability depending on the region of the participants. The group with the highest agreement between the first and second ratings comes from Europe, followed by America and Central and South America. The remaining regions performed comparably. After chi-square testing, a significant difference in the "almost perfect" intra-rater agreement group is revealed in the breakdown of participants with Europeans being the most reliable ($p = 0.017$) (Supplemental File 1).

Agreement with gold-standard

In general, there was a high agreement of the classification with the gold standard in all regions. In Europe, the agreement in both assessments was highest at about 80%.

Table 4 Results of the first evaluation for interrater-agreement

Classification	Africa	Asia	Central and South America	Europe	Middle East	North America
Globally	0.46	0.48	0.58	0.64	0.57	0.58
OF 1	0.49	0.47	0.52	0.62	0.54	0.57
OF 2	0.45	0.45	0.58	0.61	0.48	0.54
OF 3	0.36	0.36	0.41	0.54	0.39	0.41
OF 4	0.57	0.59	0.70	0.75	0.72	0.69
OF 5	0.39	0.47	0.59	0.65	0.62	0.63

Table 5 Results of the second evaluation for interrater-agreement

Classification	Africa	Asia	Central and South America	Europe	Middle East	North America
Globally	0.51	0.51	0.55	0.66	0.55	0.61
OF 1	0.67	0.50	0.53	0.66	0.50	0.61
OF 2	0.50	0.49	0.52	0.63	0.54	0.64
OF 3	0.30	0.35	0.44	0.51	0.36	0.38
OF 4	0.66	0.64	0.66	0.76	0.69	0.68
OF 5	0.42	0.50	0.56	0.68	0.58	0.66

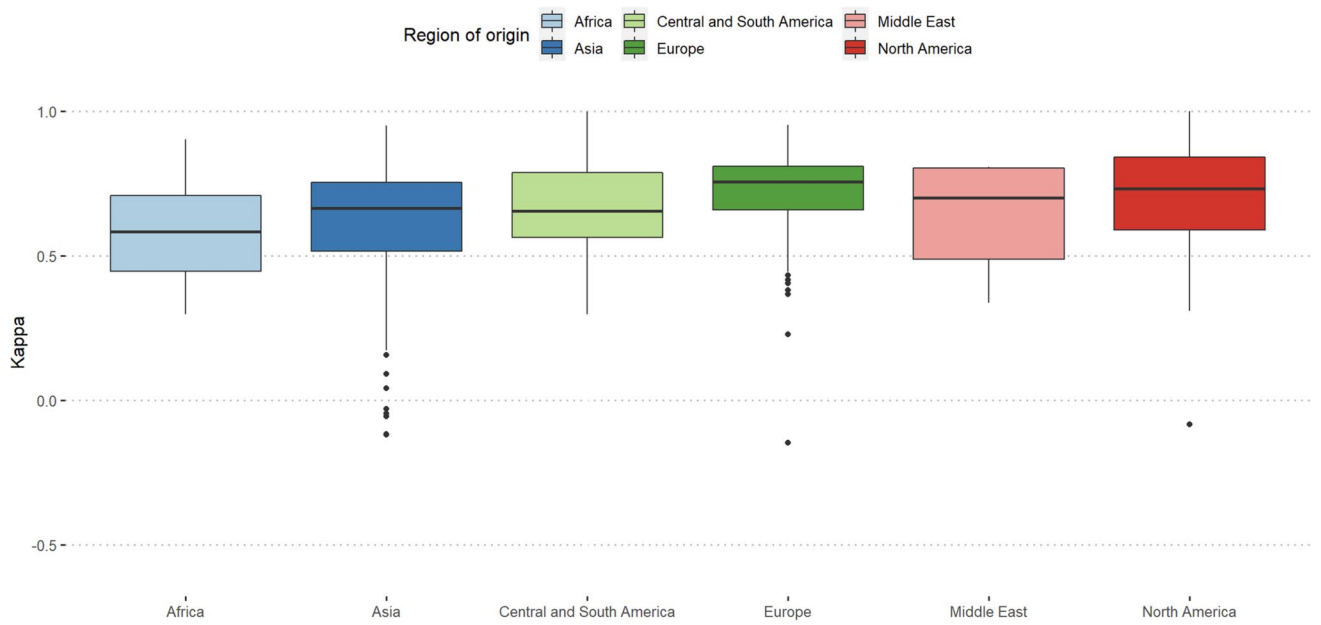


Fig. 3 Box-plot diagram showing the intra-rater reproducibility by region

The lowest level of agreement in both assessments was found among participants from Africa, at 70% in each case. Regarding the subgroups, the lowest agreement for OF 3 fractures was seen in participants from the Middle East and Africa (54% each). The greatest agreement was found for OF 5 fractures in participants from Europe with 88%. Detailed results can be found in Tables 6 and 7.

