

Toward a clinically integrated AI framework for personalized spine care

Alemu Sisay Nigru^{a,b,*,*}, Sergio Benini^a, Yao Wang^d, Federico Maffezzoni^{b,c}, Matteo Bonetti^{b,c}, Michele Frigerio^c, Graziella Bragaglio^{b,c}, Serena Miglio^{b,c}, Riccardo Leonardi^a

^a Department of Information Engineering, University of Brescia, 25133, Brescia, Italy

^b X-ray Service s.r.l., Via Guglielmo Oberdan, 126, 25128 Brescia, Italy

^c Poliambulatorio Oberdan, Via Guglielmo Oberdan, 126, 25128 Brescia, Italy

^d Electrical and Computer Engineering Department, New York University, 370 Jay Street, Brooklyn, NY 11201, USA

ARTICLE INFO

Keywords:

Low back pain
Spine care
Multimodal artificial intelligence
Patient-reported outcome measures (PROMs)
Human-in-the-loop AI
Clinical decision support systems
Large language models (LLMs)

ABSTRACT

Background and Objective: Low back pain (LBP) is the leading global cause of disability, arising from complex structural, clinical, behavioral, and psychosocial factors. Although artificial intelligence (AI) has demonstrated considerable promise in spine imaging, automated diagnosis, and outcome prediction, most existing approaches remain task-specific, with limited integration of multimodal clinical information into routine spine-care workflows. Rather than proposing another disease-specific predictive model, this work presents a modular, clinically oriented AI framework that integrates imaging analysis, patient-reported outcomes, structured clinical information, and retrieval-aware language synthesis within a unified, clinician-supervised workflow to support future multimodal decision-making in spine care.

Framework Design: The conceptual, implementation-independent architecture comprises three interoperable components: (1) an imaging-analysis module supporting automated interpretation and severity grading with human-in-the-loop validation; (2) a multimodal assessment component integrating patient-reported outcomes with demographic and clinical variables; and (3) a retrieval-aware large language model (LLM) synthesis layer generating clinician-reviewable radiological summaries and contextualized clinical considerations grounded in structured evidence. Core design principles include interoperability, explainability, governance, security, and human-centered workflow integration.

Early Feasibility Demonstration: An exploratory proof-of-concept prototype incorporating AI-assisted lumbar MRI interpretation demonstrated how selected framework components may support explainable imaging analysis, multimodal contextualization, editable reporting, and clinician supervision. This demonstration illustrates architectural feasibility and workflow integration only, and should not be interpreted as evidence of diagnostic accuracy, clinical effectiveness, or generalizable clinical performance.

Conclusions: The proposed framework provides an implementation-independent translational blueprint for future technical development, prospective multicentre validation, regulatory evaluation, and real-world deployment of clinically integrated, trustworthy multimodal AI systems for spine care.

1. Introduction

1.1. Clinical background and motivation

Low back pain (LBP) remains the leading global cause of disability, with lifetime prevalence estimates approaching 84% and more than 619 million affected individuals worldwide in 2020, a number projected to increase substantially over coming decades [1]. Beyond its epidemiological burden, LBP imposes major socioeconomic costs through healthcare expenditure, work absenteeism, and long-term functional impairment [2,3]. Its clinical management is particularly challenging

because LBP is not solely a structural disorder, but rather a multifactorial condition shaped by anatomical, clinical, behavioral, and psychosocial factors [4]. Consequently, patients with similar imaging findings may experience markedly different pain severity, disability trajectories, coping responses, and treatment outcomes [5,6].

Although spine imaging plays a central role in diagnosis and treatment planning, imaging findings alone often show limited correlation with symptom severity and functional outcomes [7]. At the same time, psychosocial and behavioral dimensions known to influence prognosis and recovery are inconsistently integrated into routine clinical

* Correspondence to: Via Branze 38, Brescia, Italy.
E-mail address: alemu.nigru@unibs.it (A.S. Nigru).

<https://doi.org/10.1016/j.imu.2026.101802>

Received 21 March 2026; Received in revised form 5 August 2026; Accepted 7 August 2026

Available online 13 August 2026

2352-9148/© 2026 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

workflows [8,9]. Patient-reported outcome measures (PROMs), including pain scales, disability indices, quality-of-life instruments, and coping assessments, provide important contextual information that is frequently evaluated separately from imaging and structured clinical data [10–12]. This fragmentation limits the development of individualized, biopsychosocially informed treatment strategies capable of capturing the full complexity of chronic low back pain.

1.2. AI and multimodal integration in spine care

Recent advances in artificial intelligence (AI), particularly deep learning methods, have demonstrated strong performance in automated spine imaging analysis. Prior studies have reported promising results in the detection and grading of disk degeneration, stenosis, Modic changes, fractures, and related spinal pathologies across large imaging datasets [13–17]. Nevertheless, most existing AI approaches remain focused on isolated tasks such as image interpretation or outcome prediction, with relatively limited integration into comprehensive clinical decision-making workflows [18,19]. Furthermore, only a limited number of proposed systems have undergone rigorous external validation or meaningful evaluation within routine clinical workflows.

Translating AI models from experimental environments into clinical practice requires more than predictive performance alone. Successful translation into routine clinical practice depends not only on predictive performance but also on interoperability with hospital infrastructure, including electronic health records (EHRs), radiology information systems (RIS), and picture archiving and communication systems (PACS), as well as compliance with privacy, governance, and regulatory requirements. Equally important is the need for clinical trust and transparency. Clinicians must be able to review, interpret, verify, and, where necessary, modify AI-generated outputs rather than relying on opaque “black-box” predictions. Explainable AI techniques, including saliency visualization, attention mapping, and feature attribution methods, have therefore become increasingly important for safe and interpretable clinical AI deployment [20–22].

In parallel, large language models (LLMs) and retrieval-augmented generation (RAG) approaches are increasingly being explored for clinical documentation and decision-support applications. Early studies suggest that such systems may assist with radiology reporting, structured summarization, administrative workflow support, and patient-oriented communication [23–25]. More recent studies have begun moving beyond single-task AI models toward multimodal and clinically oriented AI systems that integrate medical imaging, structured clinical information, language models, and explainability mechanisms. For example, recent federated multimodal transformer architectures have combined medical imaging and clinical narratives with patient-specific explainability to support privacy-preserving and explainable diagnostic workflows in oncology [26], while other studies have explored retrieval-augmented clinical reasoning, multimodal report generation, and treatment-oriented decision support in low back pain and broader healthcare settings [27–30]. Despite these advances, existing systems generally remain disease-specific, application-specific, or centered on individual components of the clinical workflow rather than providing a unified framework that integrates imaging AI, patient-reported outcome measures (PROMs), structured clinical information, clinician oversight, interoperability, governance, and language-based decision support. Consequently, although substantial progress has been achieved across individual technical domains, relatively little attention has been devoted to defining how these complementary technologies can be orchestrated into clinically deployable, human-centered AI ecosystems for spine care. To our knowledge, few framework-oriented studies have explicitly combined these elements within a modular, implementation-independent architecture designed to support future adaptation across heterogeneous clinical environments. Furthermore, concerns regarding hallucination, traceability, explainability, uncertainty calibration, and

regulatory oversight remain major barriers to clinical adoption, reinforcing the importance of human-in-the-loop validation and evidence-grounded generation workflows [31,32].

Several recent studies further illustrate the fragmented evolution of the current spine-AI landscape. Quinones et al. [33] demonstrated the value of combining PROMs with machine learning for predicting postoperative outcomes in spine surgery, but their analysis remained limited to surgical cohorts and outcome prediction tasks. Chen et al. [27] proposed a retrieval-augmented LLM framework (CLBP-GPT) for treatment decision support in low back pain; however, the system did not incorporate imaging interpretation or multimodal diagnostic integration. Conversely, imaging-centered systems such as those proposed by Lu et al. [34] and Windsor et al. [16] focus primarily on automated MRI analysis without integration of PROMs, longitudinal clinical context, or language-based synthesis. Similarly, GPT-based extraction frameworks for radiology reporting [35] remain largely confined to isolated reporting tasks rather than end-to-end clinical orchestration. Collectively, these studies, including recent reviews [36], demonstrate that, despite rapid methodological advances, current spine-care AI remains fragmented across individual technical domains, with limited integration of imaging analysis, patient-reported assessment, clinical reasoning, and workflow-oriented deployment.

1.3. Comparison with related work

Table 1 positions the proposed framework relative to representative spine-care AI systems spanning imaging analysis, PROM-based machine learning, LLM/RAG-assisted clinical support, multimodal workflow orchestration, and AI-assisted radiology reporting. SpineNetV2 and related systems demonstrate strong automated spine MRI analysis, including vertebral localization and grading of degenerative findings, but remain primarily imaging-centered rather than workflow-integrated multimodal systems [16]. PROM-based machine learning studies and reviews emphasize postoperative outcome prediction after spine surgery, but generally do not integrate imaging interpretation, clinician-supervised reporting, or multimodal orchestration [33]. CLBP-GPT represents an important step toward RAG-supported treatment-oriented reasoning in chronic low back pain, but does not incorporate automated imaging interpretation or radiologist-facing validation workflows [27].

Similarly, GPT-based radiology-report extraction and generation studies demonstrate the growing potential of LLMs for structured reporting and report synthesis, but remain primarily text-centered and do not integrate PROMs, longitudinal clinical context, or multimodal clinician-supervised decision support [35,37]. More recent multimodal workflow-oriented frameworks, such as the multi-agent radiology report generation architecture proposed by Yi et al. [30], further extend this direction by combining imaging analysis, LLM-based orchestration, and clinician-aligned workflow decomposition within a modular reporting pipeline. Although not specifically designed for low back pain or spine-care environments, such architectures demonstrate how multimodal and multi-agent orchestration strategies may be transferable to future spine-care AI ecosystems.

