Dear Editor of the World Journal of Orthopedics,

We do appreciate the efforts of the reviewers. We have tried to answer their questions. The reviewers' comments and our answers are listed below. Changes in the manuscript have been marked in grey.

On behalf of the authors,

Paul Gerdhem
Department of Reconstructive Orthopaedics, Karolinska University Hospital.

Reviewer's code: 01213075

Reviewer comments

……However, some issues needed to be addressed before considering publication in World Journal of Orthopaedics. 1. The gold standard was the key element for assessing inter- and intra-rater reliability in this study. How did the authors confirm the final "accurate" diagnosis of these studied radiograms? Did the authors check the diagnosis using other image modalities, e.g., CT scans?

Thank you for the question. The 'accurate' diagnosis was set by consensus between the Raters 1 and 2, if they did not independently choose the same classification. The raters had access to the images taken during the clinical care of the patient and consisted of plain radiographs, computer tomography and/or magnetic resonance images, images that had been accepted as sufficient for taking care of the patient. We have clarified this in the methods section on page 5, lines 91 and 92.

2. Page 12 The authors mentioned that "with specific training or many years of experience, a higher consistency in inter-rater reliability could be achieved". Therefore, the readers will be interested to know how much the difference was. In addition, can we

improve it?

We expect that experience would result in better classification reliability. For example, the difference in Kappa value between the rater with the shortest experience in the thoracolumbar classification test (Rater 7) and the rater with the highest Kappa value was 0.22 and 0.51 in the two different test efforts (Table 2). It is likely that education and possible more interactive information in the digital version of the fracture registry could result in higher Kappa values, but we have not tested this. We have added a sentence in the discussion on page 8, lines 149 - 151.

3. The results showed that the mean kappa coefficient for inter-rater reliability and intra-rater reliability ranged from 0.51 to 0.80. For the epidemiological studies, a low kappa coefficient may not reach the requirement. Therefore, it would be better to mention more detail the coefficient for various fracture types of fractures in text and provided clear statement. 4.

Answer: From the results in Table 2 it is apparent that thoracolumbar B-type injuries may be the most difficult to classify. This is not surprising since this corresponds to our clinical experience; to determine whether there is a rupture of the posterior ligament complex is not always easy. We have tried to further clarify this in the discussion. Changes made on page 8, lines 191-193.

…For study design, it would be better to do the third or more test and compared the results. This design mat help assessing the learning effect.

Answer: Unfortunately, we did not design the study for three test rounds, only two. We hope to perform another follow-up study with another design when the fracture register has been used for a while to see whether the reliability in clinical use is of the same level as in this study. As stated in the discussion, studies comparing treatment outcome should consider reclassifying images to ensure correct classification. No change has been made in the manuscript.

Reviewer's code: 02444715

Reviewer comments: The paper Inter- and intra-rater reliability of vertebral fracture

classifications in the Swedish fracture register is well written. I think it worth publication despite the limited value to many readers maybe if the authors discuss the inter and intra observer reliability in other registers and different classification systems that could add more value But I understand this data may not be available

Answer: Thank you for your comments. Compared to other classification efforts our data seem comparable if you exclude expertise settings. Our intention was to mimic the everyday clinic activity by using Raters with different experience levels. The study has taught us to be humble when using the registry data and strive for further improvement. We think that the other relevant studies have been included in the discussion. No changes have been made.

Other

For clarity, we have rephrased two sentences; page 6, lines 120 and 121, and page 8, lines 147 and 148