Method for Retrospective Analysis of Clinical Outcomes After Removal of Radiologist Assessment of AI-Positive Skeletal X-rays

Kristian Malm-Nicolaisen
Norwegian Centre for E-health Research
Tromsø, Norway
Email: kristian.nicolaisen@ehealthresearch.no

Rune Pedersen
Norwegian Centre for E-health Research
Tromsø, Norway
Email: rune.pedersen@ehealthresearch.no

Asbjørn Johansen Fagerlund
Norwegian Centre for E-health Research
Tromsø, Norway
Email: asbjorn.johansen.fagerlund@ehealthresearch.no

Abstract— **This retrospective in-progress study presents a method for evaluating the clinical impact of removing routine radiologist assessments of AI-flagged ("AI-positive") skeletal X-rays at Vestre Viken Health Trust, which recently implemented the BoneView AI system. Over a five-month period, orthopedic surgeons or radiologists prospectively identified False Positives (FPs), and comprehensive chart reviews assessed whether omitting radiologist input compromised patient safety or led to unnecessary interventions. In the current phase, an estimated 20-40 FP cases will be analyzed for treatment outcomes, resource utilization, and potential misdiagnoses. The findings will inform evidence-based strategies for integrating AI into radiological workflows, guiding institutions in balancing efficiency with the need for robust diagnostic oversight.**

Keywords-artificial intelligence; diagnostic imaging; clinical workflow; patient outcomes.

I. INTRODUCTION

Artificial Intelligence (AI)-driven technologies hold significant promise in addressing critical challenges within healthcare, including workforce shortages and the increasing demands of an aging population with complex medical needs. Despite the proliferation of commercial AI solutions globally, the integration and deployment of these technologies in clinical environments remain limited [1]. While commercial AI algorithms undergo validation for clinical performance prior to market entry, early adopter institutions must determine how to integrate these solutions into clinical workflows and evaluate the implications of workflow modifications.

A notable example is Vestre Viken Health Trust in Norway, part of the public healthcare sector, which implemented BoneView (by Gleamer) across all its hospitals in 2023. BoneView is an AI-powered tool designed to assist radiologists in the detection and assessment of fractures in X-ray examinations. By embedding BoneView into routine radiological practice, Vestre Viken aims to enhance diagnostic accuracy, reduce radiologist workload, and improve patient care through faster and more reliable fracture assessments [2]. Throughout the implementation phase, Vestre Viken has continued its legacy workflow where radiologists assess clinical images that are labeled as diagnostically positive by the AI. The so-called "AI-positive" patients are triaged to orthopedics for clinical procedures and treatment. The radiology assessment thus often occurs *after* the orthopedic procedure is complete (see Figure 1). The added clinical value of the post-procedure radiology assessment is unknown, and the hospital wants to assess whether it is viable to end the practice of radiologist/AI double-assessment of *non-complex* AI-positive cases, to free radiology resources for more complicated tasks.

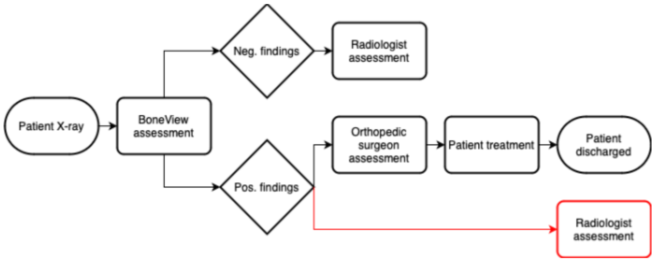

Figure 1. Current workflow, radiologist assessment step proposed removed highlighted in red.

To evaluate the clinical impact of this workflow modification in a live clinical environment, this paper outlines a retrospective observational study in progress. The study aims to determine the added clinical value of radiologists' assessments for AI-positive images by focusing on False Positive (FP) AI results identified through orthopedic or radiologic evaluation. Expert reviews of clinical documentation will further clarify the consequences for patient care, offering insights into whether radiologist oversight remains essential in AI-driven fracture detection.

The paper is structured as follows: Section 2 discusses AI integration into radiology workflows and the implications of removing radiologist assessments. Finally, Section 3 describes the study methodology, including data collection, sampling, analyses, and ethical considerations. Section 4 outlines the expected outcomes and their impact on clinical practice.

II. Background

AI integration in radiology has traditionally followed a "radiologist-in-the-loop" model, where AI serves as a supportive tool rather than an autonomous decision-maker [3]. However, recent discussions have explored whether low-complexity cases (such as straightforward fractures) could be managed without radiologist reassessment to optimize resource utilization. Studies have suggested that AI can reduce interpretation time for radiologists [4] and even improve fracture detection sensitivity compared to junior radiologists [5]. However, these studies primarily assess AI's diagnostic capabilities, not its role in workflow changes – a critical gap in current research.

