Follow us @BoneJointOpen

Freely available online OPEN ACCESS

BJO

META-ANALYSIS

# An increasing number of convolutional neural networks for fracture recognition and classification in orthopaedics

## ARE THESE EXTERNALLY VALIDATED AND READY FOR CLINICAL APPLICATION?

**L. Oliveira e Carmo,**
**A. van den Merkhof,**
**J. Olczak,**
**M. Gordon,**
**P. C. Jutte,**
**R. L. Jaarsma,**
**F. F. A. IJpma,**
**J. N. Doornberg,**
**J. Prijs,**
**Machine Learning Consortium**

*From Flinders Medical Centre and Flinders University, Adelaide, Australia*

### Aims

The number of convolutional neural networks (CNN) available for fracture detection and classification is rapidly increasing. External validation of a CNN on a temporally separate (separated by time) or geographically separate (separated by location) dataset is crucial to assess generalizability of the CNN before application to clinical practice in other institutions. We aimed to answer the following questions: are current CNNs for fracture recognition externally valid?; which methods are applied for external validation (EV)?; and, what are reported performances of the EV sets compared to the internal validation (IV) sets of these CNNs?

### Methods

The PubMed and Embase databases were systematically searched from January 2010 to October 2020 according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement. The type of EV, characteristics of the external dataset, and diagnostic performance characteristics on the IV and EV datasets were collected and compared. Quality assessment was conducted using a seven-item checklist based on a modified Methodologic Index for NOn-Randomized Studies instrument (MINORS).

### Results

Out of 1,349 studies, 36 reported development of a CNN for fracture detection and/or classification. Of these, only four (11%) reported a form of EV. One study used temporal EV, one conducted both temporal and geographical EV, and two used geographical EV. When comparing the CNN's performance on the IV set versus the EV set, the following were found: AUCs of 0.967 (IV) versus 0.975 (EV), 0.976 (IV) versus 0.985 to 0.992 (EV), 0.93 to 0.96 (IV) versus 0.80 to 0.89 (EV), and F1-scores of 0.856 to 0.863 (IV) versus 0.757 to 0.840 (EV).

### Conclusion

The number of externally validated CNNs in orthopaedic trauma for fracture recognition is still scarce. This greatly limits the potential for transfer of these CNNs from the developing institute to another hospital to achieve similar diagnostic performance. We recommend the use of geographical EV and statements such as the Consolidated Standards of Reporting Trials–Artificial Intelligence (CONSORT-AI), the Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence (SPIRIT-AI) and the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis–Machine Learning (TRIPOD-ML) to critically appraise performance of CNNs and improve methodological rigor, quality of future models, and facilitate eventual implementation in clinical practice.

**Cite this article:** *Bone Jt Open* 2021;2-10:879–885.

**Keywords:** Artificial intelligence, External validation, Convolutional neural networks, Machine learning, Deep learning

Correspondence should be sent to Jasper Prijs; email: jasperprijs@icloud.com

doi: 10.1302/2633-1462.210.BJO-2021-0133

*Bone Jt Open* 2021;2-10:879–885.