Beyond spine care, recent work has also proposed multimodal federated transformer frameworks that integrate medical imaging, clinical narratives, and explainability for leukemia diagnosis and staging under privacy-preserving training settings. Although these studies demonstrate the feasibility of clinically oriented multimodal AI systems, they remain disease-specific and algorithm-focused rather than addressing modular clinical workflow orchestration across heterogeneous data sources and decision-support components [26]. Collectively, these studies demonstrate that substantial progress has been achieved within individual methodological domains; however, none provides a comprehensive framework encompassing multimodal data integration, clinician-supervised decision support, interoperability, governance, and implementation guidance specifically for spine care.

Table 1

Comparison of representative spine-care AI approaches and the proposed modular framework.

Approach	Img	PROM	EHR	LLM	HITL	Flow	Core scope
SpineNetV2 [16]	✓	–	–	–	◦	Low	Automated MRI localization and degenerative grading.
PROM-based studies [33]	◦	✓	◦	–	–	Low	Postoperative PROM prediction in spine surgery.
CLBP-GPT [27]	–	◦	◦	✓	◦	Mod.	RAG-assisted treatment-oriented LLM support.
GPT-4 Extraction [35]	–	–	–	✓	–	Low	Structured extraction from lumbar MRI reports.
LLM MRI Reporting [37]	–	–	–	✓	◦	Low	LLM-generated lumbar MRI reporting workflows.
Multi-Agent RRG [30]	✓	–	◦	✓	✓	Mod.	Multimodal multi-agent radiology report generation with clinically aligned workflow orchestration.
Proposed Framework	✓	✓	✓	✓	✓	High	Integrated multimodal framework with imaging, PROMs, EHR, LLM/RAG, HITL, governance, and workflow orchestration.

✓: explicit integration; ◦: partial or indirect support; –: not addressed; Img: Imaging; PROM: Patient-reported Outcome Measures; EHR: Electronic Health Record; LLM: Large Language Model; HITL: Human-in-the-Loop; Flow: degree of clinical workflow integration.

Overall, existing work demonstrates that the individual technical components required for AI-assisted spine care are advancing rapidly. However, these developments remain fragmented across imaging-only models, PROM-based prediction systems, LLM-assisted reporting tools, and workflow-specific applications. The principal novelty of the present study lies not in proposing another disease-specific predictive model, but in presenting a modular, clinically oriented framework that explicitly integrates imaging AI, PROMs, structured clinical information, retrieval-augmented language synthesis, human-in-the-loop validation, interoperability, and governance within a single translational architecture. By remaining implementation-independent, the framework is intended to provide a flexible blueprint that can be adapted to diverse clinical environments while supporting future technical development, prospective validation, and real-world deployment.

1.4. Study objective and scope

Motivated by these gaps, this study proposes a modular framework for clinically integrated AI-assisted spine care that is designed to support multimodal decision-making through the integration of imaging analysis, structured clinical information, patient-reported outcome measures (PROMs), and language-based synthesis within a clinician-supervised workflow.

Importantly, the framework intentionally remains technology-neutral and does not prescribe specific imaging modalities, fusion strategies, model architectures, or deployment configurations, thereby enabling future adaptation to institutional requirements and technological advances.

Rather than presenting a fully validated clinical AI platform, this work is intended as a translational architectural blueprint accompanied by an early feasibility demonstration. To illustrate practical workflow integration, a prototype implementation using AI-assisted lumbar MRI interpretation and structured report synthesis is explored on a small set of representative clinical cases. The objective is not to establish diagnostic performance, clinical effectiveness, or generalized workflow efficiency, but rather to illustrate how multimodal integration, clinician-in-the-loop validation, and explainable reporting interfaces may be operationalized within a unified spine-care workflow. Specifically, this study makes the following contributions: (1) it proposes a modular framework for clinically integrated spine-care AI that unifies imaging analysis, PROMs, structured clinical information, and language-based synthesis; (2) it embeds explainability, interoperability, governance, and human-in-the-loop validation as core architectural principles; (3) it

illustrates one possible implementation through an exploratory proof-of-concept prototype; and (4) it establishes a translational foundation for future technical development, prospective multicentre validation, regulatory evaluation, and real-world deployment of multimodal AI systems in spine care.

2. Clinical motivation and design requirements

2.1. Stakeholder needs and clinical decision points

A clinically integrated spine-care AI framework should support multiple stakeholders across the diagnostic and therapeutic continuum of low back pain (LBP). Because different clinical specialties rely on distinct sources of information and make decisions at different stages of the care pathway, effective AI integration requires workflow-aware coordination rather than isolated algorithmic outputs.

Radiologists require reliable imaging interpretation tools that support efficient review, structured reporting, and transparent verification of automated findings. Beyond predictive performance alone, clinically usable systems should provide editable reporting interfaces, confidence estimates, and explainable visualizations that enable radiologists to validate, modify, or reject AI-generated assessments before clinical reporting.

Pain specialists (e.g., neurologists, spine surgeons, interventional pain physicians, anesthesiologists) require synthesis of multimodal information, including imaging findings, functional status, psychosocial indicators, and comorbidities, to support individualized risk stratification and treatment planning. Such integration may facilitate more informed risk stratification and support decisions regarding conservative management, interventional pain procedures, and referral for surgical evaluation.

Rehabilitation clinicians (e.g., physiotherapists, psychologists, occupational therapists, physiatrists, nurses) require longitudinal and patient-centered assessments that contextualize structural abnormalities within functional limitations, coping behavior, and quality-of-life measures. These factors influence rehabilitation planning, treatment adherence, patient engagement, and long-term recovery trajectories.

Collectively, these stakeholder requirements highlight the need for interoperable, human-centered AI systems capable of integrating heterogeneous clinical information into coherent, transparent, and clinically actionable workflows.

Low Back Pain Patient Journey Workflow

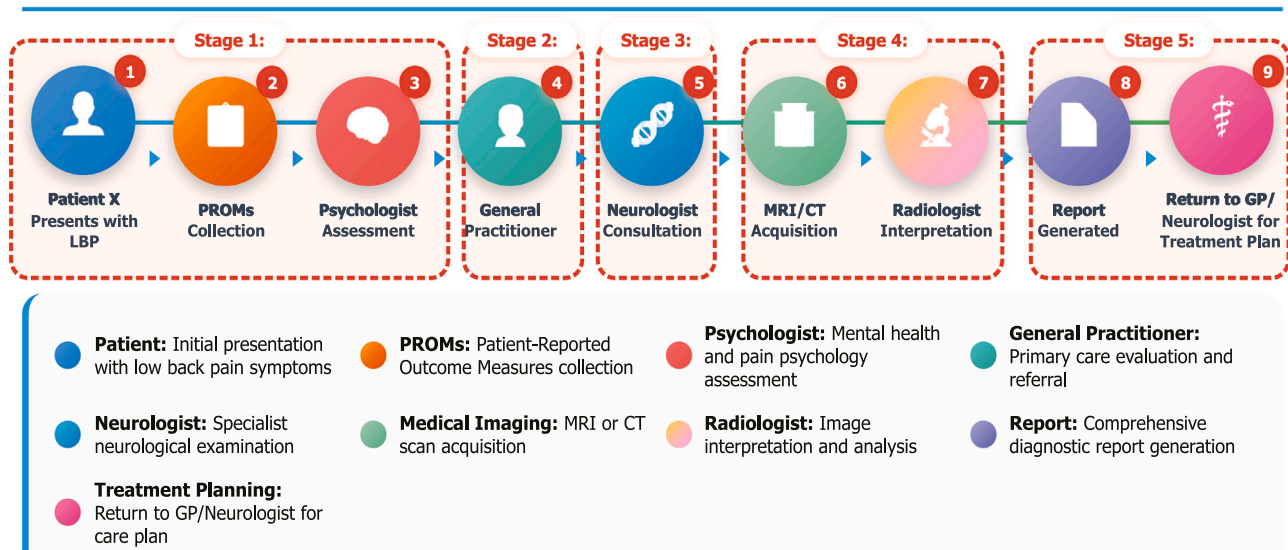


Fig. 1. Representative multidisciplinary workflow for low back pain assessment and treatment planning. The illustrated pathway summarizes major clinical stages frequently encountered in specialized spine-care settings, including patient presentation, PROMs collection, psychological assessment, primary-care evaluation, neurological consultation, imaging acquisition, radiological interpretation, report generation, and treatment planning. Actual workflows may vary across institutions and healthcare systems.

2.2. Clinical motivation and translational value

From a clinical perspective, one of the principal limitations of contemporary spine-care workflows is the fragmentation of imaging findings, functional assessments, psychosocial information, and longitudinal patient data across disconnected systems and clinical encounters. This fragmentation increases cognitive burden, introduces variability in interpretation, and complicates multidisciplinary coordination.

The proposed framework is motivated by the need to support four practical objectives relevant to real-world clinical environments:

Multimodal information synthesis. Clinicians routinely interpret imaging findings alongside functional assessments, patient-reported symptoms, comorbidities, and psychosocial context. Integrating these heterogeneous data streams within a unified clinical framework may improve contextual interpretation, reduce information fragmentation, and support more holistic clinical decision-making.

Clinician-centered decision support. Rather than replacing clinical expertise, the proposed framework is designed to augment clinical reasoning through explainable outputs, editable reports, and evidence-linked summaries while preserving clinician authority over final decisions.

Interdisciplinary workflow coordination. Low back pain management frequently involves multiple specialties operating across disconnected workflows. Shared multimodal outputs may improve communication and decision consistency across radiologists, neurologists, rehabilitation teams, spine surgeons, pain specialists, and primary-care clinicians.

Longitudinal learning and feedback integration. Clinician feedback, longitudinal follow-up data, and structured patient assessments may provide valuable signals for continuous quality improvement, post-deployment monitoring, and future learning-health-system applications while maintaining institutional governance and clinician oversight.

Collectively, these clinical considerations motivate a modular, workflow-oriented framework that is intended to facilitate future clinically interpretable, multimodal, and human-supervised AI integration across diverse spine-care environments.

2.3. Design requirements

To support safe and clinically deployable integration, the framework is designed around several core operational requirements.