AI models used in fracture detection are trained on large datasets and often exhibit high sensitivity, meaning they excel at identifying potential fractures. However, this increased sensitivity comes at the expense of higher false positive rates, where AI mistakenly flags normal findings as fractures or experiences difficulty differentiating old fractures from new ones. While previous studies have examined false negative rates (i.e., AI missing fractures) as a safety concern, less attention has been given to false positives, which may lead to unnecessary imaging, overtreatment, and increased healthcare costs. The real-world implications of radiologists reassessing AI-positive cases post-treatment have not been systematically evaluated, making it essential to investigate whether this step improves patient outcomes or merely confirms decisions already made by orthopedic surgeons.

Workflow efficiency is a key priority for radiology departments as imaging volumes continue to rise and workforce shortages persist [6]. Reducing radiologist involvement in AI-positive cases could allow radiologists to focus on more diagnostically complex or uncertain cases, thereby improving overall patient care. However, any workflow modification must be empirically validated to ensure that it does not introduce unintended clinical risks. This study aims to address this by assessing whether post-treatment radiologist assessment of AI-positive cases adds clinical value. Unlike previous studies that primarily assess AI's diagnostic accuracy, this study focuses on workflow efficiency, resource utilization, and patient outcomes.

III. Methodology and ethical consideration

A. Study design and setting

This study is based on a retrospective, exploratory design, aimed at evaluating the clinical consequences of workflow modifications introduced by integrating AI-driven applications into clinical practice. The retrospective, exploratory study design is particularly suited for initial investigations of workflow modifications, as it allows real-world data to be analyzed without artificially altering standard practice [7]. The study is conducted in a hospital setting where approximately 2,300 patients undergo AI-supported radiological evaluation over a five-month period. The population includes patients for whom the BoneView application indicates a positive finding. Among these, cases where subsequent clinical evaluation determines no need for treatment – designated as false positives – are included in the final dataset.

The use of FP cases provides a focused and practical approach to investigating the potential added clinical value of radiologists' assessments in AI-positive workflows. These cases represent scenarios where BoneView identifies a positive finding, but subsequent human evaluation by physicians concludes with no relevant clinical findings. By examining such cases, the study isolates instances where radiologists' expertise might either confirm or challenge the AI's assessment. This enables the evaluation of whether radiologists' interpretations contribute to improved diagnostic accuracy, prevent overtreatment, or identify subtleties missed by AI alone.

False negative cases were not included in this study because they are not relevant to the proposed workflow modification. The study evaluates the impact of removing radiologists' assessment of AI-positive cases, meaning that all cases in which the AI indicates a negative finding will continue to be reviewed by radiologists as part of standard practice. Since false negatives occur when the AI fails to detect a fracture, these cases would still undergo radiologist evaluation under the modified workflow. As a result, their inclusion would not contribute to answering the primary research question, which focuses on whether post-procedure radiologist assessment of AI-positive cases adds clinical value. Focusing on FP cases allows for a targeted assessment and are particularly suited for this investigation since they highlight potential discrepancies in AI performance, offering critical insight into the role of human oversight in ensuring patient safety and diagnostic rigor.

B. Data collection and sampling

FP cases will be identified prospectively at the point of care when either an orthopedic surgeon or a radiologist determines an AI-positive case to lack sufficient evidence of fracture. In these cases, patient identifiers will be recorded for retrospective analysis, see Figure 2 for study procedure. To ensure the rigor of the research, it is critical that the process of data collection remains independent of clinical decision-making and treatment. All patients will receive treatment-as-usual, following established clinical protocols, regardless of whether their case is identified and registered as FP. This independence ensures that the registration of FPs does not influence or alter the treatment decisions made by the clinicians, ensuring that the results accurately reflect the impact of workflow modifications without interference from the data collection process.

Based on expert estimates from the early implementation phase, the prevalence of FP is anticipated to be 1-2%. Based on this estimate, the study is expected to identify and include 20-40 FP cases during the five-month data collection period.

While this sample may be limited in detecting rare but high-impact clinical events, these cases represent a targeted subset derived from an estimated total of approximately 2,300 AI-evaluated X-ray examinations. The focused selection of FPs inherently constitutes a 'funnel' from a larger data pool, enhancing specificity and clinical relevance. Thus, despite the smaller number of cases included in detailed analyses, the extensive initial dataset bolsters the representativeness and applicability of our findings to broader clinical practice, and provides sufficient depth for an exploratory analysis of workflow effects.