Post hoc analysis

Our post hoc analysis revealed a significant difference in reliability between the European participants and the pooled participants from other regions. The kappa for the European participants indicated substantial agreement with a value of 0.72, while the pooled participants from

other regions also showed substantial agreement with a kappa of 0.61, albeit less in comparison to the European participants ($p < 0.001$). In the subgroup analysis of the European participants, there was also a significant difference between the German-speaking participants and the rest of the European participants ($p = 0.0026$). The German-speaking participants demonstrated a significantly better result with a kappa of 0.79 compared to the remaining European participants (kappa 0.70). Figures 4 and 5 display the results as box plots.

Table 6 Agreement to gold-standard allocated to region first assessment

Classification	Africa	Asia	Central and South America	Europe	Middle East	North America	*P*-value
Globally	69.7% (372/534)	71.6% (1508/2107)	76.5% (733/958)	80.3% (2612/3251)	74.2% (325/438)	76.5% (390/510)	0
OF 1	66.7% (28/42)	68.8% (106/154)	71.6% (53/74)	80% (188/235)	78.1% (25/32)	75% (27/36)	0.1427
OF 2	67.9% (93/137)	69.2% (379/548)	78.3% (191/244)	79% (668/846)	72.2% (83/115)	79.1% (106/134)	0.0002
OF 3	68.4% (54/79)	61.9% (193/312)	58.3% (84/144)	70.7% (342/484)	56.9% (37/65)	60% (45/75)	0.0151
OF 4	72.1% (129/179)	74.6% (523/701)	79.8% (256/321)	82.9% (898/1083)	78.8% (115/146)	78.2% (133/170)	0.0003
OF 5	70.1% (68/97)	78.3% (307/392)	85.1% (149/175)	85.6% (516/603)	81.2% (65/80)	83.2% (79/95)	0.0016

Table 7 Agreement to gold-standard allocated to region second assessment

Classification	Africa	Asia	Central and South America	Europe	Middle East	North America	*P*-value
Globally	70.7% (345/488)	72% (1211/1682)	75.5% (551/730)	80.6% (2257/2800)	74% (301/407)	77.6% (399/514)	0
OF 1	67.6% (25/37)	70.2% (87/124)	76.4% (42/55)	83.6% (173/207)	66.7% (20/30)	84.2% (32/38)	0.0217
OF 2	70.4% (88/125)	73.6% (320/435)	75.1% (142/189)	80.7% (585/725)	70.5% (74/105)	80.5% (107/133)	0.0093
OF 3	53.9% (41/76)	57.8% (145/251)	57.4% (62/108)	63.5% (265/417)	54.1% (33/61)	59.7% (46/77)	0.4117
OF 4	77.2% (125/162)	75.8% (424/559)	79.4% (193/243)	83.6% (781/934)	80.7% (109/135)	79.5% (136/171)	0.0119
OF 5	75% (66/88)	75.1% (235/313)	83% (112/135)	87.6% (453/517)	85.5% (65/76)	82.1% (78/95)	0.0001