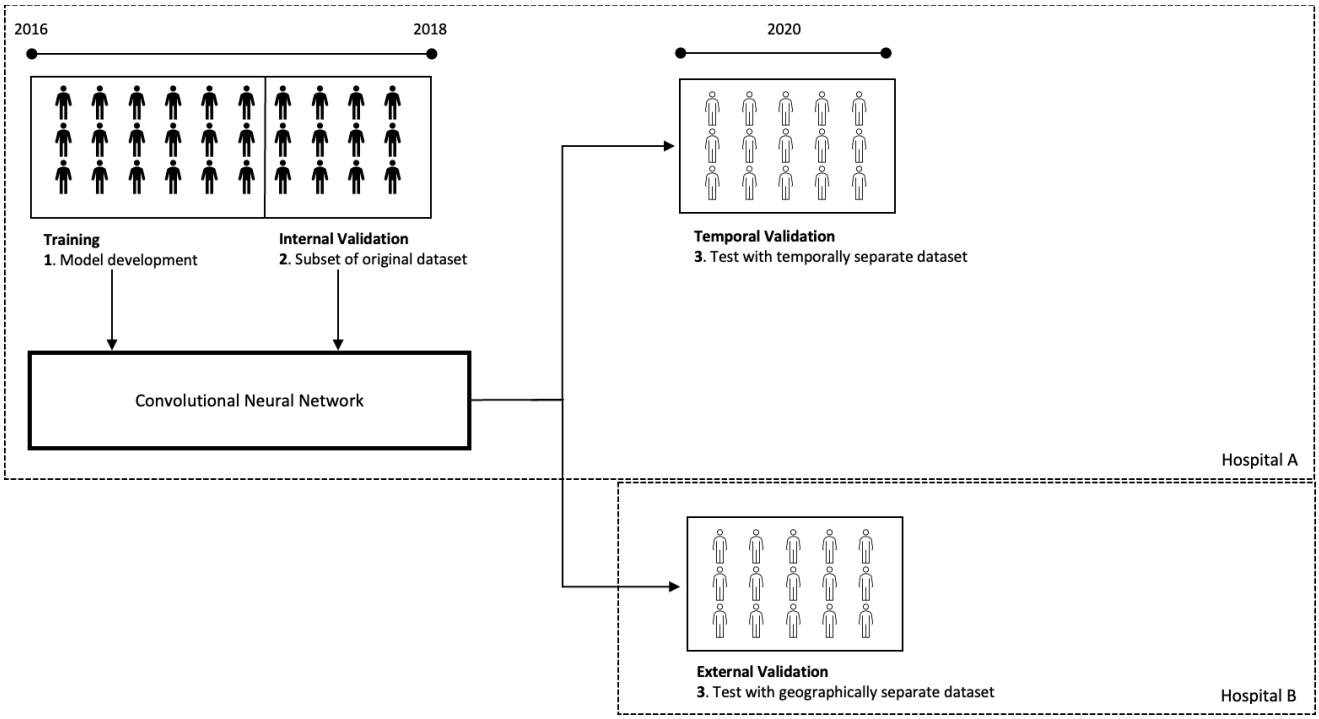


**Fig. 1**

Overview of common methodology used to develop and evaluate convolutional neural networks. Development starts with a database that is then split into a training (used for development) and internal validation set (used for evaluating performance). Subsequently, an external validation can be performed to assess generalizability of the model. This can be done using data from the same hospital but during a different time period (temporal) or, ideally, with data from another hospital (geographical).

## Introduction

An increase in the use of artificial intelligence (AI), particularly convolutional neural networks (CNNs, which mimic human visual cortex neurones), has been observed in medical imaging.[1-4] CNNs are able to process enormous volumes of data that surpass the pace of human observations, and in the field of orthopaedic trauma, CNNs have been reported to perform at the level of experienced orthopaedic surgeons and radiologists in detection and classification of distal radius, hip, proximal humerus, pelvis, and femur fractures.[5-11]

Performance of CNNs is evaluated using unseen data from the same initial longitudinal dataset used for training the CNN, called the test set or internal validation (IV) set. However, characteristics of these data are identical (i.e. same hospital and time period) to those used for model development. Algorithms generally perform poorly when externally validated with datasets from different institutions.[12-15] For example, in automated recognition of distal radius fractures, Blüthgen et al[6] reported decreased performance using the external validation (EV) set, while performance was excellent using the IV set. To explore weaknesses and generalizability of CNNs, two techniques can be used: geographical (separated by location) or temporal (separated by time) validation (Figure 1).[16] Arguably only the former truly represents EV that allows transfer of locally trained CNNs to applications in other hospitals.[17] Hence, geographical EV is considered the most stringent test of a model's performance and an important step before clinical implementation.[17-19]

Therefore, we aimed to answer the following: are CNNs for fracture recognition externally valid?; what are current methods applied for EV of CNNs for fracture recognition in the field of orthopaedic trauma?; and what are the reported performances of EV compared to the IV? To our knowledge, this is the first study to evaluate current applications of EV of CNNs used in orthopaedics for fracture detection and/or classification.

## Methods

A literature search according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement[20] (Figure 2) was conducted in the PubMed and EMBASE libraries for articles published between January 2010 and October 2020. The protocol was registered on PROSPERO (CRD42020216478) prior to screening the articles. Together with a medical librarian a search strategy was formulated (Supplementary Material).

Two reviewers (LOEC, AVDM) independently screened the titles and abstracts of the retrieved articles. They subsequently performed the full-text screening to check eligibility of articles with predetermined inclusion criteria. Disagreements between reviewers were solved by consulting a third reviewer (JP). Due to ambiguous