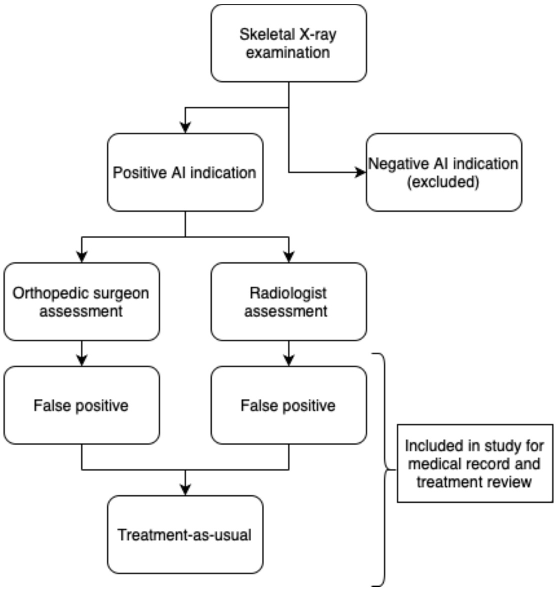
Skeletal X-ray examination
Positive AI indication
Negative AI indication (excluded)
Orthopedic surgeon assessment
Radiologist assessment
False positive
False positive
Included in study for medical record and treatment review
Treatment-as-usual

Figure 2. Study procedure flowchart.

C. Outcome measures and analysis

The primary outcome measure is the potential clinical harm or unnecessary care resulting from false positive AI assessments when radiologist evaluation is omitted. Secondary outcomes include (1) the frequency of overtreatment related to AI-positive findings, (2) patient recall rate, (3) resource utilization metrics (e.g., additional imaging or extended clinical encounters), and (4) the concordance between orthopedic and radiologist evaluations. A panel of experienced clinicians (radiologists and orthopedic surgeons) will independently review the medical records and provided treatments of the included FP cases and apply structured criteria to evaluate the degree of clinical impact using a standardized rubric (see Table 1). To minimize potential biases introduced by differing clinician thresholds for identifying false positives, all included cases will undergo independent review by an expert panel comprising both radiologists and orthopedic surgeons. For each FP case, ground truth will be retrospectively set by a specialist in musculoskeletal radiology (i.e., the *"gold standard"*)

Using the documented cases, the study will analyze the clinical pathways to identify the role of radiologist's evaluations in mitigating adverse outcomes. The analysis will focus on mapping the potential effects of radiological input on treatment decisions, providing a qualitative and quantitative basis for assessing the workflow change. Quantitative data (e.g., number of FP cases, frequency of overtreatment) will be summarized using descriptive statistics. We will also conduct comparative analyses to explore relationships between FP cases and patient demographics or fracture types. Qualitative data, including expert panel evaluations of clinical impact, will be thematically analyzed using framework method [8] to identify patterns in decision-making and diagnostic discrepancies.

TABLE 1. INFORMATION TO BE COLLECTED DURING REVIEW PHASE

Nr.	Category	Predefined items
1	Serial nr.	
2	PACS nr.	
3	Gender	Male; Female
4	Age	0-20; 20-60; 60+
5	Bone region	
6	Fracture type	
7	Other clinical findings	Bone lesion; Hydrops; Luxation
8	Patient referral from	
9	Patient sent to after X-ray examination	
10	AI fracture indication	Positive; Negative; Doubt
11	Image evaluation from radiologist	Positive; Negative; Doubt
12	Image evaluation from orthopedic surgeon	Positive; Negative; Doubt
13	Treatment implication	Overtreatment; Orthopedic FP identification; Radiologist FP identification; Patient recall

Overtreatment is defined in this study as any unnecessary medical intervention (e.g., additional imaging, orthopedic procedures) that would not have occurred had the radiologist reviewed and correctly classified the AI-positive case. Patient recall refers to instances where a patient is asked to return for further evaluation due to an AI-generated false positive result. Resource utilization encompasses additional diagnostic procedures, prolonged clinical encounters, and increased workload for radiologists and orthopedic surgeons. AI-based systems like BoneView may occasionally flag incidental findings (e.g., bone lesions, luxations) that do not directly correspond to fractures. These cases are assessed by clinical reviewers to determine whether they contribute to false positive classifications and whether their presence influences resource utilization or overtreatment.

To translate expert chart reviews into meaningful metrics, we will apply a structured rubric that categorizes clinical consequences based on severity, distinguishing cases with minor clinical implications from those leading to significant clinical harm or unnecessary interventions.

D. Ethical considerations

According to the Norwegian Act on Medical and Health Research §2 and §4, the study does not require approval from the regional ethics committee (REK). Data handling and storage will comply with institutional and national privacy standards and regulations, with approval from the hospital's Data Protection Officer. Informed consent will be obtained from patients identified as FP prior to accessing their medical records for detailed analysis. The study does not entail any change in workflow during the duration of the project, and all patients will receive treatment-as-usual.