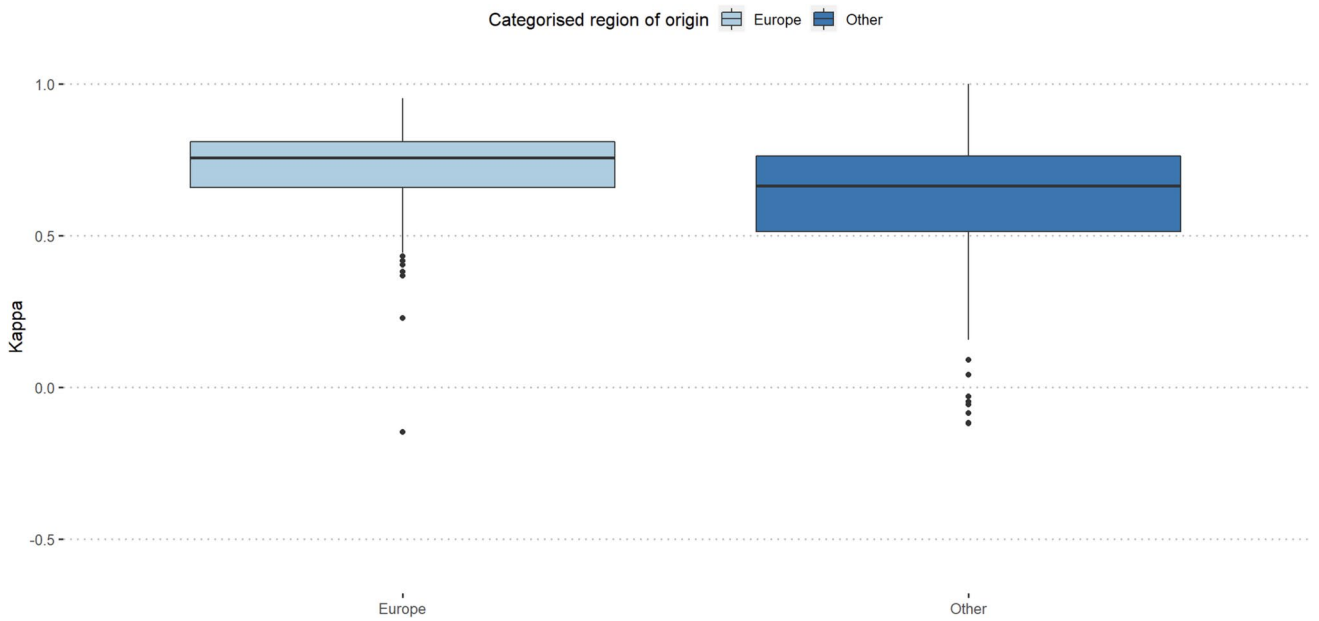


Fig. 4 Box-plot diagram showing the intra-rater reproducibility categorized by region of origin (Europe vs other)

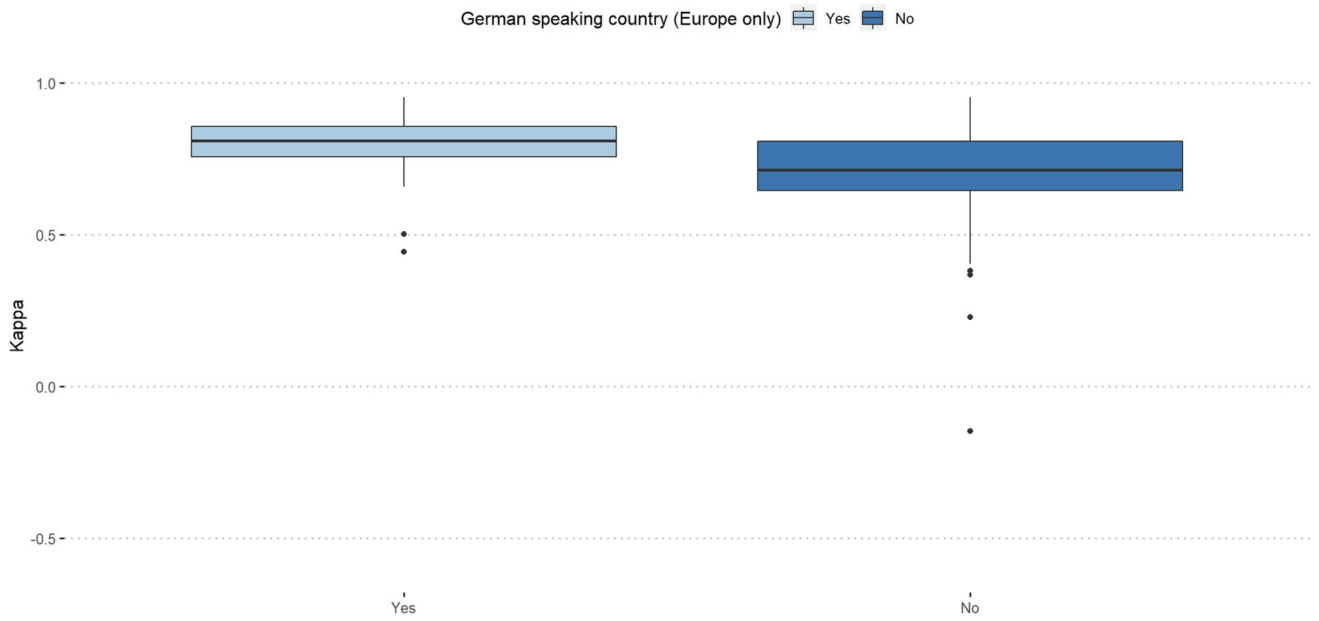


Fig. 5 Box-plot diagram showing the intra-rater reproducibility within Europe categorized by language (German vs other)