Explainability and transparency. Clinical AI systems should provide interpretable outputs that enable clinicians to understand, verify, and appropriately contextualize model predictions within the broader clinical picture. Explainability mechanisms such as saliency visualization [38], feature attribution [39], and attention-based heatmaps [40] may help identify relevant image regions, highlight contributing variables, and improve transparency during clinical review. Beyond improving usability, explainability is increasingly recognized as a fundamental requirement for regulatory compliance, clinical governance, user trust, and safe deployment of high-risk medical AI systems [21,22,41].

Human-in-the-loop validation. The proposed framework adopts a human-in-the-loop (HITL) paradigm in which AI-generated outputs remain subject to clinician review, modification, and approval before clinical use. This workflow is intended to reduce automation bias, improve accountability, and preserve clinician authority over final decisions [42–44]. In future implementations, clinician corrections and structured feedback may additionally serve as valuable signals for continual model refinement, quality assurance, post-deployment surveillance, and adaptive governance.

Interoperability and workflow integration. For practical deployment, AI systems must integrate with existing hospital infrastructure rather than operate as isolated research tools. Accordingly, the framework is designed to facilitate interoperability with electronic health records (EHRs), radiology information systems (RIS), and picture archiving and communication systems (PACS) through established standards such as HL7 FHIR and DICOM Structured Reporting (DICOM-SR) [45,46]. This design objective aims to minimize workflow disruption while supporting scalable integration into heterogeneous clinical environments through institution-specific implementation strategies.

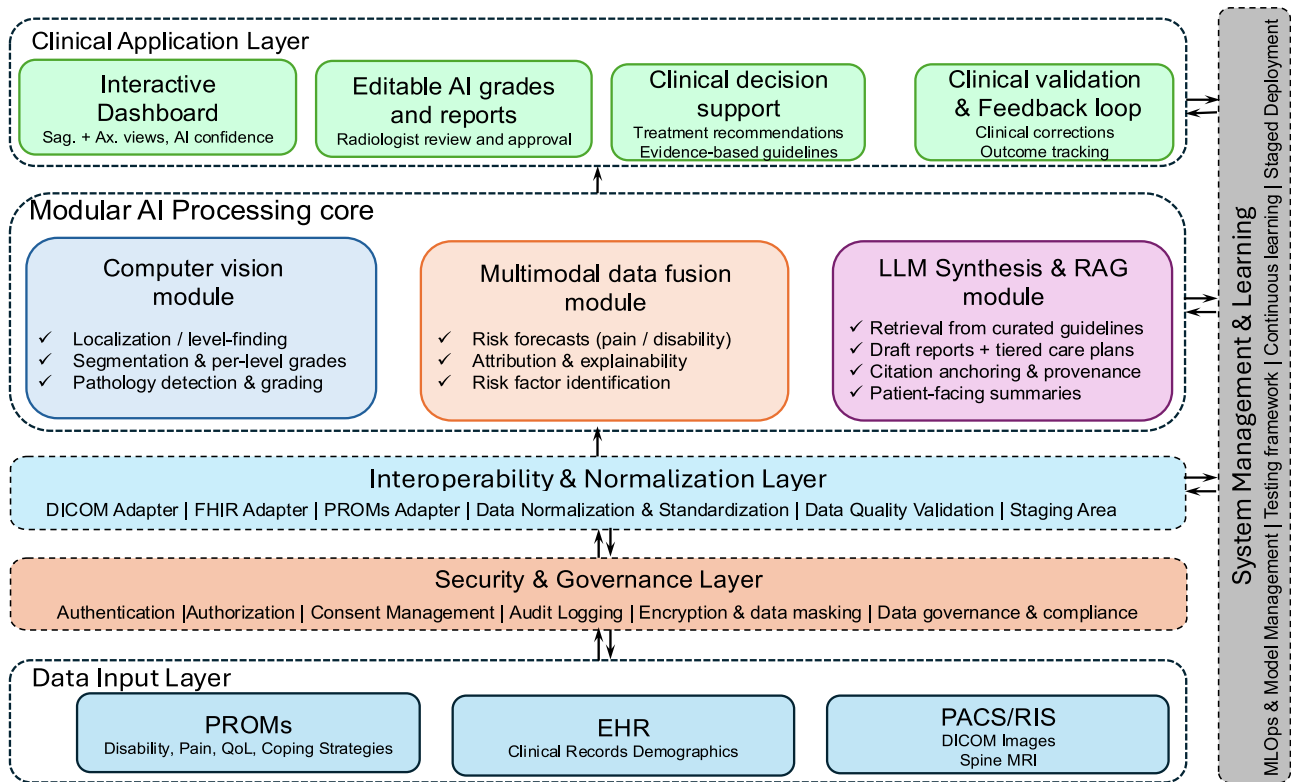


Fig. 2. Conceptual overview of the proposed modular AI framework for spine care integrating multimodal clinical data, clinician-supervised AI processing, interoperability layers, and continuous governance mechanisms. The framework is intentionally technology-neutral and designed to support future adaptation across different imaging modalities, multimodal fusion strategies, AI architectures, and clinical deployment environments.

2.4. Clinical workflow integration

Fig. 1 illustrates a representative multidisciplinary workflow for low back pain management and identifies the principal stages at which multimodal AI may augment, rather than replace, clinician decision-making. The pathway reflects a common structure observed in specialized spine and orthopedic care environments, although real-world workflows may differ substantially across institutions, healthcare systems, and resource settings.

The illustrated process includes five major stages:

1. Patient Presentation and PROMs Assessment. Patients initially present with low back pain symptoms and may undergo structured collection of patient-reported outcome measures (PROMs), including assessments of pain severity, disability, quality of life, and coping behavior. Depending on institutional practice, psychological evaluation may additionally be performed to assess psychosocial contributors influencing treatment response and recovery trajectories. Example instruments include the WHOQOL scale [47], Roland Disability Questionnaire [48], Coping Strategy Questionnaire [49], and Numerical Rating Scale (NRS) [50].

2. Primary Clinical Evaluation. General practitioners or referring clinicians document relevant medical history, demographic characteristics, comorbidities, occupational factors, and lifestyle variables that may influence risk stratification and treatment planning.

3. Neurological and Specialist Assessment. Neurological consultation evaluates motor deficits, sensory abnormalities, radicular symptoms, and other indicators relevant to imaging referral and treatment prioritization.

4. Imaging Acquisition and Interpretation. MRI or CT imaging studies are acquired and transferred through PACS infrastructure for radiological interpretation. Imaging assessment focuses on structural abnormalities, degenerative findings, stenosis, disk pathology, and related anatomical contributors.

5. Integrated Reporting and Treatment Planning. Clinical, imaging, and patient-reported information are synthesized into multidisciplinary treatment planning decisions. Within the proposed framework, this stage represents the principal integration point where imaging findings, structured clinical information, patient-reported outcome measures, and explainable AI outputs are synthesized into clinician-supervised decision support. Importantly, all AI-generated recommendations remain subject to expert review and final clinical judgement.

Although the sequence illustrated in Fig. 1 represents one possible clinical pathway, the proposed framework intentionally remains technology-neutral and adaptable across heterogeneous healthcare systems, imaging modalities, institutional workflows, and levels of digital maturity. This flexibility is intended to facilitate future implementation, prospective multicentre evaluation, and iterative refinement as clinical AI technologies continue to evolve.

3. System architecture overview

The proposed framework is a modular, interoperable clinical AI architecture designed to augment, rather than replace, clinician decision-making throughout the spine-care pathway. The architecture conceptually organizes heterogeneous clinical information into a layered workflow that integrates imaging analysis, structured clinical data, patient-reported outcomes (PROMs), and clinician-supervised language-based synthesis. Fig. 2 illustrates the conceptual system organization, emphasizing interoperability, explainability, governance, and human-in-the-loop validation.

Importantly, the framework is intentionally remains implementation-independent and does not prescribe specific imaging modalities, AI architectures, multimodal fusion strategies, or deployment configurations. This abstraction promotes extensibility, reproducibility, and adaptation across heterogeneous clinical environments while accommodating future methodological advances.

3.1. Data input layer

The framework interfaces with routine clinical data sources commonly available in hospital environments:

- **PACS/RIS:** DICOM-format imaging studies, including spine MRI or CT examinations.
- **EHR:** Structured electronic health record information such as demographics, clinical history, medications, lab tests and comorbidities.
- **PROMs:** Patient-reported assessments capturing pain severity, disability, quality of life, and coping-related measures.

Collectively, these heterogeneous data sources provide the multimodal foundation for downstream analytical processing and contextual clinical interpretation.

3.2. Security and governance layer

A dedicated governance layer mediates data access, exchange, and traceability across the framework. This layer is intended to support authentication, authorization, consent management, audit logging, encryption, and provenance tracking while maintaining compliance with institutional and regulatory requirements, including HIPAA- and GDPR-aligned data governance practices.

Conceptually, this middleware layer separates institutional data ownership from downstream AI services, thereby enabling modular integration, controlled deployment, and lifecycle governance of downstream AI components without requiring direct modification of existing hospital infrastructure.

3.3. Interoperability and normalization layer

The interoperability layer functions as an integration interface between clinical systems and AI modules. It incorporates standardized adapters and normalization pipelines designed to harmonize heterogeneous clinical inputs prior to downstream analysis. Example interfaces include DICOM connectivity for imaging data, HL7/FHIR-compatible structured clinical exchange, and standardized PROM ingestion pipelines.

Additional responsibilities include data validation, harmonization, provenance preservation, and workflow traceability to support reproducible and auditable AI-assisted clinical processes.

3.4. Modular AI processing core

The analytical core of the framework is organized into three complementary modules addressing image analysis, multimodal integration, and language-based synthesis.