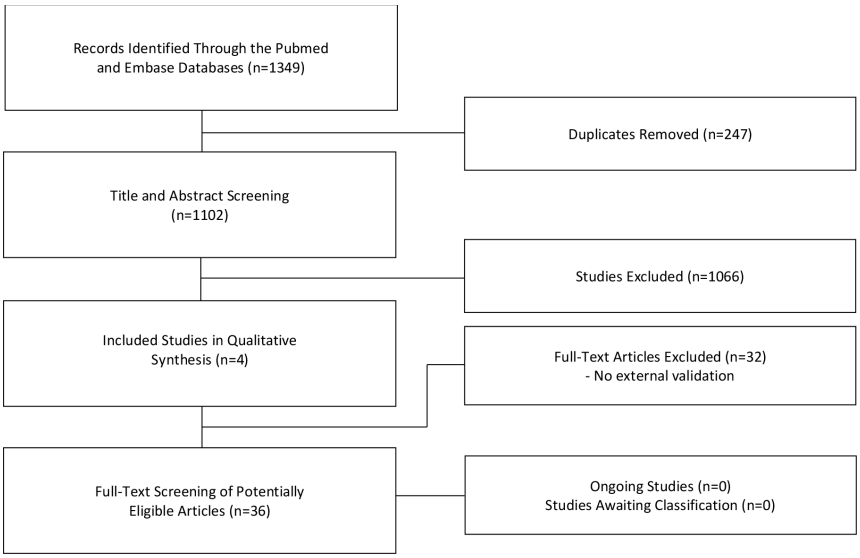


**Fig. 2**

This Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart describes the inclusion, exclusion, and selection of articles yielded in our search.

and unclear reporting of 'external validation' in articles found during the preliminary searches, all articles that reported the use of a CNN in orthopaedic trauma were selected for full-text review. From the full-text review, only articles that evaluated their CNN on a separate dataset—geographically or temporally—from that used during the CNN development (the "external validation") were included.

Inclusion criteria were journal articles reporting the use of a CNN in orthopaedic trauma including a form of EV, studies published after 2010, and written in English, Dutch, French, Portuguese, or Spanish. Exclusion criteria were the use of a CNN outside of an orthopaedic trauma setting, studies evaluating robot-assisted surgery techniques, studies with mixed cohort without clear subgroup reporting, review articles, letters to the editor, meeting abstracts, technique papers, and animal and cadaveric studies.

The search strategy yielded a total of 1,349 articles. After removal of duplicates, a total of 1,102 articles were screened. Overall, 36 studies reported the use of a CNN for fracture detection and/or classification and were selected for full-text review. Of these, four studies reported a form of EV (Figure 2). Additionally, no new studies were identified after manually screening the reference lists of included studies.

Quality assessment was performed by two independent reviewers (LOEC, AVDM). Disagreement was solved through discussion with a third reviewer (JP). The Quality Assessment of Diagnostic Accuracy Studies (QUADAS) criteria, a tool designed for the assessment of published diagnostic studies in systematic reviews, was not used because it was previously considered difficult to apply in machine-learning studies.[21,22] Due to lack of suitable tools assessing risk of bias for machine learning studies, we modified the Methodologic Index for NOn-Randomized Studies (MINORS) instrument, commonly used to assess the quality of cohorts or case-control studies.[23] The modified MINORS included the following items: disclosure, study aim, input features, ground truth, EV method, EV dataset, and performance metric. Screening and full-text review were conducted using Covidence (Veritas Health Innovation, Australia). Standardized forms were used to extract and record data (Excel v. 16.21; Microsoft, USA).

**Outcome measures.** To answer the primary research question, EV was defined as verification of model performance on a separate dataset, geographically or temporally, from that used for model development. To answer the secondary research question, type and characteristics of the EV set (dataset used, number of images, location and date of collection) were collected from the included articles. To answer the tertiary research question, performances of the CNN on the IV and EV datasets were collected and compared. All four studies were used to answer both secondary and tertiary research questions.

The following items were collected from all included studies: authors, year of publication, input feature (e.g. radiographs), radiological views if applicable (e.g. anteroposterior (AP)), anatomical location, output classes, ground truth label assignment, CNN model used, size, source and date of the initial dataset used for development, performance on IV set (e.g. area under the curve (AUC)), method of EV (temporal or geographical), size, source and date of EV set, and performance on the EV set (e.g. AUC).

Three studies[6,24,25] reported the area under the receiving operating characteristics curve (AUC-ROC) to evaluate IV and EV performance. The AUC is a common

