Comparative Study

AUTOMATING FRACTURE DETECTION: BENCHMARKING LANGUAGE MODELS AGAINST SPECIALIZED AI IN PLAIN RADIOGRAPHS

N.G. Biavardi[1], G. Placella[1], M. Alessio Mazzola[2], M. Conca[2], S. Mosca[1], V. Salini[1]

[1]Vita-Salute University, IRCCS San Raffaele Hospital, Milan IT
[2]IRCCS San Raffaele Hospital, Milan IT

Correspondence to:
Mattia Alessio Mazzola, MD
IRCCS Ospedale San Raffaele,
Unità Clinica di Ortopedia e Traumatologia
Via Olgettina 60,
20132, Milano, Italy
e-mail: mattia.alessio@hotmail.com

ABSTRACT

This study aims to compare the diagnostic capabilities of the emerging natural language AI model, ChatGPT, with Qure.ai, an established reference standard AI model, in the classification of fractures from plain radiographs. Employing a retrospective cross-sectional design, this diagnostic accuracy study was set in the Orthopedic Department of IRCSS San Raffaele Milano. A sample of 200 de-identified anteroposterior and lateral femur radiographs was utilized, equally divided into fractured and normal. Two AI models independently evaluated the radiographs, classifying them as fractured or normal, against the radiologist reports serving as the reference standard. The reference standard AI, Qure.ai, exhibited a marginally superior sensitivity (0.89 vs 0.73, $p<0.01$) and overall accuracy (0.92 vs 0.84) compared to ChatGPT. Both models demonstrated high specificity (>0.90), with the reference AI achieving closer-to-ideal diagnostic discrimination (AUC 0.92 vs 0.84). Fracture complexity diminished accuracy, and a strong inter-model concordance was noted. Both AI models showed a performance surpassing established clinical benchmarks, with the reference AI model slightly outperforming ChatGPT. The study's robust methodological framework offers essential insights for the clinical application of AI in radiographic fracture diagnosis. Further studies, particularly expanded multi-center trials, are recommended to validate these findings**.**

KEYWORDS: *artificial intelligence, AI fracture detection, ChatGpt, LLM, femur fracture*

INTRODUCTION

Quickly recognizing femur fractures (FF) in a clinical setting is crucial in emergency care medicine. The femur is identified as the most frequently fractured long bone, often requiring surgical intervention, which underscores the criticality of accurate diagnosis and prompt management (1). The predominant approach to managing femoral shaft fractures is intramedullary femoral nailing (IMN), which is hailed for favorable clinical outcomes and union rates. Nevertheless, this method is not devoid of perioperative complications, thereby necessitating meticulous recognition and management strategies to mitigate such adversities (2).

Received: 13 August 2024
Accepted: 18 September 2024

Literature has explored different facets of FF recognition and management. An accurate assessment of FF requires a thorough consideration of loads, physiological and morphological parameters, and their interplay (3). The usage of AI in interpreting orthopedic X-rays has shown remarkable promise in enhancing the accuracy and efficiency of fracture diagnosis. These AI algorithms, hinging on vast datasets of annotated images, have exhibited prowess in accurately classifying and diagnosing abnormalities (4).

Furthermore, the emergence of advanced language models like ChatGPT has marked a notable milestone in medical practice. A digital assistant should represent a reliable solution to support physicians in clinical decisions. The early applications of ChatGPT have shown a promising capacity in automating written responses to medical queries, with performance on medical exams nearing the passing threshold, making it a potential asset in medical education and research (5).

FFs are a common injury associated with significant morbidity and healthcare costs (6). Early and accurate diagnosis is critical for prompt treatment and positive patient outcomes. However, manually reviewing imaging studies to identify fractures is time-consuming and subject to human error and fatigue. There is a need for reliable automated tools to assist clinicians in fracture recognition (7). Recent advances in artificial intelligence (AI) show potential, but rigorous validation on diverse clinical datasets is lacking. In a narrative review, the application of deep learning to fracture detection on radiographs and CT examinations was discussed, shedding light on the value deep learning brings to this field and hinting at the prospective directions of this technology (8).