Discussion

A unified classification for injuries is essential. A classification should facilitate discussion among involved professionals, thus creating a common language. However, a classification also requires clinical significance. Ideally, it should include information on the severity of the injury and allow conclusions about the outcome and prognosis. It must aid in clinical decision-making. As proposed by Audigé et al., a new classification system should undergo a three-step validation process: (1) expert category descriptions, (2) a multicenter agreement assessment by a representative panel of potential scheme users, and (3) a prospective clinical study assessing its clinical value [17]. The study presented here represents step two to validate the AO Spine-DGOU Osteoporotic Fracture Classification in an international setting.

This study showed moderate to substantial agreement with the gold standard in a large international validation, along with generally high intra-rater reliability. Since its publication in 2018, the OF Classification has been successfully applied in a larger prospective study, where Ulrich et al. [18] compared various treatment concepts for osteoporotic vertebral fractures.

In addition to morphological fracture classification, a score accounting for patient-specific factors should be considered to guide treatment decisions. This score is already utilized, particularly within German-speaking physicians, with a planned multicenter validation.

In general, the results of this study are consistent with the results of studies that have also examined regional differences in reliability [19, 20]. The commonalities are a weaker performance of participants from Africa and the Middle East compared to participants from Europe and North America. Unfortunately, little is known and researched about the causes and background of this tendency. Possible causes could lie in variations in the structure, focus, or methodology of surgical training across regions. These differences should be investigated in further studies.

Osteoporosis

Osteoporosis in general and fracture classification in particular both have some peculiarities in themselves. Fracture diagnosis might be challenging [21]. The occurrence of osteoporotic fractures is a particular problem in the population over 60 years of age. Regions with a higher prevalence of osteoporotic fractures may have healthcare professionals who are more accustomed to dealing with these types of fractures, potentially leading to improved classification accuracy. Due to the higher proportion of the elderly population in the industrialized nations, treating physicians in these countries are confronted with these fractures more frequently [2]. This effect could have an influence on the findings presented in this article. Furthermore, in more developed regions, health care systems offer more treatment options, and patients themselves may demand more extensive treatment.

Our post hoc analysis showed a significant difference between the German-speaking participants and the remaining participants. This validation study is the first to investigate a connection with inter-rater reliability in relation to the associated language area and showed a higher interrater

reliability within the region of origin. The proposed classification was first presented in 2017 by Schnake et al. [22]. The classification was developed and promoted by the Spine Section of the German Society for Orthopaedic and Trauma Surgery (DGOU). The origin of this nationally very strong and far-reaching organization has led to a faster and higher penetration within German-speaking countries. This in turn may have an impact on the willingness and quality of study participation. In this context, it is interesting to note that there was an improvement in reliability from the first to the second assessment. It is possible that a more in-depth examination of the classification in the second assessment led to an improvement. It would be interesting to evaluate the learning effect of further validation rounds.

Subgroup considerations

The present study showed both regional differences and differences between subgroups. It seems to have been a challenge for the participants to correctly identify OF 3 injuries. In the proposed classification, the difference between OF 2 and OF 3 is the extent of involvement of the posterior wall of the vertebral body. This aligns with qualitative feedback from participants, who reported particular difficulty in distinguishing OF 2 from OF 3. A possible explanation could be the method of validation. In the study performed here, the assessment was performed online. The images of the patients were played back at a standardized speed, and there was no possibility to view individual sequences again or for a longer time. It was also not possible to measure the posterior wall in comparison to the adjacent vertebral bodies. In the past, it was shown that this kind of validation is well suited to reach a high and diverse number of practitioners, but the above-mentioned limitations could explain some effects. Another conceivable explanation why OF 3 fractures were classified worse in comparison may be the frequency of occurrence in everyday life compared to this validation study. To achieve good statistical power, the groups were distributed more or less equally. In our clinical experience, more OF 3 fractures occur in everyday life. Therefore, clinical experience may have had an influence on the expectation of the participants and may have led to an under/over classification of OF 3 fractures.