**Table I.** Method of external validation, characteristics of datasets, and performance.

| Study | Anatomical location | AI model used | Input feature and imaging direction | Output classes | Ground truth label assignment | Performance | | | Training set size | IV set size | EV type | EV set size |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Metric | IV | EV | | | | |
| Lindsey et al, 2018[24] | Wrist* | CNN | Radiograph; AP, lat | 2 | 1 or 2 orthopaedic surgeons | AUC | 0.967 | 0.975 | 31,490 | Set 1: 3,500 Set 2: 1,400 | T | 1,400 |
| Choi et al, 2020[25] | Elbow/distal humerus | CNN | Radiograph; AP, lat | 2 | 2 paediatric radiologists | AUC | 0.976 | T: 0.985; G: 0.992 | 1,012 | Not performed | T+G | T: 258 G: 95 |
| Bluthgen et al, 2020[6] | Wrist/distal radius | DLS† | Radiograph; AP, lat, combined | 2 | 2 radiology residents, reports, available CTs | AUC | Model 1: 0.93 Model 2: 0.96 | Model 1: 0.80 Model 2: 0.89 | 524 | 100 | G | 100 |
| Zhou et al, 2020[27] | Ribs | CNN | CT; Axial | 3 | 2 musculoskeletal radiologists, 2 senior radiologists, thoracic surgeon | F1 score | 1: 0.863 2: 0.856 | 3: 0.840 4: 0.811 5: 0.757 | 876 | 30 | G | 173 |

EV types: temporal (T), geographical (G).
*The model was also trained for the foot, elbow, shoulder, knee, spine, femur, ankle, humerus, pelvis, hip, and tibia.
†Deep learning system ViDi Suite v. 2.0 (ViDi Systems, Switzerland).
AP, anteroposterior; CNN, convolutional neural network; DLS, deep learning system; EV, external validation; G, geographical; IV, internal validation; Lat, lateral; T, temporal.

metric to report CNN performance,[26] where a value of 1.0 indicates perfect discriminatory performance, whereas 0.5 indicates a prediction equal to that of chance. One study used F1-score to evaluate model performance.[27] The F1-score (scored between 0 and 1) is a harmonic mean of precision (positive predictive value) and recall (sensitivity), where it requires both to be high for the F1-score to be high.

**EV dataset characteristics and CNN features.** All studies addressed AI models for fracture detection. In addition, one also used localization of fractures on images.[6] Zhou et al[27] addressed both fracture detection and classification. The CNNs detected fractures on a single anatomical location like the wrist,[6,24] elbow,[25] or ribs.[27] Input features of three studies[6,24,25] were conventional radiographs; one study used CT scans.[27] All four studies reported the use of IV, with sets ranging from 98 CT scans[27] to 3,500 radiographs.[24]

**Quality appraisal.** All studies reported disclosure. Study aim was clearly stated in all included studies, thereby reducing the possibility of outcome bias. All four studies clearly described the size, time, and location of collection of the EV dataset used, how the performance of the AI model was determined, and the ground truth (the reference standards used in AI). Out of the four studies included, three studies clearly stated the EV method used.[6,24,25] One study used external data to improve model robustness and generalizability, however this was done before internally validating the model performance on the test set.[27] The inclusion and exclusion criteria for input features were clearly described in three studies.[6,25,27] However, for one of the studies it was unclear which eligibility criteria were used for included radiographs.[24]

**Statistical analysis.** Performance metrics used in each study were described, as well as its values for fracture detection and classification tasks. The values were given for both IV and EV set whenever available. Descriptive statistics such as size of the EV, training, and IV set were reported.

## Results

To answer the primary research question, which CNNs for fracture recognition are externally valid and thus available for transfer from the developer to another hospital: four of 36 (11%) studies to date reported the use of EV (Table I).

To answer the second research question (which methods of EV for fracture recognition CNNs are currently used in the field of orthopaedic trauma), the following methodologies were identified (Table I).

CNNs deployed by Lindsey et al[24] were trained and internally validated on 31,490 and 3,500 respective radiographs between September 2000 and March 2016, and temporal EV performed with 1,400 radiographs from July to September 2016 from the same hospital. No geographical EV was applied.

Choi et al[25] conducted both temporal and geographical EV and used 258 patients for their temporal EV, which were collected between January and December 2018, and 95 patients collected at another hospital for their geographical EV. The CNN was trained and internally validated on 1,012 and 257 radiographs from their institution collected between January 2013 and December 2017.

The final two studies used geographical EV exclusively. Zhou et al[27] reported the use of a total of 75 patients for the geographical EV from three different respective hospitals with the original model trained and internally validated on 876 and 98 patients respectively, while Blüthgen et al[6] randomly selected 100 patients from the MURA dataset[28] with the index CNN trained and internally validated on 166 and 42 patients from the authors' local institution.

**Performance of CNN on EV compared to test set.** To answer the third study question on performance of CNNs for fracture recognition on test set versus EV, this systematic review yielded four studies.