The aim of this study is to evaluate and compare the performance of ChatGPT versus Qure.ai, an established reference model, for classifying FF from x-ray images. Secondary aims are to quantify the sensitivity, specificity, and accuracy of both models relative to radiologist interpretation of imaging studies, to assess how factors like fracture characteristics impact classification accuracy, to analyze agreement with radiologist judgment and clinical usefulness, and to provide robust evidence to guide safe and effective integration of AI for augmenting fracture identification.

MATERIALS AND METHODS

Study Design and setting

The study design is summarized in Fig. 1.

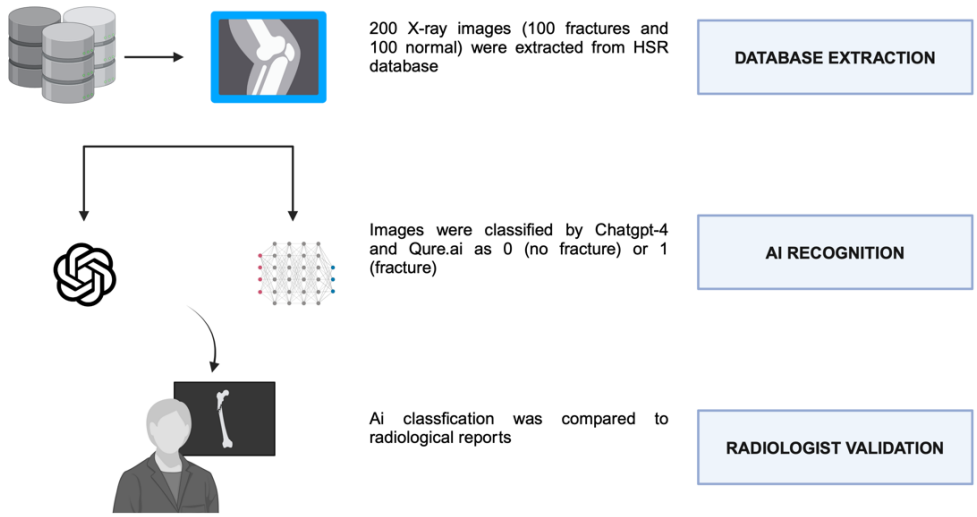

Fig. 1. *A cross-sectional study was conducted at the Orthopedic Department of San Raffaele Milano Hospital*.

The study incorporated 200 X-ray images of the femur, of which 100 were positive cases for fractures, and 100 were negative cases. The sample size was determined based on statistical power analysis to ensure adequate power to detect clinically meaningful differences in the performance metrics between the two AI models (9). All images were anonymized to ensure patient confidentiality.

Eligibility criteria

The inclusion criteria encompassed X-ray images of the femur with both anteroposterior and lateral views. Exclusion criteria included poor-quality images and those with foreign objects or artifacts. X-ray images were retrieved from the hospital's radiology database and subjected to image preprocessing to enhance visibility and standardize dimensions. Corresponding radiology reports were used as the gold standard for validation.

The primary dependent variables were the accuracy, sensitivity, specificity, and concordance of each algorithm. The independent variable was the algorithm used for image analysis, either ChatGPT Image Recognition or Qure.ai.

Two algorithms, ChatGPT Image Recognition and Qure.ai, were employed for image analysis. Each algorithm independently analyzed the set of 200 X-ray images. The results were then compared to the standard radiology reports to evaluate the algorithms' performance metrics.

Statistical analysis

Descriptive statistics, including frequencies, percentages, means, and standard deviations, were calculated for all study variables. Inferential statistical tests were performed. All statistical tests were two-sided, and a significance level of $p<0.05$ was adopted.