3.4.1. Computer vision module

The computer vision component is designed to support automated analysis of spine imaging studies through tasks such as vertebral localization, segmentation, pathology detection, and severity grading. Depending on implementation, the module may operate on MRI, CT, X-ray, or other imaging modalities relevant to spine-care workflows. The framework intentionally does not prescribe a specific computer vision architecture, allowing future implementations to employ convolutional, transformer-based, hybrid, or other emerging imaging models according to the intended clinical application.

Outputs are structured as level-specific imaging descriptors (e.g., stenosis severity or disk degeneration grading) accompanied by visual overlays, confidence estimates, or uncertainty indicators intended to support radiologist verification and triage prioritization. Rather than functioning autonomously, the module is intended to assist radiologists by providing explainable, confidence-aware imaging assessments that remain subject to clinician verification before clinical reporting.

3.4.2. Multimodal data fusion module

The multimodal integration module combines heterogeneous data streams, including imaging-derived features, structured EHR variables, and PROM-related assessments, into unified patient-level representations. The framework does not constrain the underlying fusion methodology; depending on the implementation context, integration may be achieved using early-fusion, late-fusion, attention-based, graph-based, probabilistic, or hybrid strategies. This flexibility is intended to encourage adaptation of the framework across different institutional requirements, available data modalities, and future methodological advances.

The primary objective of this module is to contextualize structural imaging findings within broader clinical, functional, and psychosocial profiles thereby supporting more comprehensive risk stratification and personalized clinical decision support. Example outputs may include attributed risk summaries, longitudinal progression indicators, or multimodal contextual profiles linking predictions to contributing data sources.

3.4.3. LLM synthesis and RAG module

The language-generation component is designed to transform structured outputs from upstream modules into clinician-reviewable summaries, draft radiology reports, and contextualized clinical recommendations. To improve traceability and reduce unsupported generation, the framework conceptually incorporates retrieval-augmented generation (RAG) principles [51], in which relevant clinical guidelines, institutional protocols, or curated reference documents may be retrieved prior to text synthesis.

Rather than relying exclusively on parametric knowledge encoded within the language model, the proposed design seeks to ground generated outputs in externally retrieved evidence, thereby improving factual consistency, transparency, and clinician confidence. Depending on implementation, retrieved references may be linked to generated summaries to support clinician verification and transparency during review.

Because generative AI systems remain susceptible to hallucination, factual inconsistency, uncertainty, and bias, all outputs generated by this module are intended to remain subject to clinician review and approval prior to clinical use. The proposed framework further emphasizes evidence-grounded generation, provenance tracking, uncertainty communication, and auditability as important safeguards for future clinical deployment. Accordingly, the LLM synthesis layer is intended to function as an assistive documentation and clinical communication tool rather than an autonomous decision-making system, ensuring that responsibility for final interpretation remains with qualified clinicians.

3.5. Clinical application layer

The clinician-facing layer provides interfaces intended to support supervised interaction with AI-generated outputs and multimodal patient summaries. Example functionalities include:

- **Interactive imaging dashboard:** Visualization of sagittal and axial imaging studies with confidence indicators and explainability overlays.
- **Editable AI-assisted reporting:** Structured imaging summaries and draft reports subject to clinician review, correction, and approval.
- **Multimodal decision support:** Contextualized summaries integrating imaging, PROMs, and clinical variables to support treatment planning.
- **Feedback and validation logging:** Capture of clinician edits, overrides, and downstream outcomes for quality monitoring and future refinement workflows.

This clinician-facing layer operationalises the human-in-the-loop paradigm by preserving clinician authority over interpretation, report approval, and treatment recommendations while enabling structured AI assistance throughout routine clinical workflows.

3.6. System management and learning

The framework additionally incorporates a conceptual model-management and monitoring layer supporting deployment governance and iterative refinement. Example functions include version control, staged testing, performance monitoring, auditability, and controlled feedback-driven retraining.

Importantly, continuous learning is conceptualized as a governed process requiring institutional oversight, predefined validation checkpoints, regulatory compliance, and deployment safeguards rather than unrestricted autonomous model adaptation.

Collectively, these architectural layers define a modular workflow for integrating multimodal clinical information, clinician-supervised AI assistance, interoperable data exchange, and governance mechanisms within spine-care environments. The framework therefore provides a translational architectural foundation intended to support future implementation, prospective evaluation, and safe clinical deployment of multimodal AI systems.

3.7. Illustrative prototype implementation

To illustrate how selected components of the proposed framework may operate within a clinician-supervised workflow, an exploratory prototype platform, termed *LumbarAI*, was developed as a proof-of-concept implementation focused on AI-assisted lumbar spine MRI interpretation, PROM-based risk stratification, and structured report generation. The prototype was designed primarily to demonstrate workflow feasibility, multimodal integration, explainability mechanisms, editable clinician interaction, and human-in-the-loop validation rather than to establish diagnostic accuracy, clinical effectiveness, or workflow superiority.

The current exploratory prototype implementation employs a SWIN3D-based transformer architecture trained on sagittal lumbar spine MRI sequences for automated intervertebral disk (IVD) localization, vertebral labeling, and pathology grading. Vertebral body detection and anatomical labeling components were adapted from the SpineNetV2 framework proposed by Windsor et al. [16], while a separate SWIN-transformer-based grading model was developed for pathology severity assessment. The grading component was trained using a combination of publicly available lumbar spine MRI datasets, including RSNA, SPIDER, and Mendeley spine MRI datasets, together with an additional internally curated institutional dataset. The system processes sagittal MRI inputs and generates structured predictions across multiple lumbar levels, including degenerative grading and related pathological findings. Although the current prototype focuses on sagittal lumbar MRI analysis, the broader framework remains intentionally modality-independent and readily extensible to axial MRI, CT, radiography, and future multimodal imaging pipelines.

The imaging-analysis component generates structured tabular predictions for multiple pathological categories, including Pfirrmann degeneration grade, disk narrowing, central canal stenosis, spondylolisthesis, vertebral endplate defects, marrow signal abnormalities, foraminal stenosis, and disk herniation across lumbar levels. To enhance interpretability and clinician trust, the interface incorporates multiple explainability mechanisms, including vertebral localization, pathology-specific overlays, and attention-based visualizations that spatially associate AI predictions with the underlying anatomical structures. Vertebral bodies and intervertebral levels are automatically localized and highlighted directly on MRI images, while pathology-specific overlays and attention-based grading visualizations provide localized anatomical context for AI-generated findings. Interactive focus mechanisms allow clinicians to select specific lumbar levels within the structured prediction table and immediately visualize the corresponding highlighted anatomical regions and pathology overlays within the MRI viewer.

The prototype additionally incorporates confidence-aware safety mechanisms intended to support clinician oversight. Predictions below

a predefined confidence threshold are automatically flagged within a dedicated “Safety Alert” review panel, prompting clinicians to perform targeted secondary review before report approval. These confidence-aware alerts function as workflow-oriented safety checkpoints intended to encourage targeted clinician review rather than autonomous acceptance or rejection of AI-generated findings.

A FastAPI-based interactive user interface was implemented to support editable clinician review and structured correction workflows. Radiologists can directly modify predicted pathology grades, revise structured findings, and validate or reject AI-generated outputs before finalization. Importantly, modifications made within the structured prediction table dynamically update associated visual overlays and downstream report-generation components in real time. This bidirectional synchronization was designed to maintain consistency between structured findings, imaging visualization, and generated narrative reports throughout the review process.

In addition to imaging interpretation, the prototype incorporates a PROM-based patient risk screening module designed to operationalize the biopsychosocial dimension of the proposed framework. This module was derived from exploratory unsupervised clustering analyses performed on patient-reported outcome measures, including pain intensity (NRS), functional disability (Roland–Morris Disability Questionnaire), psychological quality-of-life metrics, and coping-strategy assessments. Exploratory unsupervised clustering analyses consistently identified two broad biopsychosocial patient profiles characterized by relatively adaptive and maladaptive patterns of pain, disability, psychological well-being, and coping behavior.

To facilitate practical clinical interpretation, the cluster assignments were subsequently translated into a continuous numerical risk score ranging from 0 to 200 using a logistic-regression-based predictive model. The interactive screening interface provides confidence-aware cluster prediction, numerical risk estimation, and feature-contribution visualization designed to highlight the relative influence of individual PROM variables on the predicted risk profile. These visualizations are intended to support clinician understanding of the dominant psychosocial and functional contributors potentially relevant for multidisciplinary intervention planning and treatment personalization.

All clinician modifications and validation actions within both the imaging and PROM interfaces are recorded through an audit-oriented change logging mechanism. These audit logs preserve both the original AI outputs and clinician-corrected decisions, thereby supporting future model calibration, quality assurance, post-deployment monitoring, and governed continual learning under appropriate institutional oversight.

The validated structured outputs are subsequently passed to a table-to-report synthesis module that automatically generates editable radiological narrative reports. The generated report remains fully clinician-editable, allowing radiologists to refine terminology, modify impressions, adjust phrasing, or incorporate additional clinical observations prior to approval. Finalized reports can then be exported into institutionally branded PDF templates for clinical documentation and communication workflows.

Representative screenshots of the prototype interface are shown in Figs. 3 and 4, including explainability overlays, confidence-aware safety alerts, editable structured prediction tables, PROM-based risk screening, and AI-assisted report generation workflows. Additional details regarding the exploratory proof-of-concept integration performed on representative patient cases are provided in the **Supplementary Material**.