Comparing the CNNs' performance on the IV versus EV set, the following values are found: AUC of 0.967 vs 0.975 for distal radius fracture recognition,[24] AUC of 0.976 versus 0.985 (temporal) and 0.992 (geographical) for paediatric supracondylar fracture recognition,[25] AUC of 0.93 to 0.96 versus 0.80 to 0.89 for recognition of distal radius fractures,[6] as well as an F1-score of 0.856 to 0.863 versus 0.757 to 0.840 for rib fracture recognition and classification on thoracic CT scans.[27]

Lindsey et al[24] reported slightly improved performance (AUC 0.967 vs 0.975) upon temporal EV. Choi et al[25] reported an increase of the AUC when geographically externally validated, a decrease of 10% accuracy detecting normal elbows, and an increase of 5% accuracy in detecting fractures, whereas the temporal EV set accuracy performed similarly to the IV set. Blüthgen et al[6] report a decrease in performance, for which the decrease in AP view of the distal radius is statistically significant ($p = 0.008$ to 0.021); however, calculating p-values in comparing AUCs has limited value. In Zhou et al,[27] a decreased F1-value is reported for the geographical EV sets.

## Discussion

There has been a significant increase in the use of CNNs in the field of orthopaedics the past few years.[1–11] Papers tout promising results, however careful evaluation of performance and clinical utility of CNNs is warranted. EV is one of the crucial steps to secure generalizability of CNNs developed to detect pathoanatomy,[19,29-31] prior to implementation into clinical practice. As many studies in our field now claim to have developed CNNs that perform at least on par with radiologists and orthopaedic surgeons,[6,8,9,24,25] we aimed to review if these CNNs for fracture recognition are indeed externally valid and thus ready for clinical application; and secondly which methods of EV were used. Just four out of 36 full-text reviewed studies report any form of EV, and three applied and tested their algorithm to a geographically different dataset. None of the current CNNs have been prospectively tested in clinical practice.

This study has several strengths and weaknesses: first, an appropriate risk of bias assessment tool currently does not exist for studies reporting the use of a CNN, therefore we modified the MINORS tool. Second, although a broad search strategy encompassing two large databases was used, potentially relevant publications or algorithms developed for commercial purposes might have been missed. Third, comparability of the diagnostic performance characteristics between studies is limited as studies developed CNNs recognizing different types of fractures, however this factor did not affect answering our research questions.

Although EV of CNNs for fracture recognition is scarce in orthopaedic trauma, authors of four included studies did stress the importance of EV.[6,24,25,27] They discussed the use of EV in evaluating CNNs, to discover generalizability and real-world performance. Indeed, EV evaluates the performance of CNNs in a different clinical context, a crucial step prior to implementation in clinical practice.[18] In other fields of medicine, this step is believed to be paramount before translation to clinical practice.[30] EV is considered the sequel to IV in evaluating a model, as it addresses transportability, rather than reproducibility.[32] The effect of factors, such as differences in demographics, operator-dependent radiological variances (for example, angle, rotation, and radiation dosage when performing a radiograph or CT), and brand and quality of radiograph machines on performance of the CNN need to be evaluated before one can transport any CNN to another institution.[18,33] This is highlighted by Raisuddin et al,[34] who advocate for in-depth analyses of artifical intelligence models, as reported in their paper where their model had great performance on radiographs from the general population, but significantly reduced performance on cases that were deemed hard for diagnosis by clinicians.

In general, true model performance as tested via EV is lower than the performance assessed with the dataset used for model development.[13,35,36] In this review for fracture detection and classification, studies conducting temporal EV reported similar or slightly improved performance compared to the IV set.[24,25] In contrast, studies using a geographically split dataset reported a decrease in performance with use of EV,[6,25,27] indicating the superiority of geographical over temporal validation. Blüthgen et al[6] explains that the decrease in performance observed indicates that the "variance" in images differed significantly between the IV and geographical EV sets, emphasizing the importance of geographical EV.

Not only variances in data, but also variation in labelling, can lead to varying performance: label noise can severely impact performance of CNNs,[37] and radiology reports are often based on only one observer.[38] In addition, these reports can have a variety of expertise and accuracy depending on who interprets the images.[24] Data labelling performed by a single expert carries significant risk of developing a biased CNN, catered to the opinion of one observer. Expert consensus can also be used, based on the assumption that agreement implies accuracy.[39] Nonetheless, limited availability of qualified experts to provide accurate image labels is a challenging problem when developing CNNs.[38] Although the input of experts—especially regarding evaluation of model predictions—is imperative to ensure clinical accuracy and relevance, reference standards such as follow-up imaging and surgical confirmation are considered the most

accurate method to train CNNs.[38] However, these are not always available, especially in simple fractures.