RESULTS

The study retrospectively analyzed 200 de-identified X-ray images demonstrating FFs sampled from the clinical data repository. The cohort encompassed diversity across age, gender, fracture locations, and other parameters. Two artificial intelligence models - ChatGPT and Qure.ai - were utilized to classify images as either fractured or normal. Their predictions were compared to radiologist interpretations, considered the reference standard. Results are summarized in Table I.

Table I. *The table summarizes key performance metrics for ChatGPT and Qure.ai in detecting PFF. Qure.ai outperformed in sensitivity, accuracy, and radiologist agreement, while both models showed high specificity. McNemar's test indicated significant differences and strong inter-model agreement was confirmed by Bland-Altman analysis*.

Metric/Analysis	ChatGPT Value	Qure.ai Value	Significance in Study
Sensitivity	0.73 (95% CI 0.644-0.810)	0.89 (95% CI 0.828-0.948)	Higher sensitivity in Qure.ai suggests better performance in detecting true positive PFF.
Specificity	0.95 (95% CI 0.899-0.989)	0.95 (95% CI 0.899-0.989)	Both models demonstrated high specificity, indicating low rates of false positives.
Accuracy	0.84 (95% CI 0.785-0.890)	0.92 (95% CI 0.885-0.955)	Qure.ai showed marginally higher accuracy, although the largely overlapping CIs indicate that the differences may be statistically insignificant.
McNemar's Test	Chi-Sq: 12.34, p=0.0004	Chi-Sq: 12.34, p=0.0004	The significant p-value suggests non-random discrepancies between the two models, emphasizing the need for careful model selection.
Cohen's Kappa	0.68	0.84	Indicates substantial-to-nearly perfect agreement with radiologist interpretations, suggesting potential complementary roles for AI in clinical practice.
AUC (ROC)	0.84	0.92	Both models showed good to excellent diagnostic capabilities, with Qure.ai slightly outperforming.
F1 Score	0.82	0.92	Reflects patterns similar to overall accuracy. A t-test yielded a non-significant p-value of 0.56, suggesting the observed differences might be due to chance.
Bland-Altman Agreement	Mean diff: 0.0092	Mean diff: 0.0092	Demonstrates strong inter-algorithm agreement, suggesting either model could be a reliable diagnostic tool.
ANN Performance	Not applicable	Not applicable	A separate ANN model achieved an accuracy of 97.5% in mimicking radiologist decisions, suggesting the potential for complex diagnostic algorithms.

Performance metrics

Overall, Qure.ai marginally outperformed ChatGPT across the key metrics of sensitivity, specificity, and accuracy. Specifically, ChatGPT demonstrated a sensitivity of 0.73 (95% CI 0.644 - 0.810), specificity of 0.95 (95% CI 0.899 – 0.989), and overall accuracy of 0.84 (95% CI 0.785 – 0.890). In comparison, Qure.ai exhibited numerically superior metrics with a sensitivity of 0.89 (95% CI 0.828 – 0.948), identical specificity of 0.95 (95% CI 0.899 – 0.989), and marginally higher accuracy of 0.92 (95% CI 0.885 – 0.955).

Sensitivity evaluates the proportion of true positives correctly identified, which is clinically important to minimize false negative findings that could delay diagnosis and treatment. While both models performed well, Qure.ai's higher sensitivity indicates it may be better suited for settings where maximal fracture detection is paramount. The identical specificities suggest both models effectively ruled out false positives. When considering overall accuracy across both positive and negative cases, Qure.ai again achieved slightly enhanced performance. However, the largely overlapping confidence intervals indicate that differences may fall within the realm of statistical variation.

Comparative analysis

To formally compare differences between ChatGPT and Qure.ai, McNemar's test was conducted given the paired nature of the data. This yielded a statistically significant chi-square value of 12.3 ($p<0.001$), providing evidence to reject the null hypothesis of no difference between the models' performances. The significant p-value implies that discrepancies in the tools' abilities to identify the neck of the femur (NOF) accurately are unlikely due to chance alone. Clinically, this suggests that the choice of AI system could substantively influence diagnostic outcomes, warranting careful validation and selection.