Importantly, this prototype represents only one illustrative implementation of selected components of the proposed framework and should be interpreted exclusively as an exploratory demonstration of workflow feasibility rather than a clinically validated diagnostic platform. The current implementation has not undergone prospective multicentre validation, formal reader studies, regulatory evaluation, or real-world deployment assessment. Future work will focus on comprehensive quantitative evaluation of diagnostic performance, workflow

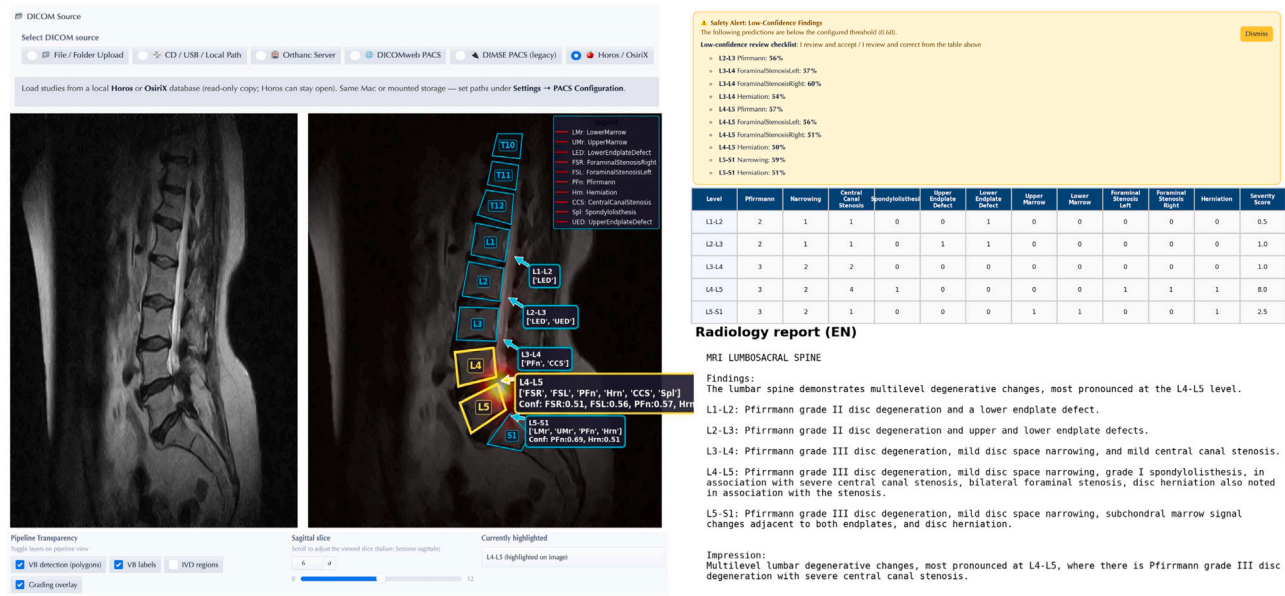


Fig. 3. Illustrative imaging-analysis and reporting workflow of the exploratory LumbarAI prototype. The interface demonstrates automated lumbar MRI interpretation with vertebral localization, pathology-specific explainability overlays, editable structured grading tables, confidence-aware safety alerts for low-certainty findings, and AI-assisted radiology report synthesis. Clinician modifications dynamically propagate across visualization and reporting components to support human-supervised workflow consistency.

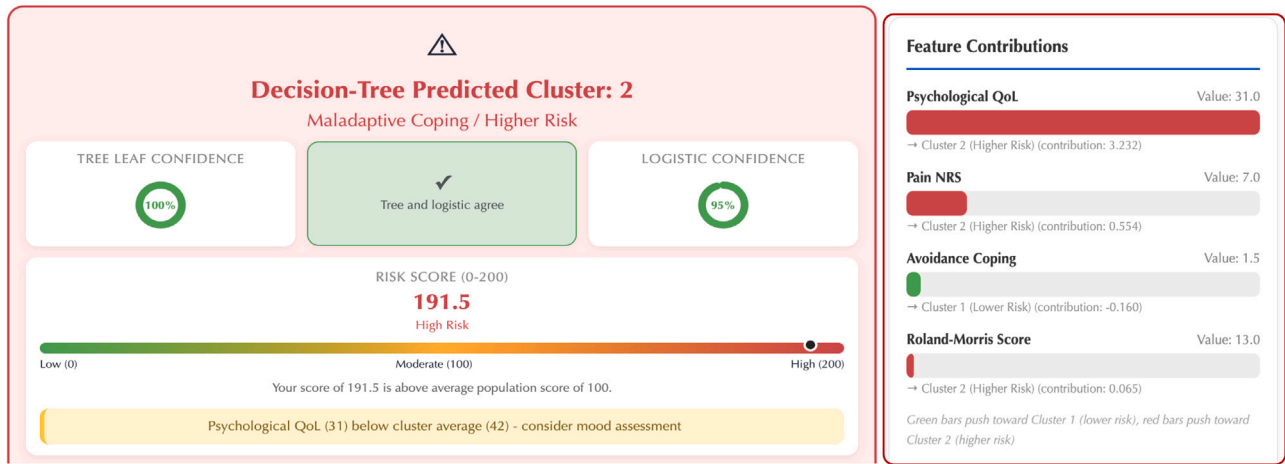


Fig. 4. PROM-based biopsychosocial risk-screening interface integrated within the exploratory framework. The module visualizes cluster-based patient stratification, interpretable numerical risk scoring, confidence estimation, and variable-level feature contributions derived from patient-reported outcome measures including pain intensity, functional disability, psychological quality of life, and coping behavior.

efficiency, clinician acceptance, interoperability, safety, and clinical utility across diverse healthcare environments. The primary purpose of the present prototype is therefore to demonstrate how explainable imaging AI, confidence-aware review mechanisms, PROM-integrated risk stratification, clinician-supervised editing, structured reporting, and longitudinal governance may be operationalized within a modular clinical workflow.

4. Illustrative clinical use scenarios

The following representative scenarios illustrate how selected components of the proposed framework may support multimodal, clinician-supervised workflows across different spine-care contexts. These examples are intended to demonstrate the translational role of the proposed architecture rather than the performance of any individual AI model. These examples are hypothetical and are intended to clarify potential interactions between imaging analysis, patient-reported outcomes

(PROMs), structured clinical information, and language-based synthesis within the proposed architecture. They are not intended as formal clinical validation studies or quantitative performance evaluations.

Scenario 1: Multimodal contextualization in chronic low back pain

Consider a 45-year-old office worker presenting with chronic low back pain. Imaging demonstrates moderate L4–L5 disk degeneration and mild central canal stenosis. In a conventional imaging-centered workflow, automated grading systems may provide structured pathology descriptors that the radiologist must manually interpret and contextualize within the broader clinical picture. This process can introduce variability in reporting and may require clinicians to reconcile imaging findings with functional and psychosocial information distributed across separate clinical systems.

Within one possible implementation of the proposed framework, imaging findings, PROMs, and structured clinical history are integrated

Table 2
Representative clinical use cases illustrating potential workflow enhancement following illustrative deployment of selected components of the proposed modular AI framework.

Use case	Current clinical practice	Potential AI-enabled workflow
Case 1: Chronic Low Back Pain	<i>Imaging-driven reports; limited context integration.</i> • Structured image grades require manual contextualization. • High inter-reader variability and time-intensive reporting.	<i>Context-aware multimodal synthesis.</i> • Fuses spine imaging, EHR, and PROMs for risk-graded, evidence-linked draft summaries. • Illustrative workflow objectives: reduced reporting burden, improved reporting consistency, and enhanced workflow efficiency.
<i>Illustrative evaluation metrics: time-to-report, edit frequency, report concordance</i>		
Case 2: Treatment Planning Support	<i>Structure-dominant decision pathways.</i> • Imaging severity may outweigh psychosocial context. • Limited support for holistic patient assessment and pre-treatment optimization.	<i>Risk-adjusted, evidence-grounded recommendations.</i> • Multimodal analytics highlight comorbidities and PROM-derived psychosocial risk factors. • Illustrative workflow objectives: more comprehensive patient assessment and improved multidisciplinary treatment planning.
<i>Illustrative evaluation metrics: referral appropriateness, optimization rate, clinician acceptance</i>		
Case 3: Urgent Escalation	<i>Manual escalation; inconsistent traceability.</i> • Relies on clinician vigilance for urgent routing. • Limited auditability of decisions and timing.	<i>Automated, auditable escalation support.</i> • Potential alerts and escalation pathways may support timely review of high-priority findings. • Provenance tracking records AI outputs, clinician review, and subsequent actions. • Illustrative workflow objectives: improved escalation workflows, enhanced safety, and greater accountability.
<i>Illustrative evaluation metrics: time-to-escalation, false-alarm rate, provenance completeness</i>		

into a unified multimodal summary. In this illustrative scenario, the combined profile suggests a lower-risk presentation favoring conservative management strategies such as rehabilitation-focused therapy and functional monitoring. The language-generation layer may then generate a clinician-reviewable draft summary incorporating both imaging findings and contextual patient information. The radiologist remains responsible for reviewing imaging overlays, editing generated content where necessary, and approving the final report.

This example illustrates how multimodal contextualization may facilitate more holistic interpretation by integrating structural imaging findings with functional and psychosocial information, thereby supporting clinician-centered decision-making beyond conventional imaging-only workflows.

Scenario 2: Risk-aware surgical decision support

A second illustrative case involves a 62-year-old manual laborer presenting with acute-on-chronic low back pain. MRI demonstrates multilevel degeneration and Modic type II changes. In conventional structure-dominant workflows, severe imaging findings may contribute to early surgical referral consideration.

Within a multimodal workflow, however, imaging findings are interpreted alongside disability scores, coping-related assessments, psychological quality-of-life indicators, and metabolic comorbidities such as obesity and diabetes. In this scenario, the integrated patient profile highlights both structural disease burden and potentially modifiable psychosocial or functional risk factors relevant to treatment planning.

The clinician-reviewable AI-generated summary may assist clinicians by surfacing contextualized management considerations derived from retrieved clinical guidance, including rehabilitation strategies, metabolic optimization, psychosocial support, or staged multidisciplinary evaluation prior to definitive intervention. Final treatment decisions remain entirely clinician-directed.

This scenario illustrates how multimodal contextualization may support more individualized treatment planning by integrating structural imaging findings with patient-reported outcomes, psychosocial factors, and relevant clinical comorbidities.

Scenario 3: Safety-critical escalation workflow

A third scenario illustrates a potential urgent-care pathway. A 38-year-old patient presents with acute bilateral weakness and severe radicular pain. In this context, rapid imaging analysis identifies findings suspicious for a compressive lesion requiring urgent specialist review.