Smaller, monocentric or regional studies also showed that distinguishing between OF 2 and OF 3 is challenging [23, 24]. A more discriminating definition may be developed in the future.

Limitations

This study has several limitations. Firstly, the online assessment method may have impacted the accuracy of classifications, as participants could not view individual sequences multiple times or measure the posterior wall precisely. Additionally, the sample may have selection bias and limited case scenarios, suggesting a need for a more diverse participant pool and broader range of fracture cases in future research. Exploring language, cultural factors, and learning effects through standardized assessment conditions and incorporating qualitative participant feedback could further refine the classification system. The relatively small sample sizes in some regions (e.g., Africa, Middle East) limit the confidence in region-specific conclusions and may reduce generalizability. Despite these limitations, this study presents valuable opportunities for future research and improvement. Selection bias and limited case scenarios can be addressed by recruiting a more diverse sample and using a wider range of fracture cases. Ensuring standardized assessment conditions and exploring the impact of language, cultural factors, and learning effects can help refine the classification system. Additionally, incorporating qualitative data, such as participant feedback, can provide insights into specific challenges and areas for enhancement. By acknowledging and addressing these limitations, the study's findings can contribute to the ongoing refinement and validation of the AO Spine-DGOU Osteoporotic Fracture Classification System.

Conclusion

This study demonstrates that the AO Spine-DGOU Osteoporotic Fracture Classification System has overall moderate to substantial reliability and reproducibility across different geographic regions. However, regional differences in classification performance were observed, suggesting that factors such as demographics, location, surgical training, subspecialty, and clinical experience may influence the system's application. European participants showed the highest agreement with Germany outperforming all other regions. This study underscores the need for addressing regional factors to improve the system's global clinical utility. Despite some limitations, the AO Spine-DGOU Osteoporotic Fracture Classification System provides a valuable tool for assessing osteoporotic thoracolumbar vertebral body fractures, informing therapeutic decisions, and facilitating international collaboration in the management of osteoporotic patients. Future research should focus on addressing the identified limitations and further validating the classification system, as well as comparing its performance with other existing classification systems to ensure the most effective and evidence-based care for patients with osteoporotic fractures.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s00198-026-08112-4.

Acknowledgements The authors of the manuscript would like to thank Olesja Hazenbiller for her assistance in developing the methodology

and providing support during the validation. This study was organized and funded by AO Spine through the AO Spine Knowledge Forum Trauma & Infection, a focused group of international spinal experts. AO Spine is a clinical division of the AO Foundation, which is an independent medically guided not-for-profit organization. Study support was provided directly through AO Network Clinical Research.