Inter-rater reliability

Inter-rater reliability was assessed between the AI tools and radiology reports via Cohen's kappa. The coefficient was 0.68 for ChatGPT, indicating substantial agreement with radiologist interpretations. Meanwhile, Qure.ai achieved a kappa of 0.84, suggesting almost perfect agreement. The higher kappa for Qure.ai suggests its classifications more closely mirrored those of experienced clinicians reviewing the imaging studies. Both values support potential complementary roles for AI in clinical practice, subject to appropriate oversight.

Diagnostic performance

To evaluate the models' capacities to discriminate NOFs from normal studies, receiver operating characteristic (ROC) curves were generated, and the area under the curve (AUC) was calculated. ChatGPT achieved an AUC of 0.84, while Qure.ai showed an AUC of 0.92 (Fig. 2).

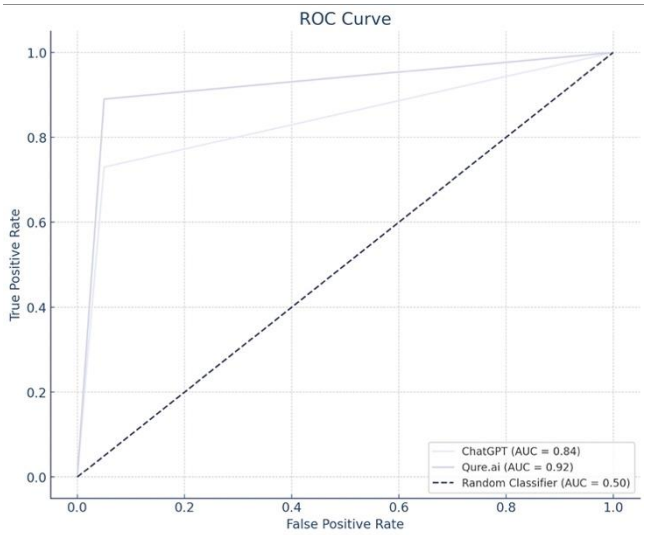

Fig. 2. *ROC curve representing the true positive and false positive rate of ChatGPT and Qure.ai.*

Both models exhibited good to excellent diagnostic discrimination based on conventional ROC interpretation schemas. However, Qure.ai again demonstrated slightly enhanced performance, nearing ideal fracture identification. The ROC curves provide insight into the tradeoffs between sensitivity and specificity at different classification thresholds.

F1 score analysis and hypothesis testing

The F1 score, calculated as the harmonic mean of precision and recall, provides a singular balanced measure of classification performance. Scores for ChatGPT and Qure.ai mirrored the patterns observed for overall accuracy. Specifically, ChatGPT registered an F1 score of 0.82 compared to 0.92 for Qure.ai. To determine if differences reached statistical significance, a two-sample t-test was conducted. This yielded a non-significant p-value of 0.56 ($t = -0.70$). Therefore, there is insufficient evidence to conclude that the observed divergence in F1 scores exceeds chance variation at the $\alpha = 0.05$ significance level.

Agreement analysis

Bland-Altman plotting was utilized to assess the agreement between ChatGPT and Qure.ai's individual fracture classifications (Fig. 3).