Within the proposed framework, the workflow may support prioritization of the case for expedited clinician review, with AI-generated summaries routed through clinician-facing alert workflows. The language-synthesis layer may generate a concise draft escalation summary referencing relevant institutional or guideline-based recommendations retrieved through the RAG pipeline. Example outputs may include draft neurosurgical referral documentation, urgent imaging confirmation recommendations, or structured escalation templates.

All generated outputs remain subject to clinician verification and approval. In addition, the framework conceptually supports audit logging of clinician interactions, timestamps, edits, and escalation decisions to facilitate traceability and quality-review processes.

This example demonstrates how the proposed architecture could support time-sensitive clinical workflows while preserving clinician oversight, auditability, and procedural traceability throughout the escalation pathway.

Potential evaluation endpoints

Although these scenarios are illustrative rather than evaluative, they suggest several measurable endpoints that could be examined in future implementation and validation studies. Examples include reporting turnaround time, clinician edit frequency, report concordance, diagnostic agreement, referral appropriateness, treatment escalation patterns, time-to-escalation in urgent workflows, clinician acceptance, usability metrics, calibration stability, provenance completeness, and workflow safety indicators.

Together, these examples are intended to operationalize selected concepts of the proposed framework and motivate future prospective evaluation in real-world clinical environments.

Table 2 summarizes the representative workflow differences illustrated across the three scenarios.

5. Discussion

This work presents a modular and clinically oriented framework intended to support future integration of multimodal AI systems within spine-care workflows. Rather than describing a fully deployed clinical platform, the study outlines a translational architectural model combining multimodal data integration, clinician-supervised AI assistance, interoperability mechanisms, and evidence-aware language synthesis within a unified workflow structure.

5.1. Principal contributions

The proposed framework is intended to contribute to the emerging field of clinically integrated spine-care AI in five primary ways.

First, it proposes a multimodal integration paradigm that treats imaging findings, structured clinical variables, and patient-reported outcome measures (PROMs) as complementary sources of information rather than isolated analytical inputs. This design reflects the multifactorial and biopsychosocial nature of low back pain.

Second, the framework incorporates a retrieval-aware language synthesis layer intended to support structured report drafting and contextualized clinical summaries grounded in retrievable evidence sources rather than unrestricted free-text generation. This design is motivated by current concerns regarding hallucination, traceability, and factual consistency in clinical generative AI systems.

Third, the proposed architecture embeds human-in-the-loop validation throughout the workflow. AI-generated outputs remain subject to clinician review, correction, and approval, while structured feedback mechanisms support future monitoring and controlled refinement processes.

Fourth, the framework emphasizes interoperability and deployment governance through modular middleware design, standardized interfaces, auditability, staged deployment concepts, and traceable data-flow management.

Finally, the study outlines a conceptual evaluation and monitoring strategy spanning technical validation, workflow-oriented assessment, prospective deployment monitoring, and post-deployment governance considerations. Proposed evaluation dimensions include reporting efficiency, clinician edit patterns, calibration stability, referral appropriateness, fairness monitoring, and provenance completeness.

Taken together, the principal contribution of this work lies not in proposing another disease-specific predictive model, but in presenting a clinically oriented, modular framework that explicitly integrates multimodal data fusion, clinician-supervised AI, retrieval-augmented language synthesis, interoperability, governance, and explainability within a unified translational architecture for spine care.

5.2. Illustrative feasibility demonstration

To explore the practical feasibility of the proposed workflow, an early proof-of-concept integration study was conducted using retrospective real-world data collected at [X-ray Service s.r.l.](#), a specialized spine-care clinic in Brescia, Italy. The exploratory prototype implementation was intended to demonstrate the practical feasibility of integrating explainable imaging AI, PROM-informed contextualization, editable reporting, and clinician-supervised workflow orchestration within a unified proof-of-concept environment.

In this preliminary workflow-oriented assessment, two experienced radiologists independently reviewed conventional imaging interpretations and subsequently evaluated AI-assisted reporting outputs incorporating explainable imaging overlays and editable structured summaries. The exploratory prototype observations illustrated how AI-assisted reporting, clinician review, and PROM-contextualized information may support reporting workflows and treatment-oriented reasoning. Observed reporting-time differences and qualitative concordance were

exploratory workflow observations and should not be interpreted as evidence of generalized workflow efficiency or clinical effectiveness.

Importantly, this exploratory feasibility demonstration was not designed to establish diagnostic accuracy, workflow superiority, or clinical effectiveness. Rather, it illustrates how the proposed architectural principles may be operationalized within a prototype implementation while motivating future prospective technical and clinical validation. Instead, the objective was to illustrate how multimodal integration, clinician-supervised reporting, explainability mechanisms, and structured workflow orchestration could be operationalized within a unified prototype environment. Detailed implementation characteristics and observational findings are provided in Supplementary Material 1.

5.3. Relationship to existing work

Most existing AI applications in spine care remain focused on isolated analytical tasks such as imaging interpretation, outcome prediction, or report generation. Prior systems have demonstrated important advances in automated spinal imaging analysis, predictive modeling, and large-language-model-assisted reporting; however, these approaches are typically developed as modality-specific or workflow-isolated tools.

The proposed framework differs fundamentally from most existing multimodal clinical AI systems in that its primary contribution is architectural rather than algorithmic. Rather than optimizing a single diagnostic task, it provides an integrated organizational blueprint that combines multimodal contextualization, clinician-supervised language synthesis, interoperability, governance, explainability, and workflow orchestration within a unified translational framework.

Importantly, the framework does not prescribe a fixed implementation strategy or model architecture. Instead, it is intended to provide a flexible organizational blueprint through which future researchers and institutions may adapt imaging pipelines, multimodal fusion methods, language models, and deployment configurations according to local clinical and technical requirements.

Recent multimodal diagnostic systems have increasingly incorporated privacy-preserving learning, explainability, and heterogeneous data integration. The proposed framework builds upon these developments by explicitly addressing how these complementary technologies can be orchestrated into a clinically deployable, workflow-oriented ecosystem for spine care rather than existing as isolated technical components [26]. These developments suggest that clinically deployable AI increasingly requires not only strong predictive performance, but also transparent reasoning, secure data management, and workflow integration. The proposed framework complements these efforts by emphasizing implementation-agnostic workflow orchestration, clinician supervision, interoperability, and patient-centered integration within spine care. Rather than proposing another disease-specific predictive architecture, it provides a modular blueprint that can accommodate different imaging modalities, AI models, multimodal fusion strategies, and deployment settings while supporting future prospective validation and real-world clinical implementation.

5.4. Limitations and translational challenges

Several important limitations and translational challenges must be acknowledged.

First, the proposed framework remains conceptual and implementation-oriented rather than a clinically validated deployment framework. Although an exploratory prototype is presented, comprehensive technical validation, prospective multicentre evaluation, regulatory assessment, and clinical effectiveness studies remain future work.

Second, real-world deployment of multimodal clinical AI systems introduces substantial challenges related to data heterogeneity, missingness, institutional variability, interoperability constraints, and governance requirements. Robust deployment would require rigorous external validation, harmonization strategies, and carefully controlled implementation protocols.

Third, although retrieval-augmented generation (RAG) mechanisms may improve traceability and contextual grounding, large language models remain susceptible to hallucination, factual inconsistency, uncertainty miscalibration, and bias. Consequently, evidence-grounded generation, transparent provenance tracking, uncertainty communication, continuous monitoring, and clinician supervision remain essential prerequisites for trustworthy clinical deployment.

Finally, infrastructure and computational requirements may limit accessibility in resource-constrained settings. Federated learning strategies, lightweight deployment configurations, and staged implementation pathways may help reduce these barriers in future deployments.

5.5. Evaluation and deployment roadmap

The proposed framework suggests a staged translational pathway spanning technical development, prospective validation, clinical deployment, and post-deployment governance.

Stage 1 — Technical validation: Early studies should establish diagnostic performance, robustness, calibration, explainability, and external generalizability using retrospective multicentre datasets.

Stage 2 — Workflow validation: Subsequent evaluations should examine reporting efficiency, clinician edit behavior, usability, trust calibration, and human–AI interaction using simulated or silent-deployment environments.

Stage 3 — Prospective pilot implementation: Pilot clinical deployment should evaluate workflow integration, referral appropriateness, treatment planning, escalation pathways, and operational safety within routine clinical practice.

Stage 4 — Multicentre clinical evaluation: Large-scale prospective studies should assess diagnostic performance, patient-centered outcomes, workflow efficiency, fairness, cost-effectiveness, regulatory readiness, and long-term post-deployment monitoring before widespread clinical adoption.

5.6. Open research directions

Several research directions remain open for future investigation. These include privacy-preserving and federated learning strategies, longitudinal multimodal modeling, integration of wearable and remote-monitoring signals, domain-specific evaluation frameworks for clinical LLMs, and structured assessment of clinician acceptance and human–AI collaboration dynamics.

More broadly, future progress in clinically integrated spine-care AI will depend not only on advances in predictive modeling but also on rigorous prospective validation, interoperability, governance, clinician acceptance, regulatory compliance, and safe integration into routine healthcare workflows.

6. Conclusion

This work presents a modular, clinically oriented, and implementation-independent framework intended to support the future integration of multimodal AI systems into routine spine-care workflows. Rather than describing a fully implemented clinical platform, the study proposes a translational architectural model that combines imaging analysis, patient-reported outcomes, structured clinical information, and clinician-supervised language-based synthesis within a unified and interoperable workflow structure.

The proposed framework emphasizes several core principles relevant to clinically integrated AI deployment, including multimodal contextualization, human-in-the-loop validation, interoperability with clinical infrastructure, explainability, auditability, and staged governance-aware implementation. By intentionally remaining implementation-agnostic, the architecture is designed to support adaptation across different imaging modalities, institutional infrastructures, multimodal fusion strategies, and deployment environments.