AO Spine-DGOU international validation Group Maximo-Alberto Diez-Ullloa, Marco Burkhard, Mahmoud, Haribabu Subramaniam, Alsammak Wael, Panneerselvam Keerthivasan, Luis Maria Romero-Munoz, Sokol Trungu, Sebastian F Bigdon, Chee Huan Pan, Banerjee Purnajyoti, Kumar Sharma, Mantu Jain, Ehab Shiban, Risto Todorov, Adetunji Toluse, Juliano Chaves, Maria Celeste Zamora, Satish Muthu, Roberto Berebichez-Fridman, Miltiadis Georgiopoulos, Eduardo Laos, Waeel Hamouda, Hüseyin Sina Coşkun, Margetis Konstantinos, Jaime Segura, Sung-Joo Yuh, Davide Bizzoca, Adesh Shrivastava, Valentin Neuhaus, Nasser Khan, Mengistu M Mengesha, Elena Virginia Colombo, Jagadish Thapa, Michal Ziga, Pragnesh Bhatt, Rohan Gala, Robert Huang, Benito Dario, Arshad Ayesha, Ateef Ataulla Khan, Manjul Tripathi, Luís Marques, Ataollah Shahbandi, Harsh Deora, Luigi Pansini, Arne Mehrkens, Paloma Bas, Federico Sartor, Juan Lourido, Bishnu Dev Sharma, Marcello Barbato, Ricky Rasschaert, Inan Uzunoglu, Mahmoud Alkharsawi, Sergio Zylbersztejn, Dimitrios Ntourantonis, Ikwutah Emmanuel, Parmenion Tsitsopoulos, Selim Ayhan, Francisco Verdu-Lopez, Konstantinos Gousias, Luis Miguel Duchén, Woon Tak Yuh, Nader Hejrati, Peramunugamage Dilanka, Falethu Sukati, Omar Elwardany, Alexander Durst, Rafid Al-Mahfoudh, Eugen Cezar Popescu, Yasunori Sorimachi, Manoj Khatri, André Barros, Konstantinos Sidiropoulos, Reuben Soh, Jose-Carlos Sauri-Barraza, Sayid omar Mohamed, Murray Echt, Andrea Redaelli, Guilherme Godoy, Stipe Corluka, Senan Faruk, Paulo Pereira, David Dalton, Paul Thorpe, Anubhav Amin, Prasad Shetty Ajoy, Tarun Suri, Ratko Yurac, Petr Vachata, Lukas Dominik Burger, Wongthawat Liawrungrueang, Ahmed Hassan, Abduljabbar Alhammoud, Vijay Kumar Loya, Leon-Gordian Koepke, Madhivanan Karthigeyan, Martin Nikolaus Stienen, Bernhard Ullrich, Dave Anthony Dizon, Ahmad Nigad, Mark Murerwa, Erkin Sönmez, Nicolas Lauper, Salvatore Russo, Peter Ferlic, Mohammed Riaz, Michele Parolin, Lady Lozano Cari, Etxeberria Igone, Dilip Gopalakrishnan, Joost Rutges, Richard A Lindtner, Joseph Asuquo, Vito Fiorenza, Peter Loughenbury, Pumibal Wetpiriyakul, Jarkko Halme, Emiliano Vialle, Mario Ganau, Derek Cawley, Juan Gabriel Limpin De Leon, Michel Triffaux, Eran Steinberg, Sanjay Hajare Swapnil, Alvin Pun, Peter Obid, Lincoln Liu, Monsef Jad Bou, Babak Shariati, Nilay Chhasatia, Mohammed Imran, Akash Mohapatra, Sultan Alkalbani, Claudio Lamartina, Omar Marroquin-Herrera, Ralph T. Schär, Giorgio Lofrese, Muñoz M Juan, Shaffaf Kareem, Shah Waliullah, Mateusz M. Weclewicz, Kim-Soon Oh, Mohamed A.R. Soliman, Madhan Jeyaraman, Arulkumar Nallakumarasamy, Antonio Krieger, Stefano Carnesecchi, Charles A Carazzo, Alessandro Ramieri, Pradeep Kafle, Kiran Gurung, Fergus McCabe, Irfan Shujah, Rakesh Kumar, Konstantinos Paterakis, Stacey Darwish, Alecio Barcelos, Kyaw Linn Linn, James Bernatz, Zdenek Klezl, Shukla Devansh, FERAS QUMQUMJI, Elma Suljagic-Ljubovic, Hector Aceituno, Ramy Sherif, Daniel Cruz, Barry Kweh, Kamil Krzysztof Siedlecki, Ugochukwu Uchendu, Javier Altamirano, Najib El Tecle, Takeshi Aoyama, Martin Tejeda, Ravi Sreenivasan, Amitava Biswas, Juan A Uribe, Gumar Davletgaleev, Julian Scherer, Joshua Decruz, PS Khoh, Ivan Marintschev, Chumpon Jetjumnong, Stavridis Stavros I, Paula Valerie Ter Wengel, Pathinathan Kalaventhan, Manuel Tavares de Matos, Jeasson Javier Pérez Rios, Pasquale Donnarumma, Udit Agrawal, Girinivasan Chellamuthu, Yu-Mi Ryang, Christogeorgos Paris, Dan Schneider, Daniel Sauer, Nicola Montemurro, Aravind Ravichandran, Mohammad Elbaroody, Gerardo Zambito-Brondo, Ousmane Toumany Cherif Haidara, Hossein Elgafy, Lykourgos Kollintzas, Koopong Siribumrungwong, Hitesh Dabasia, Pedro Kalo, Sergey Mlyavykh, Philipp Schleicher, Oscar Garcia-Casas, Mike Achor, Nikolaos Gkantsinikoudis, Darko Perovic, Ahmed Ezzat, Giuseppe Francesco Papalia, Zion Hwang, Michael S Hisey, Pedro Luis Bazán, Lukas Grassner, Federico Landriel, Youheng Ou Yang, Zahid Rehan, Hugo Marcelo Wirz, Samy Samer, Victor Martin-Gorgojo, Alon Grundshtein, Dana Niry, Santiago Hem, Thami Benzakour, Raju Reddycherla Naga, Hansdieter Neumann.