IV. EXPECTED OUTCOMES AND STUDY IMPLICATIONS

Randomized Controlled Trials (RCTs) are the gold standard for investigating clinical outcomes from clinical interventions. However, the number of studies adhering to this methodology that address clinical outcomes of AI implementations in healthcare are presently limited and concentrated around a few geographical clusters [9]. While clinical evidence for AI in healthcare is certainly in demand, we encourage an effort to develop study designs that can explore the clinical proxy outcomes of AI implementation in locations that have implemented AI in clinical practice. The outcomes of this study may provide critical insights into the integration of AI applications into clinical workflows, addressing a key gap in the existing knowledge [10]. While pre-market certifications validate the safety and clinical performance of AI applications, it does not address their operational downstream effects on clinical workflows and resource allocation. This study bridges this gap by exemplifying a study design to evaluate workflow modifications and their implications, addressing the growing need for post-deployment validation frameworks.

We expect the final findings to contribute to understanding how to optimize resource allocation without compromising patient safety. Specifically, the study aims to identify the extent to which radiological evaluations influence treatment decisions and mitigate risks in specific patient cases. Moreover, the study's findings could inform triage protocols by identifying which clinical cases warrant immediate radiologist input and which can safely proceed without additional radiologist review. By discerning patterns in AI performance – particularly for different fracture types or demographic groups – clinicians and administrators could prioritize high-risk patient cohorts for expedited evaluation. By specifically focusing on FP cases, this study isolates the cohort of patients most vulnerable to overtreatment and misdiagnosis in the absence of radiological review – thereby providing a high-fidelity assessment of the necessity for radiologist oversight in image assessment in non-complex facture cases.

This approach aligns with the broader literature emphasizing the need to balance efficiency and safety in healthcare AI deployment [11][12]. Additionally, the proposed methodology supports scalability and adaptability, offering a replicable method for other institutions to assess AI-induced workflow changes, and underscores the importance of clinician involvement to ensure that workflow adjustments maintain transparency and clinical relevance.

REFERENCES

[1] A. Muley, P. Muzumdar, G. Kurian, and G. P. Basyal, "Risk of AI in Healthcare: A comprehensive literature review and study framework," *arXiv preprint arXiv:2309.14530,* 2023.

[2] L. Silsand, M. Kanneløning, G.-H. Severinsen, and G. Ellingsen, "Enabling AI in Radiology: Evaluation of an AI Deployment Process," *Studies in health technology and informatics,* vol. 316, pp. 580-584, 2024.

[3] R. Najjar, "Redefining radiology: a review of artificial intelligence integration in medical imaging," *Diagnostics,* vol. 13, no. 17, p. 2760, 2023.

[4] S. Jeong *et al.*, "The Impact of Artificial Intelligence on Radiologists' Reading Time in Bone Age Radiograph Assessment: A Preliminary Retrospective Observational Study," *Journal of Imaging Informatics in Medicine,* pp. 1-9, 2024.

[5] R. Y. Kuo *et al.*, "Artificial intelligence in fracture detection: a systematic review and meta-analysis," *Radiology,* vol. 304, no. 1, pp. 50-62, 2022.

[6] R. Bruls and R. Kwee, "Workload for radiologists during on-call hours: dramatic increase in the past 15 years," *Insights into imaging,* vol. 11, pp. 1-7, 2020.

[7] P. Aithal and S. Aithal, "Redefining Experimental, Empirical, and Exploratory Research in AI Era," *Poornaprajna International Journal of Emerging Technologies (PIJET),* vol. 1, no. 1, pp. 90-136, 2024.

[8] N. K. Gale, G. Heath, E. Cameron, S. Rashid, and S. Redwood, "Using the framework method for the analysis of qualitative data in multi-disciplinary health research," *BMC medical research methodology,* vol. 13, no. 1, pp. 1-8, 2013.

[9] T. Y. Lam, M. F. Cheung, Y. L. Munro, K. M. Lim, D. Shung, and J. J. Sung, "Randomized controlled trials of artificial intelligence in clinical practice: systematic review," *Journal of Medical Internet Research,* vol. 24, no. 8, p. e37188, 2022.

[10] A. Jackson and B. Hirsch, "Changing the workflow–Artificial intelligence in radiologic sciences," vol. 55, ed: Elsevier, 2024, p. 101710.

[11] M. A. Sujan, "Looking at the Safety of AI from a Systems Perspective: Two Healthcare Examples," in *Safety in the Digital Age: Sociotechnical Perspectives on Algorithms and Machine Learning*: Springer Nature Switzerland Cham, 2023, pp. 79-90.

[12] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine,* vol. 25, no. 1, pp. 44-56, 2019.

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-270-8