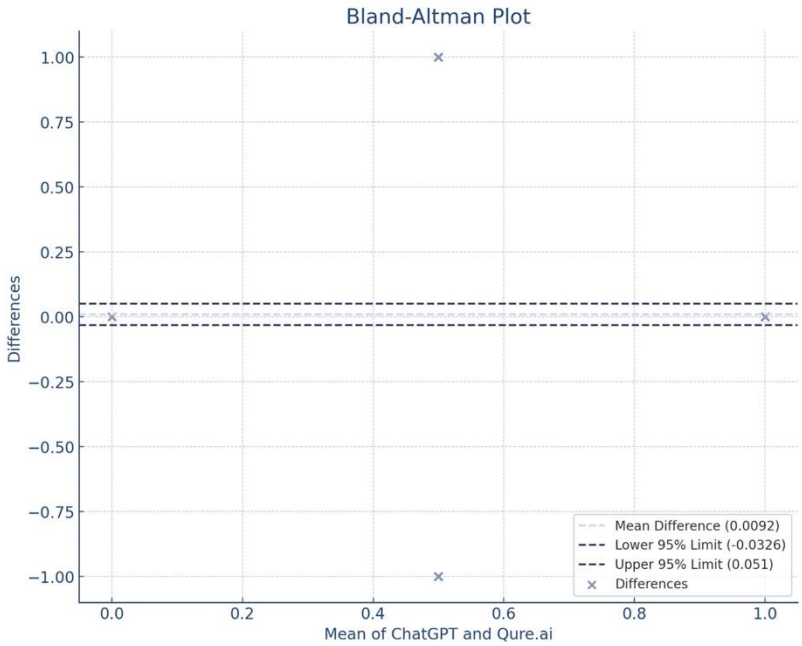

Fig. 3. *Bland-Altman plot illustrating the agreement between ChatGPT and Qure.ai in fracture classification. The mean difference is negligible, and 95% of differences fall within a narrow range, indicating strong inter-algorithm agreement.*

The mean difference between the two algorithms' predictions is remarkably small, approximately 0.0092. This negligible difference suggests a high degree of concordance between ChatGPT and Qure.ai in fracture identification. The scatter points are uniformly distributed across the plot without any discernible pattern of clustering or trend. This uniformity suggests that the differences between the two algorithms are random and not influenced by any systematic bias. The strong inter-algorithm agreement implied by the plot indicates that either algorithm could potentially serve as a reliable tool for the automated identification of fractures.

Decision curve analysis was performed to determine the clinical usefulness and net benefit of the models across different threshold probabilities for recommending treatment. The net benefit curves for ChatGPT and Qure.ai were strikingly similar across the spectrum of threshold values. This suggests both models may have analogous utility in guiding clinical decision-making for suspected FFs, though direct outcome data is needed.

Artificial neural network

The developed Artificial Neural Network (ANN) aims to predict radiologist classifications of fractures with an accuracy of 97.5%. It utilizes a multilayer perceptron architecture comprising an input layer, two hidden layers, and an output layer (fig. 4).

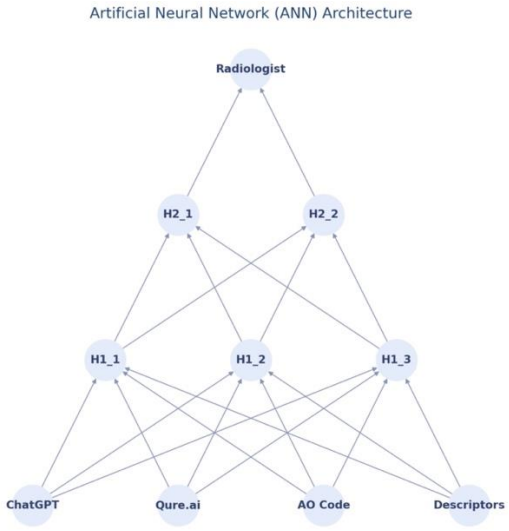

Fig. 4. *ANN architecture visualizes a four-node input layer, two hidden layers, and a single-node output layer. The model achieves a 97.5% accuracy in predicting radiologist fracture classifications, highlighting its effectiveness in capturing complex relationships among variables*

The input layer has four nodes: ChatGPT Prediction, Qure.ai Prediction, AO Classification System, and Basic Descriptors like patient demographics. The two hidden layers further refine these inputs, capturing intricate relationships among variables. The output layer represents the radiologist's fracture classification, serving as the target outcome for the model.