Funding Open access funding provided by University of Bern.

Data availability The data from the study is available from the corresponding author upon reasonable request.

Declarations

Conflicts of interest None.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc/4.0/.

References

1. Lippuner K, Johansson H, Kanis JA, Rizzoli R (2009) Remaining lifetime and absolute 10-year probabilities of osteoporotic fracture in Swiss men and women. Osteoporos Int 20(7):1131–1140. https://doi.org/10.1007/s00198-008-0779-8
2. Kanis JA, Odén A, McCloskey EV et al (2012) A systematic review of hip fracture incidence and probability of fracture worldwide. Osteoporos Int 23(9):2239–2256. https://doi.org/10.1007/s00198-012-1964-3
3. Bigdon SF, Saldarriaga Y, Oswald KAC et al (2022) Epidemiologic analysis of 8000 acute vertebral fractures: evolution of treatment and complications at 10-year follow-up. J Orthop Surg Res 17(1):270. https://doi.org/10.1186/s13018-022-03147-9
4. Genant HK, Wu CY, van Kuijk C, Nevitt MC (2009) Vertebral fracture assessment using a semiquantitative technique. J Bone Miner Res 8(9):1137–1148. https://doi.org/10.1002/jbmr.5650080915
5. Vaccaro AR, Lehman RA, Hurlbert RJ et al (2005) A new classification of thoracolumbar injuries: the importance of injury morphology, the integrity of the posterior ligamentous complex, and neurologic status. Spine 30(20):2325–2333. https://doi.org/10.1097/01.brs.0000182986.43345.cb
6. Kepler CK, Vaccaro AR, Koerner JD et al (2016) Reliability analysis of the AOSpine thoracolumbar spine injury classification system by a worldwide group of naïve spinal surgeons. Eur Spine J 25(4):1082–1086. https://doi.org/10.1007/s00586-015-3765-9
7. Schnake KJ, Blattert TR, Hahn P et al (2018) Classification of osteoporotic thoracolumbar spine fractures: recommendations of the spine section of the German Society for Orthopaedics and Trauma (DGOU). Glob Spine J 8(2 Suppl):46S-49S. https://doi.org/10.1177/2192568217717972
8. Karamian BA, Schroeder GD, Lambrechts MJ et al (2022) The influence of regional differences on the reliability of the AO Spine Sacral Injury Classification System. Glob Spine J. https://doi.org/10.1177/21925682211068419