Although the importance of and need for EV is highlighted by many studies,[18,40-45] this review shows that EV of fracture recognition CNNs remains scarce. In addition, there is a lack of uniformity in the method of conducting and reporting of EV, such as defining ground truth. We therefore recommend readers to be cautious in interpreting performance when evaluation is limited to an internal or temporal validation set—as performance may vary when encountering data with different characteristics—and ideally geographical EV should be used to assess 'true' performance and generalizability. In addition, we advise the development and use of standardized methodology such as the recently published statements like the Clinical Artificial Intelligence Research (CAIR) checklist,[46] Standard Protocol Items: Recommendations for Interventional Trials – Artificial Intelligence (SPIRIT-AI),[47] and CONsolidated Standard for Reporting Trials – Artificial Intelligence (CONSORT-AI).[48] Several announced statements are still in development, like the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Machine Learning (TRIPOD-ML)[4] and the Standards for Reporting Diagnostic Accuracy –Artificial Intelligence (STARD-AI). Using these standardized statements will improve methodological rigor, quality of future models, and facilitate eventual implementation in clinical practice.

## Take home message

- We recommend readers to be cautious in interpreting performance when evaluation is limited to an internal or temporal validation set — as performance may vary when encountering data with different characteristics — and ideally geographical external validation should be used to assess 'true' performance and generalizability.

## Twitter

Follow University Medical Centre, University of Groningen @umcg
Follow University of Groningen @univgroningen
Follow Flinders University @Flinders
Follow Danderyd University Hospital @Danderydssjukh

## Supplementary material

Full search strategy as performed for the PubMed and Embase databases.

## References


1. **Topol EJ**. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44–56.
2. **Choy G**, **Khalilzadeh O**, **Michalski M**, **et al**. Current Applications and Future Impact of Machine Learning in Radiology. *Radiology*. 2018;288(2):318–328.
3. **Liu X**, **Rivera SC**, **Faes L**, **et al**. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med*. 2019;25(10):1467–1468.
4. **Collins GS**, **Moons KGM**. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393(10181):S0140-6736(19)30037-6):1577–1579:.
5. **Adams M**, **Chen W**, **Holcdorf D**, **McCusker MW**, **Howe PDL**, **Gaillard F**. Computer vs human: Deep learning versus perceptual training for the detection of neck of femur fractures. *J Med Imaging Radiat Oncol*. 2019;63(1):27–32.
6. **Blüthgen C**, **Becker AS**, **Vittoria de Martini I**, **Meier A**, **Martini K**, **Frauenfelder T**. Detection and localization of distal radius fractures: Deep learning system versus radiologists. *Eur J Radiol*. 2020;126(8):108925.
7. **Urakawa T**, **Tanaka Y**, **Goto S**, **Matsuzawa H**, **Watanabe K**, **Endo N**. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. *Skeletal Radiol*. 2019;48(2):239–244.
8. **Chung SW**, **Han SS**, **Lee JW**, **et al**. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop*. 2018;89(4):468–473.
9. **Tomita N**, **Cheung YY**, **Hassanpour S**. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Comput Biol Med*. 2018;98:8–15.
10. **Yamada Y**, **Maki S**, **Kishida S**, **et al**. Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: ensemble decision-making with antero-posterior and lateral radiographs. *Acta Orthop*. 2020;91(6):699–704.
11. **Kalmet PHS**, **Sanduleanu S**, **Primakov S**, **et al**. Deep l*earning in fracture detection: a narrative review. Acta Orthop*. 2020;91(2):215–220.
12. **Bongers MER**, **Thio QCBS**, **Karhade AV**, **et al**. Does the SORG algorithm predict 5-year survival in patients with chondrosarcoma? An external validation. *Clin Orthop Relat Res*. 2019;477(10):2296–2303.
13. **Siontis GCM**, **Tzoulaki I**, **Castaldi PJ**, **Ioannidis JPA**. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*. 2015;68(1):25–34.
14. **Liu K-L**, **Wu T**, **Chen P-T**, **et al**. Deep learning to distinguish pancreatic cancer tissue from non-cancerous pancreatic tissue: a retrospective study with cross-racial external validation. *Lancet Digit Health*. 2020;2(6):e313e303):e303–e313:.
15. **Gertych A**, **Swiderska-Chadaj Z**, **Ma Z**, **et al**. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Sci Rep*. 2019;9(1):1483.
16. **Steyerberg EW**. *Clinical Prediction Models*. Springer.
17. **König IR**, **Malley JD**, **Weimar C**, **Diener H-C**, **Ziegler A**, **German Stroke Study Collaboration**. Practical experiences on the necessity of external validation. *Stat Med*. 2007;26(30):5499–5511.
18. **Oosterhoff JHF**, **Doornberg JN**, **Machine Learning Consortium**. Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of Gartner's hype cycle. *EFORT Open Rev*. 2020;5(10):593–603.
19. **Ho SY**, **Phua K**, **Wong L**, **Bin Goh WW**. Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. *Patterns (N Y)*. 2020;1(8):100129.
20. **Moher D**, **Liberati A**, **Tetzlaff J**, **Altman DG**, **Group P**. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097.
21. **Whiting PF**, **Rutjes AWS**, **Westwood ME**, **et al**. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529–536.
22. **Pellegrini E**, **Ballerini L**, **Hernandez MDCV**, **et al**. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review. *Alzheimers Dement (Amst)*. 2018;10:519–535.
23. **Slim K**, **Nini E**, **Forestier D**, **Kwiatkowski F**, **Panis Y**, **Chipponi J**. Methodological index for non-randomized studies (minors): development and validation of a new instrument. *ANZ J Surg*. 2003;73(9):712–716.
24. **Lindsey R**, **Daluiski A**, **Chopra S**, **et al**. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A*. 2018;115(45):11591–11596.
25. **Choi JW**, **Cho YJ**, **Lee S**, **et al**. Using a Dual-Input Convolutional Neural Network for Automated Detection of Pediatric Supracondylar Fracture on Conventional Radiography. *Invest Radiol*. 2020;55(2):101–110.
26. **Langerhuizen DWG**, **Janssen SJ**, **Mallee WH**, **et al**. What Are the Applications and Limitations of Artificial Intelligence for Fracture Detection and Classification in Orthopaedic Trauma Imaging? A Systematic Review. *Clin Orthop Relat Res*. 2019;477(11):2482–2491.
27. **Zhou QQ**, **Wang J**, **Tang W**, **et al**. Automatic Detection and Classification of Rib Fractures on Thoracic CT Using Convolutional Neural Network: Accuracy and Feasibility. *Korean J Radiol*. 2020;21(7):869–879.
28. **Rajpurkar P**, **Irvin J**, **Bagul A**, **et al**. MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. Cornell University. 2017. https://arxiv.org/abs/1712.06957
29. **Baldwin DR**, **Gustafson J**, **Pickup L**, **et al**. External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. *Thorax*. 2020;75(4):306–312.