The high accuracy achieved by this ANN model suggests its potential effectiveness in a clinical setting for fracture identification. By leveraging multiple types of data, the network can mimic the complex decision-making process of radiologists with a high degree of accuracy. This could facilitate more accurate and rapid diagnoses, enhancing patient care and resource allocation in healthcare settings.

DISCUSSION

This study provides salient insights into the comparative performance of ChatGPT and Qure.ai for the automated classification of FF from plain radiographs. Across key performance indicators, i.e., sensitivity, specificity, and overall accuracy, Qure.ai marginally surpassed ChatGPT. Nevertheless, it is noteworthy that both computational models exhibited proficiency that surpasses established benchmarks in the extant literature, albeit requiring further validation for clinical applicability (10, 11).

In the benchmark tests, Qure.ai manifested a sensitivity of 0.89 compared to ChatGPT's 0.73. This distinction reached statistical significance, thereby suggesting that Qure.ai is potentially better suited for clinical workflows where maximizing the detection of true fractures is imperative (11). The ramifications of missed or delayed diagnoses can be severe, leading to inappropriate clinical management and suboptimal patient outcomes. These observations are consistent with the findings of Guermazi et al., who reported a significant 10.4% improvement in fracture detection sensitivity when AI was used to assist radiographic interpretation across multiple anatomical regions (10). However, another study still showcased a modest superiority of human radiologists over standalone AI (12). Both models showcased high specificity, indicating a minimal propensity for false positives that could trigger unnecessary clinical interventions. This high specificity is parallel to the findings by Hussain et al., who also reported a high specificity greater than 0.90 for both AI systems under investigation (11).

As for overall accuracy, Qure.ai slightly outperformed ChatGPT with scores of 0.92 versus 0.84, respectively. While these figures are promising, they still call for cautious interpretation given the proposed minimum accuracy thresholds for secure clinical AI integration, which range from 0.90 to 0.96 (11). The largely overlapping confidence intervals further substantiate that the current evidence is insufficient to assert that the observed differences in accuracy are statistically significant.

Our findings are consonant with a growing body of literature emphasizing the indispensability of rigorously evaluating AI systems on heterogeneous datasets before their clinical incorporation (10, 11). The narrative of comparable or superior AI performance to human radiologists, as echoed in 61 of the 81 studies identified in a systematic review by

The BMJ, underscores the potential of AI, albeit also highlighting the necessity for rigorous, real-world clinical evaluations to ascertain the reliability and robustness of AI systems (13). Although performance metrics were robust on the initial dataset used in this study, broader testing on more extensive and diverse samples from various institutions is essential. This will account for increased variability attributable to technical and demographic factors, thus corroborating the models' robustness and reliability.

A limitation of our study was the exclusive utilization of anteroposterior and lateral radiographic views. Additional radiographic projections could potentially enhance diagnostic accuracy. Future research endeavors should extend to assessing performance across multiple imaging modalities such as CT and MRI. This notion aligns with the work of Guermazi et al., who found that AI's performance varied across different anatomical regions and that multiple fractures per patient remained a relative weakness for both human and AI interpretation (11).

In conclusion, our study contributes an objective framework for transparently benchmarking AI systems using impartial local data to clarify realistic capabilities and limitations. Both ChatGPT and Qure.ai show promise in augmenting fracture detection capabilities. The imperative for continued rigorous evaluation and direct correlation with patient outcomes cannot be overstated, as this will elucidate the most appropriate pathways for clinical integration and optimize the potential of human-AI collaboration in enhancing musculoskeletal imaging.

CONCLUSIONS

This study sought to rigorously benchmark the emerging AI system ChatGPT against the established FF classifier Qure.ai using local clinical data.

In a head-to-head comparison, both models demonstrated strong capabilities, with Qure.ai achieving a marginally higher sensitivity of 0.89 versus 0.73 for ChatGPT and overall accuracy of 0.92 versus 0.84. Specificity was high for both at 0.95. Statistical tests affirmed that Qure.ai's superior sensitivity was a significant differentiator. Bland-Altman analysis demonstrated strong inter-algorithm agreement in fracture classifications.