An accompanying exploratory prototype further illustrates how selected components of the proposed framework may be operationalized within a clinician-supervised workflow, while deliberately avoiding claims of diagnostic or clinical validation. Although preliminary and limited in scope, this illustrative implementation provides an initial translational example intended to motivate future technical and clinical evaluation.

Several important challenges remain before clinically integrated multimodal AI systems can be safely deployed, including prospective multicentre validation, interoperability standardization, workflow evaluation, fairness assessment, regulatory approval, and long-term governance. Consequently, the framework should be viewed primarily as a structured foundation for future research, implementation, and evaluation rather than a finalized clinical solution.

More broadly, this work highlights a transition from isolated AI algorithms toward clinically integrated, multimodal, and workflow-oriented AI ecosystems for spine care. Future progress will depend not only on advances in predictive modeling, but also on rigorous prospective validation, transparent governance, interoperability, clinician-centered design, and sustained collaboration among clinicians, engineers, healthcare organizations, patients, and regulatory authorities.

CRedit authorship contribution statement

Alemu Sisay Nigru: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Sergio Benini:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yao Wang:** Writing – review & editing, Supervision, Formal analysis, Conceptualization. **Federico Maffezzoni:** Writing – review & editing, Writing – original draft, Visualization, Resources, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Matteo Bonetti:** Writing – review & editing, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Michele Frigerio:** Writing – review & editing, Resources, Methodology, Formal analysis, Conceptualization. **Graziella Bragaglio:** Writing – review & editing, Supervision, Resources, Project administration, Formal analysis, Conceptualization. **Serena Miglio:** Writing – review & editing, Validation, Resources, Formal analysis, Conceptualization. **Riccardo Leonardi:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Formal analysis, Data curation, Conceptualization.

Ethical statement

This study was conducted within the doctoral research project “*Protocolli di AI volti all’analisi delle necessità dei pazienti afferenti a percorsi sanitari e allo sviluppo di percorsi terapeutici smart*”. All procedures were performed in accordance with institutional guidelines, applicable national regulations, and the European Union General Data Protection Regulation (GDPR).

The research underwent a formal Data Protection Impact Assessment (DPIA) in accordance with Article 35 of Regulation (EU) 2016/679 (GDPR) and received institutional approval (DPIA: *Documento di Valutazione di Impatto*, approved on 01/08/2025).

All patient data were de-identified prior to analysis, and informed consent for the use of clinical, imaging, and patient-reported data was obtained through institutional data-collection procedures. Data handling complied with GDPR principles of privacy protection, security, and data minimization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge X-Ray Service s.r.l. clinic and its staff for their invaluable contributions to this work. We thank the clinical team for generously sharing their expertise and insights, and for their contributions to the development and evaluation of the prototype components presented in this study.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.imu.2026.101802>.

References

- [1] Yang Xuanzhe, Zhang Yuming, Li Yi, Wu Zixiang. Performance of artificial intelligence in diagnosing lumbar spinal stenosis: A systematic review and meta-analysis. *Spine* 2024;49(23):1623–32. <http://dx.doi.org/10.1097/BRS.0000000000005174>.
- [2] Chen Ningjing, Fong Daniel Yee Tak, Wong Janet Yuen Ha. The global health and economic impact of low-back pain attributable to occupational ergonomic factors in the working-age population by age, sex, geography in 2019. *Scand J Work Environ Health* 2023;49(7):487–95. <http://dx.doi.org/10.5271/sjweh.4116>.
- [3] Fatoye Francis, Gebrye Tadesse, Ryan Cormac G, Useh Ushotane, Mbada Chidozie. Global and regional estimates of clinical and economic burden of low back pain in high-income countries: A systematic review and meta-analysis. *Front Public Health* 2023;11:1098100. <http://dx.doi.org/10.3389/fpubh.2023.1098100>.
- [4] Knezevic Nebojsa Nick, Candido Kenneth D, Vlaeyen Johan W S, Zundert Jan Van, Cohen Steven P. Low back pain. *Lancet* 2021;398(10294):78–92. [http://dx.doi.org/10.1016/s0140-6736\(21\)00733-9](http://dx.doi.org/10.1016/s0140-6736(21)00733-9).
- [5] Cholewicki Jacek, Pathak Pramod, Reeves N Peter, Popovich John M. Simulation of multifactorial causes of low back pain. *Spine J* 2016;16(10):S277. <http://dx.doi.org/10.1016/j.spinee.2016.07.192>.
- [6] Carragee Eugene J, Alamin Todd F, Miller Jonathan L, Carragee John M. Discographic, MRI and psychosocial determinants of low back pain disability and remission: A prospective study in subjects with benign persistent back pain. *Spine J* 2005;5(1):24–35. <http://dx.doi.org/10.1016/j.spinee.2004.05.250>.
- [7] McNee Philip, Shambrook James, Harris E Clare, Kim Miranda, Sampson Madeleine, Palmer Keith T, et al. Predictors of long-term pain and disability in patients with low back pain investigated by magnetic resonance imaging: A longitudinal study. *BMC Musculoskelet Disord* 2011;12(1):234. <http://dx.doi.org/10.1186/1471-2474-12-234>.
- [8] George Steven Z, Lentz Trevor A, Beneciuk Jason M, Bhavsar Nrupen A, Mundt Jennifer M, Boissoneault Jeff. Framework for improving outcome prediction for acute to chronic low back pain transitions. *PAIN Rep* 2020;5(2):e809. <http://dx.doi.org/10.1097/pr.0000000000000809>.
- [9] Nicholas Michael K, Linton Steven J, Watson Paul J, Main Chris J. Early identification and management of psychological risk factors ("yellow flags") in patients with low back pain: A reappraisal. *Phys Ther* 2011;91(5):737–53. <http://dx.doi.org/10.2522/ptj.20100224>.
- [10] Beighley Adam, Zhang Allen, Huang Brendan, Carr Christopher, Mathkour Mansour, Werner Cassidy, et al. Patient-reported outcome measures in spine surgery: A systematic review. *J Craniovertebral Junction Spine* 2022;13(4):378–89. <http://dx.doi.org/10.4103/jcvjs.jcvjs.101.22>.
- [11] Bydon Mohamad, Dominari Asimina. Patient-reported outcomes in spine surgery: The wealth of data. *Neurosurgery* 2025;96(3S):S148–53. <http://dx.doi.org/10.1227/neu.0000000000003333>.
- [12] Holmes Michelle Marie. Reconceptualising patient-reported outcome measures as active components of specialist musculoskeletal care for back pain (Ph.D. thesis), University of Southampton; 2022, URL <http://eprints.soton.ac.uk/id/eprint/468587>.
- [13] Subramanian Bargava, Kumarasami Naveen, Shastry Praveen, Sri-padraj Raghotham, Sivasailam Kalyan, D Anandakumar, et al. AI-driven MRI spine pathology detection: A comprehensive deep learning approach for automated diagnosis in diverse clinical settings. 2025, <http://dx.doi.org/10.48550/arXiv.2503.20316>, URL <https://arxiv.org/abs/2503.20316>. arXiv preprint Preprint.
- [14] Shastry Praveen, Sonawane Bhawana, Mohan Kavya, Kumarasami Naveen, Sri-padraj Raghotham, D Anandakumar, et al. AI and deep learning for automated segmentation and quantitative measurement of spinal structures in MRI. 2025, <http://dx.doi.org/10.48550/arXiv.2503.11281>, URL <https://arxiv.org/abs/2503.11281>. arXiv preprint arXiv:2503.11281 [Preprint].
- [15] Tumko Vladislav, Kim Jack, Uspenskaia Natalia, Honig Shaun, Abel Frederik, Lebl Darren R, et al. A neural network model for detection and classification of lumbar spinal stenosis on MRI. *Eur Spine J* 2024;33(3):941–8. <http://dx.doi.org/10.1007/s00586-023-08089-2>.
- [16] Windsor Rhydan, Jamaludin Amir, Kadir Timor, Zisserman Andrew. Automated detection, labelling and radiological grading of clinical spinal MRIs. *Sci Rep* 2024;14(1):14993. <http://dx.doi.org/10.1038/s41598-024-64580-w>.
- [17] Huang Jia-Wei, Zhang Yu-Long, Li Kai-Yu, Li Hua-Lin, Ye Hao-Bo, Chen Yu-Han, et al. Deep learning-based lightweight model for automated lumbar foraminal stenosis classification: Sagittal CT diagnostic performance compared to clinical subspecialists. *Eur Spine J* 2026;35(2):583–90. <http://dx.doi.org/10.1007/s00586-025-09281-2>.
- [18] Constant Caroline, Aubin Carl-Eric, Kremers Hilal Maradit, Garcia Diana V Vera, Wyles Cody C, Rouzrokh Pouria, et al. The use of deep learning in medical imaging to improve spine care: A scoping review of current literature and clinical applications. *North Am Spine Soc J (NASSJ)* 2023;15:100236. <http://dx.doi.org/10.1016/j.nxspj.2023.100236>.
- [19] Chang Yu-Cherng, Del Toro Cinthia, Gjollaj Joseph P, Braga Thiago A, Subhawong Ty K. Can artificial intelligence in spine imaging affect current practice? practical developments and their clinical status. *North Am Spine Soc J (NASSJ)* 2025;23:100621. <http://dx.doi.org/10.1016/j.nxspj.2025.100621>.
- [20] Samek Wojciech, Montavon Grégoire, Vedaldi Andrea, Hansen Lars Kai, Müller Klaus-Robert. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, vol. 11700, Springer Nature; 2019, <http://dx.doi.org/10.1007/978-3-030-28954-6>.
- [21] Amann Julia, Blasimme Alessandro, Vayena Effy, Frey Dietmar, Madai Vince I, the Precise4Q consortium. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med Inform Decis Mak* 2020;20(1):310. <http://dx.doi.org/10.1186/s12911-020-01332-6>.
- [22] Rezaeian Olya, Bayrak Alparslan Emrah, Asan Onur. Explainability and AI confidence in clinical decision support systems: Effects on trust, diagnostic performance, and cognitive load in breast cancer care. 2025, <http://dx.doi.org/10.48550/arXiv.2501.16693>, URL <https://arxiv.org/abs/2501.16693>. arXiv preprint arXiv:2501.16693 [Preprint].
- [23] Kalidindi Sadhana, Prasanna Vadhana RV. Overview of the potentially transforming role of large language models (LLMs) in creating patient-friendly summaries for radiology reports. *J Radiol Clin Imaging* 2024;7(4). <http://dx.doi.org/10.26502/jrci.2809097>.
- [24] Park Jiwoo, Oh Kangrok, Han Kyunghwa, Lee Young Han. Patient-centered radiology reports with generative artificial intelligence: Adding value to radiology reporting. *Sci Rep* 2024;14(1):13218. <http://dx.doi.org/10.1038/s41598-024-63824-z>.
- [25] Kao Jun-Ping, Kao Huan-Tang. Large language models in radiology: A technical and clinical perspective. *Eur J Radiol Artif Intell* 2025;2:100021. <http://dx.doi.org/10.1016/j.ejrai.2025.100021>.
- [26] Parwez Khadija, Sohail Syed Irfan, Akram Arslan, Rashid Javed, Atteia Ghada, Sarwar Nadeem. Explainable federated transformer framework for joint leukemia classification and stage prediction. *Sci Rep* 2026;16(1):4493. <http://dx.doi.org/10.1038/s41598-025-34715-8>.
- [27] Chen Rong, Zhang Siyun, Zheng Yiyi, Yu Qiuhua, Wang Chuhuai. Enhancing treatment decision-making for low back pain: A novel framework integrating large language models with retrieval-augmented generation technology. *Front Med* 2025;12:1599241. <http://dx.doi.org/10.3389/fmed.2025.1599241>.
- [28] Mullappilly Sahal Shaji, Kurpath Mohammed Irfan, Pieri Sara, Alseier Saeed Yahya, Cholakkal Shanavas, Aldahmani Khaled M, et al. BiMediX2: Bio-medical expert LMM for diverse medical modalities. In: Findings of the association for computational linguistics: EMNLP 2025. 2025, p. 14051–71. <http://dx.doi.org/10.18653/v1/2025.findings-emnlp.756>.
- [29] Hasan Md Mehedi, Hossain Md Abir, Sayem Farman Hossain, Paul Bikash Kumar, Rahman Ziaur, Uddin Mohammad Shorif, et al. CLIN-LLM: A safety-constrained hybrid framework for clinical diagnosis and treatment generation. 2025, <http://dx.doi.org/10.48550/arXiv.2510.22609>, URL <https://arxiv.org/abs/2510.22609>. arXiv preprint arXiv:2510.22609 [Preprint].
- [30] Yi Ziruo, Xiao Ting, Albert Mark V. A multimodal multi-agent framework for radiology report generation. 2025, <http://dx.doi.org/10.48550/arXiv.2505.09787>, URL <https://arxiv.org/abs/2505.09787>. arXiv preprint arXiv:2505.09787 [Preprint].