30. **Milea D**, **Najjar RP**, **Zhubo J**, **et al**. Artificial Intelligence to Detect Papilledema from Ocular Fundus Photographs. *N Engl J Med*. 2020;382(18):1687–1695.
31. **Nam JG**, **Park S**, **Hwang EJ**, **et al**. Development and Validation of Deep Learning-based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. *Radiology*. 2019;290(1):218–228.
32. **Justice AC**, **Covinsky KE**, **Berlin JA**. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130(6):515–524.
33. **Park SH**, **Han K**. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology*. 2018;286(3):800–809.
34. **Raisuddin AM**, **Vaattovaara E**, **Nevalainen M**, **et al**. Critical evaluation of deep neural networks for wrist fracture detection. *Sci Rep*. 2021;11(1):6006.
35. **Moons KGM**, **Kengne AP**, **Grobbee DE**, **et al**. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98(9):691–698.
36. **Zendel O**, **Murschitz M**, **Humenberger M**, **Herzner W**. How Good Is My Test Data? Introducing Safety Analysis for Computer Vision. *Int J Comput Vis*. 2017;125(1–3):95–109.
37. **Karimi D**, **Dou H**, **Warfield SK**, **Gholipour A**. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Med Image Anal*. 2020;65:101759.
38. **England JR**, **Cheng PM**. Artificial Intelligence for Medical Image Analysis: A Guide for Authors and Reviewers. *AJR Am J Roentgenol*. 2019;212(3):513–519.
39. **Kundel HL**, **Polansky M**. Measurement of observer agreement. *Radiology*. 2003;228(2):303–308.
40. **Weikert T**, **Noordtzij LA**, **Bremerich J**, **et al**. Assessment of a Deep Learning Algorithm for the Detection of Rib Fractures on Whole-Body Trauma Computed Tomography. *Korean J Radiol*. 2020;21(7):891–899.
41. **Thian YL**, **Li Y**, **Jagmohan P**, **Sia D**, **Chan VEY**, **Tan RT**. Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs. *Radiol Artif Intell*. 2019;1(1):e180001.
42. **Lee C**, **Jang J**, **Lee S**, **Kim YS**, **Jo HJ**, **Kim Y**. Classification of femur fracture in pelvic X-ray images using meta-learned deep neural network. *Sci Rep*. 2020;10(1):13694.
43. **Al-Helo S**, **Alomari RS**, **Ghosh S**, **et al**. Compression fracture diagnosis in lumbar: a clinical CAD system. *Int J Comput Assist Radiol Surg*. 2013;8(3):461–469.
44. **Badgeley MA**, **Zech JR**, **Oakden-Rayner L**, **et al**. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med*. 2019;2(1):31.
45. **Derkatch S**, **Kirby C**, **Kimelman D**, **Jozani MJ**, **Davidson JM**, **Leslie WD**. Identification of Vertebral Fractures by Convolutional Neural Networks to Predict Nonvertebral and Hip Fractures: A Registry-based Cohort Study of Dual X-ray Absorptiometry. *Radiology*. 2019;293(2):405–411.
46. **Olczak J**, **Pavlopoulos J**, **Prijs J**, **et al**. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. *Acta Orthop*. 14, 2021.
47. **Cruz Rivera S**, **Liu X**, **Chan A-W**, **et al**. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351–1363.
48. **Cruz Rivera S**, **Liu X**, **Chan A-W**, **Denniston AK**, **Calvert MJ**, **et al**, **SPIRIT-AI and CONSORT-AI Working Group**. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health*. 2020;2(10):e548e537):e549–e560:.