9. Karamian BA, Schroeder GD, Levy HA et al (2021) The influence of surgeon experience and subspeciality on the reliability of the AO Spine Sacral Classification System. Spine 46(24):1705–1713. https://doi.org/10.1097/BRS.0000000000004199
10. Canseco JA, Schroeder GD, Patel PD et al (2021) Regional and experiential differences in surgeon preference for the treatment of cervical facet injuries: a case study survey with the AO Spine Cervical Classification Validation Group. Eur Spine J 30(2):517–523. https://doi.org/10.1007/s00586-020-06535-z
11. Lambrechts MJ, Schroeder GD, Karamian BA et al (2023) Effect of surgical experience and spine subspecialty on the reliability of the AO Spine Upper Cervical Injury Classification System. J Neurosurg Spine 38(1):31–41. https://doi.org/10.3171/2022.6.SPINE22454
12. Lambrechts MJ, Schroeder GD, Karamian BA et al (2022) The AO spine upper cervical injury classification system: do work setting or trauma center affiliation affect classification accuracy or reliability? Injury 53(10):3248–3254. https://doi.org/10.1016/j.injury.2022.08.030
13. Scherer J, Bigdon SF, Camino-Willhuber G et al (2025) Validation of the AOSpine-DGOU Osteoporotic Fracture Classification – effect of surgical experience, surgical specialty, work-setting and trauma center level on reliability and reproducibility. Glob Spine J. https://doi.org/10.1177/21925682251331945
14. Lambrechts MJ, Schroeder GD, Karamian BA et al (2022) Development of online technique for international validation of the AO Spine Subaxial Injury Classification System. Glob Spine J. https://doi.org/10.1177/21925682221098967
15. Gwet KL (2014) Handbook of inter-rater reliability, 4th edition: the definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC
16. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174
17. Audigé L, Bhandari M, Hanson B, Kellam J (2005) A concept for the validation of fracture classifications. J Orthop Trauma 19(6):401–406. https://doi.org/10.1097/01.bot.0000155310.04886.37
18. Ullrich BW, Schenk P, Scheyerer MJ et al (2022) Georg Schmorl prize of the German spine society (DWG) 2022: current treatment for inpatients with osteoporotic thoracolumbar fractures—results of the EOFTT study. Eur Spine J. https://doi.org/10.1007/s00586-022-07519-x
19. Mitani K, Takahashi T, Tokunaga S et al (2023) Therapeutic prediction of osteoporotic vertebral compression fracture using the AO Spine-DGOU osteoporotic fracture classification and classification-based score: a single-center retrospective observational study. Neurospine 20(4):1166–1176. https://doi.org/10.14245/ns.2346776.388
20. Ullrich BW, Schnake KJ, Schenk P et al (2023) Clinical evaluation of the osteoporotic fracture treatment score (OF-Score): results of the evaluation of the osteoporotic fracture classification, treatment score and therapy recommendations (EOFTT) study. Glob Spine J 13(1_suppl):29S-35S. https://doi.org/10.1177/21925682221148133
21. Rutsch N, Amrein P, Exadaktylos AK et al (2023) Cervical spine trauma – evaluating the diagnostic power of CT, MRI, X-Ray and LODOX. Injury 54(7):110771. https://doi.org/10.1016/j.injury.2023.05.003
22. Schnake KJ, Bula P, Spiegl UJ et al (2017) Thorakolumbale Wirbelsäulenfrakturen beim alten Menschen: Klassifikation und Therapie. Unfallchirurg 120(12):1071–1085. https://doi.org/10.1007/s00113-017-0435-x
23. Schönrogge M, Lahodski V, Otto R et al (2022) Inter- and intraobserver reliabilities and critical analysis of the osteoporotic fracture classification of osteoporotic vertebral body fractures. Eur Spine J 31(9):2431–2438. https://doi.org/10.1007/s00586-022-07201-2
24. Quinteros G, Cabrera JP, Urrutia J et al (2022) Reliability evaluation of the new AO spine-DGOU classification for osteoporotic thoracolumbar fractures. World Neurosurg 161:e436–e440. https://doi.org/10.1016/j.wneu.2022.02.029

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Sebastian Frederick Bigdon[1,2] · Julian Scherer[3,4] · Gaston Camino-Willhuber[5] · Ulrich Spiegl[6] · Andrei Fernandes Joaquim[7] · Chhabra Harvinder Singh[8] · Marcel Dvorak[9] · Gregory Schroeder[10] · Mohammad El-Sharkawi[11] · Richard Bransford[12] · Lorin Michael Benneker[2] · Klaus John Schnake[13,14] · the AO Spine-DGOU international validation Group

✉ Sebastian Frederick Bigdon
Sebastian.bigdon@insel.ch

1 Department for Orthopaedics and Traumatology, Inselspital, University Hospital Bern, Bern, Switzerland

2 Department for Spine Surgery, Sonnenhof Spital, University of Bern, Bern, Switzerland

3 General Medicine & Global Health (GMGH), Department of Medicine, Division of Orthopaedic Surgery, Faculty of Health Sciences, Orthopaedic Research Unit (ORU), University of Cape Town, Cape Town, South Africa

4 Department of Traumatology, University of Zurich, Zurich, Switzerland

5 Policina Gipuzkoa, San Sebastian, Spain

6 Klinik Für Unfallchirurgie Und Orthopädie, Klinik München Harlaching, Sanatoriumspl. 2, Munich 81545, Germany

7 Neurosurgery Division, Department of Neurology, State University of Campinas, Campinas-Sao Paulo, Brazil

8 Sri Balaji Action Medical Institute, New Delhi, India

9 Combined Neurosurgical and Orthopedic Spine Program, Vancouver General Hospital, University of British Columbia, Vancouver, Canada

10 Department of Orthopaedic Surgery, Rothman Institute, Thomas Jefferson University Hospital, Philadelphia, PA, USA

11 Department of Orthopaedic and Trauma Surgery, Faculty of Medicine, Assiut University, Asyut, Egypt

12 Department of Orthopaedics and Sports Medicine, Harborview Medical Center, University of Washington, Seattle, WA, USA

13 Center for Spinal and Scoliosis Surgery, Malteser Waldkrankenhaus St. Marien Erlangen, Erlangen, Germany

14 Department of Orthopedics and Traumatology, Paracelsus Private Medical University Nuremberg, Nuremberg, Germany