- [31] Kumar Sushant, Datta Sumit, Singh Vishakha, Datta Deepanwita, Kumar Singh Sanjay, Sharma Ritesh. Applications, challenges, and future directions of human-in-the-loop learning. *IEEE Access* 2024;12:75735–60. <http://dx.doi.org/10.1109/access.2024.3401547>.
- [32] Drori Iddo, Te'eni Dov. Human-in-the-loop AI reviewing: Feasibility, opportunities, and risks. *J Assoc Inf Syst* 2024;25(1):98–109. <http://dx.doi.org/10.17705/1jais.00867>.
- [33] Quinones Christian, Kumbhare Deepak, Guthikonda Bharat, Hoang Stanley. Scoping review of machine learning and patient-reported outcomes in spine surgery. *Bioengineering* 2025;12(2):125. <http://dx.doi.org/10.3390/bioengineering12020125>.
- [34] Lu Jen-Tang, Pedemonte Stefano, Bizzo Bernardo, Doyle Sean, Andriole Katherine P, Michalski Mark H, et al. Deep spine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. In: *Machine learning for healthcare conference. Proceedings of Machine Learning Research*; 2018, p. 403–19, URL <https://proceedings.mlr.press/v85/lu18a.html>.
- [35] Ziegeler Katharina, Kreutzinger Virginie, Tong Michelle W, Chin Cynthia T, Bahroos Emma, Wu Po-Hung, et al. Information extraction from lumbar spine MRI radiology reports using GPT4: Accuracy and benchmarking against research-grade comprehensive scoring. *Diagnostics* 2025;15(7):930. <http://dx.doi.org/10.3390/diagnostics15070930>.
- [36] Lee Aric, Ong Wilson, Makmur Andrew, Ting Yong Han, Tan Wei Chuan, Lim Shi Wei Desmond, et al. Applications of artificial intelligence and machine learning in spine MRI. *Bioengineering* 2024;11(9):894. <http://dx.doi.org/10.3390/bioengineering11090894>.
- [37] Zanardo Moreno, Albano Domenico, Molinari Valentina, Fabrizio Renato, Conca Martina, Asmundo Luigi, et al. Can AI write reports like a radiologist? A blinded evaluation of large language model-generated lumbar spine MRI reports. *Eur Radiol Exp* 2026;10(1):16. <http://dx.doi.org/10.1186/s41747-026-00682-6>.
- [38] Arun Nishanth, Gaw Nathan, Singh Praveer, Chang Ken, Aggarwal Mehak, Chen Bryan, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol: Artif Intell* 2021;3(6):e200267. <http://dx.doi.org/10.1148/ryai.2021200267>.
- [39] Alam Kazi Nabiul, Zadeh Pooneh Bagheri, Sheikh-Akbari Akbar. Attribution-based explainability in medical imaging: A critical review on explainable computer vision (x-CV) techniques and their applications in medical AI. *Electronics* 2025;14(15):3024. <http://dx.doi.org/10.3390/electronics14153024>.
- [40] Francis Jiztom Kavalakkatt, Darr Matthew J. Interpretable AI for time-series: Multi-model heatmap fusion with global attention and NLP-generated explanations. 2025, <http://dx.doi.org/10.48550/arXiv.2507.00234>, URL <https://arxiv.org/abs/2507.00234>. arXiv preprint [arXiv:2507.00234](https://arxiv.org/abs/2507.00234) [Preprint].
- [41] Gefer Warren B, Prokop Mathias, Seo Joon Beom, Raoof Suhail, Langlotz Curtis P, Hatabu Hiroto. Human-AI symbiosis: A path forward to improve chest radiography and the role of radiologists in patient care. *Radiology* 2024;310(1):e232778. <http://dx.doi.org/10.1148/radiol.232778>.
- [42] Herington Jonathan, McCradden Melissa D, Creel Kathleen, Boellaard Ronald, Jones Elizabeth C, Jha Abhinav K, et al. Ethical considerations for artificial intelligence in medical imaging: Deployment and governance. *J Nucl Med* 2023;64(10):1509–15. <http://dx.doi.org/10.2967/jnumed.123.266110>.
- [43] Maadi Mansoureh, Akbarzadeh Khorshidi Hadi, Aickelin Uwe. A review on human-AI interaction in machine learning and insights for medical applications. *Int J Environ Res Public Health* 2021;18(4):2121. <http://dx.doi.org/10.3390/ijerph18042121>.
- [44] Coppola Francesca, Faggioni Lorenzo, Gabelloni Michela, De Vietro Fabrizio, Mendola Vincenzo, Cattabriga Arrigo, et al. Human, all too human? An all-around appraisal of the "artificial intelligence revolution" in medical imaging. *Front Psychol* 2021;12:710982. <http://dx.doi.org/10.3389/fpsyg.2021.710982>.
- [45] Mudgal Keshav Shree, Das Neelanjan. The ethical adoption of artificial intelligence in radiology. *BJR|Open* 2020;2(1):20190020. <http://dx.doi.org/10.1259/bjro.20190020>.
- [46] Kovacs Mark D, Mesterhazy Joseph, Avrin David, Urbania Thomas, Mongan John. Correlate: A PACS- and EHR-integrated tool leveraging natural language processing to provide automated clinical follow-up. *RadioGraphics* 2017;37(5):1451–60. <http://dx.doi.org/10.1148/rg.2017160195>.
- [47] Skevington SM, Lotfy M, O'Connell KA. The world health organization's WHOQOL-BREF quality of life assessment: Psychometric properties and results of the international field trial. A report from the WHOQOL group. *Qual Life Res* 2004;13(2):299–310. <http://dx.doi.org/10.1023/b:qure.0000018486.91360.00>.
- [48] Padua R, Padua L, Ceccarelli E, Romanini E, Zanolì G, Bondì R, et al. Italian version of the roland disability questionnaire, specific for low back pain: Cross-cultural adaptation and validation. *Eur Spine J* 2002;11(2):126–9. <http://dx.doi.org/10.1007/s005860100262>.
- [49] Sica Claudio, Magni Cristina, Ghisi Marta, Altoè Gianmardo, Sighinolfi Cecilia, Chiri Luigi Rocco, et al. Coping orientation to problems experienced-nuova versione Italiana (COPE-NVI): Uno strumento per la misura degli stili di coping. *Psicoter Cogn E Comport*. 2008;14(1):27, URL <https://hdl.handle.net/11577/2266717>.
- [50] Hartrick Craig T, Kovan Juliann P, Shapiro Sharon. The numeric rating scale for clinical pain measurement: A ratio measure? *Pain Pr* 2003;3(4):310–6. <http://dx.doi.org/10.1111/j.1530-7085.2003.03034.x>.
- [51] Lewis Patrick, Perez Ethan, Piktus Aleksandra, Petroni Fabio, Karpukhin Vladimir, Goyal Naman, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Proceedings of the 34th international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc.; 2020, p. 793. <http://dx.doi.org/10.5555/3495724.3496517>.