**Author information:**
- L. Oliveira e Carmo, BSc, Medical Student
- P. C. Jutte, MD, PhD, Professor of Orthopaedics
  Department of Orthopaedic Surgery, University Medical Centre, University of Groningen, Groningen, Groningen, Netherlands.
- A. van den Merkhof, BSc, Medical Student
- R. L. Jaarsma, MD, PhD, FRACS, Professor of Orthopaedics and Trauma Surgery
  Department of Orthopaedic Surgery, Flinders Medical Centre, Bedford Park, Adelaide, South Australia, Australia; Flinders University, Bedford Park, Adelaide, South Australia, Australia.
- J. Olczak, MD, Orthopaedic Resident
- M. Gordon, MD, PhD, Consultant Orthopaedic Surgeon
  Institute of Clinical Sciences, Danderyd University Hospital, Karolinska Institute, Stockholm, Sweden.
- F. F. A. IJpma, MD, PhD, Consultant Trauma Surgeon, Department of Trauma Surgery, University Medical Centre Groningen, University of Groningen, Groningen, Groningen, Netherlands.
- J. N. Doornberg, MD, PhD, Professor of Orthopaedic Trauma
- J. Prijs, BSc, PhD Candidate
  Department of Orthopaedic Surgery, University Medical Centre, University of Groningen, Groningen, Groningen, Netherlands; Department of Orthopaedic Surgery, Flinders Medical Centre, Bedford Park, Adelaide, South Australia, Australia; Flinders University, Bedford Park, Adelaide, South Australia, Australia; Department of Trauma Surgery, University Medical Centre Groningen, University of Groningen, Groningen, Groningen, Netherlands.



**Author contributions:**
- L. Oliveira e Carmo: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing.
- A. van den Merkhof: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – review & editing.
- J. Olczak: Writing – review & editing.
- M. Gordon: Writing – review & editing.
- P. C. Jutte: Writing – review & editing.
- R. L. Jaarsma: Writing – review & editing.
- F. F. A. IJpma: Writing – review & editing.
- J. N. Doornberg: Conceptualization, Formal analysis, Supervision, Writing – original draft, Writing – review & editing.
- J. Prijs: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

**Funding statement:**
- No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article. Open access was funded by University Medical Centre Groningen.

**Acknowledgements:**
- No funding has been received for this study. We would like to acknowledge Karin Sijtsma for her help in building the final search strategy.
  Paul Algra; Michel van den Bekerom; Mohit Bhandari; Michiel Bongers; Charles Court-Brown; Anne-Eva Bulstra; Geert Buijze; Sofia Bzovsky; Joost Colaris; Neil Chen; Job Doornberg; Andrew Duckworth; J. Carel Goslings; Max Gordon; Benjamin Gravesteijn; Olivier Groot; Gordon Guyatt; Laurent Hendrickx; Beat Hintermann; Dirk-Jan Hofstee; Frank IJpma; Ruurd Jaarsma; Stein Janssen; Kyle Jeray; Paul Jutte; Aditya Karhade; Lucien Keijser; Gino Kerkhoffs; David Langerhuizen; Jonathan Lans; Wouter Mallee; Matthew Moran; Margaret McQueen; Marjolein Mulders; Rob Nelissen; Miryam Obdeijn; Tarandeep Oberai; Jakub Olczak; Jacobien H.F. Oosterhoff; Brad Petrisor; Rudolf Poolman; Jasper Prijs; David Ring; Paul Tornetta III; David Sanders; Joseph Schwab; Emil H. Schemitsch; Niels Schep; Inger Schipper; Bram Schoolmeesters; Joseph Schwab; Marc Swiontkowski; Sheila Sprague; Ewout Steyerberg; Vincent Stirler; Paul Tornetta; Stephen D. Walter; Monique Walenkamp; Mathieu Wijffels.