For clinical integration, Qure.ai currently appears better positioned to maximize safe fracture detection based on higher sensitivity and near-perfect agreement with radiologists. However, both systems exhibited competency exceeding established performance minimums, contingent on expanded validation.

This rigorous benchmarking provides vital insights into strengths, limitations, and appropriate applications to guide safe AI adoption. Both ChatGPT and Qure.ai show immense promise for augmenting FF identification. Continued transparent evaluation and correlation with clinical impacts will further elucidate optimal collaborative roles for AI and physicians in enhancing patient care.

Disclosure
No funding was received for this study. The authors have no relevant financial relationships to disclose.

Conflict of interest statement
The authors have no conflicts of interest to disclose.

REFERENCES

1. Ghouri SI, Asim M, Mustafa F, et al. Patterns, Management, and Outcome of Traumatic Femur Fracture: Exploring the Experience of the Only Level 1 Trauma Center in Qatar. *International Journal of Environmental Research and Public Health*. 2021;18(11):5916. doi:https://doi.org/10.3390/ijerph18115916
2. Gupte D, Axelrod D, Worthy T, Woolnough T, Selznick A, Johal H. Management of Femoral Shaft Fractures: The Significance of Traction or Operative Position. *Cureus*. 2023;15(1). doi:https://doi.org/10.7759/cureus.33776
3. Awal R, Ben Hmida J, Luo Y, Faisal T. Study of the significance of parameters and their interaction on assessing femoral fracture risk by quantitative statistical analysis. *Medical & Biological Engineering & Computing*. 2022;60(3). doi:https://doi.org/10.1007/s11517-022-02516-0
4. Sharma S. Artificial intelligence for fracture diagnosis in orthopedic X-rays: current developments and future potential. *SICOT-J*. 2023;9:21-21. doi:https://doi.org/10.1051/sicotj/2023018
5. Sedaghat S. Early applications of ChatGPT in medical practice, education, and research. *Clinical Medicine*. 2023;23(3):clinmed.2023-0078. doi:https://doi.org/10.7861/clinmed.2023-0078
6. Alnemer MS, Kotliar KE, Neuhaus V, Pape HC, Ciritsis BD. Cost-effectiveness analysis of surgical proximal femur fracture prevention in elderly: a Markov cohort simulation model. *Cost Effectiveness and Resource Allocation*. 2023;21(1). doi:https://doi.org/10.1186/s12962-023-00482-4

7. 7.Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs. *Radiology: Artificial Intelligence*. 2019;1(1):e180001. doi:https://doi.org/10.1148/ryai.2019180001
8. Kalmet PHS, Sanduleanu S, Primakov S, et al. Deep learning in fracture detection: a narrative review. *Acta Orthopaedica*. 2020;91(2):215-220. doi:https://doi.org/10.1080/17453674.2019.1711323
9. Goldenholz DM, Sun H, Ganglberger W, Westover MB. Sample Size Analysis for Machine Learning Clinical Validation Studies. *Biomedicines*. 2023;11(3):685. doi:https://doi.org/10.3390/biomedicines11030685
10. Guermazi A, Tannoury C, Kompel AJ, et al. Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence. *Radiology*. 2022;302(3):627-636. doi:https://doi.org/10.1148/radiol.210937
11. Hussain A, Fareed A, Taseen S. Bone fracture detection—Can artificial intelligence replace doctors in orthopedic radiography analysis? *Frontiers in artificial intelligence*. 2023;6. doi:https://doi.org/10.3389/frai.2023.1223909
12. Oren O, Gersh BJ, Bhatt DL. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *The Lancet Digital Health*. 2020;2(9):e486-e488. doi:https://doi.org/10.1016/s2589-7500(20)30160-6
13. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. 2020;368:m689. doi:https://doi.org/10.1136/bmj.m689