



**Università
di Brescia**

Department of Clinical and Experimental Sciences

**PhD in: Artificial Intelligence in Medicine and Innovation in Clinical
Research and Methodology**

(Artificial Intelligence in Medicine)

IINF-03/A Telecommunications

scientific and disciplinary sector

CYCLE: XXXVIII

**AI tools for enhanced follow-up of patients with
low-back pain**

Alemu Sisay Nigru

**Tutor: Prof. Riccardo Leonardi
Co-supervisor: Prof. Sergio Benini**

May, 2026
Brescia, Italy

*Dedicated to **Emahoy Mulu Fentabil**, the foundation of my journey and the
reason for who and where I am today. Rest in peace!*

Acknowledgements

This journey would not have been possible without the grace of Almighty God and the intercession of His mother, the Virgin Mary.

My deepest admiration and gratitude goes to my PhD supervisor, Prof. Riccardo Leonardi, for his broad vision, kind guidance, and for the freedom and trust he placed in me throughout this PhD.

I am sincerely grateful to Prof. Sergio Benini, my co-supervisor and mentor. He supported me through every step of this journey, always making time for meetings, thoughtful discussions, and on-point feedbacks.

My heartfelt thanks go to X-Ray Service s.r.l. and Dott.ssa Graziella Bragaglio for funding my PhD and granting access to the invaluable data used in this research. I also thank all the PO staff for their hospitality and administrative support during the data collection phase. I extend my deep respect and gratitude to Dott. Matteo Bonetti and Dott. Michele Frigerio for their expert radiological annotations and clinical insights, which provided the foundational ground truth for this work. I am especially grateful to Dott. Federico Maffezzoni, my immediate advisor and collaborator, for our productive partnership, for his insightful psychological interpretation of the data, and for fostering such a positive and welcoming working environment.

I extend my sincere gratitude to Prof. Binyam C. Tilahun, Director of the Center for Digital Health and Implementation Science (CDHI) in Ethiopia, and to all the CDHI staff for hosting my three-month research stay, for their exceptional hospitality, and for the enriching experience I shared with them.

I also thank Prof. Yao Wang and the members of the Video Lab at New York University (NYU) for hosting my six-month research stay and for welcoming me into an open, inspiring, and diverse research environment.

This journey would have been far more difficult without the constant support of my wife, Betelehem Dagnaw, my technical advisor, closest friend, and strongest supporter. Her patience, encouragement, and positivity, even during long periods of distance, allowed me to focus fully on my work. This dissertation is, in many ways, the fruit of her love and patience.

Finally to all my family, my friends, and to the entire UNIBS & Brescia Habesha community: thank you for your unconditional support, encouragement, and for all the moments we shared along the way.

Abstract

Chronic low back pain (LBP) is a leading global cause of disability. Its effective management is complicated by a persistent clinical gap in which objective imaging findings are often weakly aligned with patients' subjective experiences, limiting the delivery of personalized, biopsychosocial care. This dissertation addresses this challenge by proposing and systematically evaluating an integrative artificial intelligence (AI) framework designed to combine spine imaging analysis with patient-reported outcome measures (PROMs) into clinically meaningful decision support. The research makes five primary contributions. First, it introduces a modular architectural framework for multimodal integration that emphasizes explainability, clinician oversight, and human-in-the-loop governance. A proof-of-concept evaluation involving radiologists suggests that such integration can improve workflow efficiency and inform treatment prioritization in selected clinical scenarios. Second, the dissertation advances automated imaging analysis through LumbarAI, a multi-view transformer-based model incorporating curriculum learning for the grading of eleven lumbar spinal pathologies. Third, it establishes a methodological approach for patient phenotyping using unsupervised clustering to derive clinically interpretable patient groups from combined imaging and psychosocial data. The resulting patterns provide empirical support for the biopsychosocial model by revealing meaningful heterogeneity in pain-related presentation that is not fully explained by structural MRI findings alone. Fourth, this analysis highlights coping-related psychosocial factors, particularly Avoidance and Positive Attitude, as influential maladaptive and adaptive coping strategies associated with patient stratification. Fifth, to facilitate clinical translation, the dissertation develops a parsimonious predictive model aimed at reducing assessment burden while preserving discriminative capability. Building on the full variable clustering framework, a simplified four-variable model was derived based on statistical relevance, clinical interpretability, and feasibility. This model demonstrated strong predictive performance and enabled the formulation of transparent decision rules and a risk-scoring scheme suitable for early screening and stratified intervention.

By progressing from critical validation of existing imaging technologies, through the development of an improved core imaging model, to the principled integration of psychosocial data, this dissertation offers both a technical foundation and a clinical roadmap. It demonstrates how AI can support the integration of structural, functional, and psychosocial information while preserving clinician agency, contributing to more holistic, evidence-informed, and patient-centered spine care.

Abstract [Italian]

La lombalgia cronica è una delle principali cause di disabilità a livello globale. La sua gestione è complicata da un persistente divario clinico, in cui i risultati oggettivi delle immagini diagnostiche risultano spesso debolmente allineati con l'esperienza soggettiva dei pazienti, limitando l'erogazione di cure personalizzate di tipo biopsicosociale. Questa tesi affronta tale sfida proponendo e valutando sistematicamente un quadro integrativo di intelligenza artificiale (IA) progettato per combinare l'analisi delle immagini della colonna vertebrale con le misure degli esiti riferiti dai pazienti in un supporto decisionale clinicamente significativo. La ricerca apporta cinque contributi principali. In primo luogo, introduce un quadro architetturale modulare per l'integrazione multimodale, che enfatizza la spiegabilità, la supervisione clinica e la governance human-in-the-loop. Una valutazione proof-of-concept con radiologi suggerisce potenziali miglioramenti nell'efficienza del flusso di lavoro e nella prioritizzazione terapeutica in specifici scenari clinici. In secondo luogo, la tesi avanza l'analisi automatizzata delle immagini attraverso LumbarAI, un modello multi-vista basato su architetture transformer per la valutazione di undici patologie della colonna lombare. In terzo luogo, viene definito un approccio metodologico per la fenotipizzazione dei pazienti mediante clustering non supervisionato, finalizzato all'identificazione di gruppi clinicamente interpretabili a partire da dati di imaging e variabili psicosociali. I risultati forniscono supporto empirico al modello biopsicosociale, evidenziando una variabilità clinica non pienamente spiegata dai soli reperti strutturali della risonanza magnetica. In quarto luogo, l'analisi evidenzia il ruolo di fattori psicosociali legati alle strategie di coping, in particolare l'evitamento e l'atteggiamento positivo, quali dimensioni disadattive e adattive associate alla stratificazione dei pazienti. Infine, per favorire la traduzione clinica, la tesi sviluppa un modello predittivo semplificato volto a ridurre il carico valutativo mantenendo una solida capacità discriminatoria. A partire dal modello multimodale completo, è stato derivato un modello a quattro variabili, selezionate in base alla rilevanza statistica, all'interpretabilità clinica e alla fattibilità, che consente la definizione di regole decisionali trasparenti e di uno schema di valutazione del rischio per lo screening precoce e l'intervento stratificato.

Nel complesso, questa tesi fornisce una base tecnica e una roadmap clinicamente fondata, dimostrando come l'intelligenza artificiale possa supportare l'integrazione di informazioni strutturali, funzionali e psicosociali, preservando l'autonomia decisionale del clinico e contribuendo a una cura della lombalgia più olistica, basata su evidenze scientifiche e incentrata sul paziente.

Contents

1	Introduction	1
1.1	The Global Burden of Low Back Pain	1
1.2	The Clinical Challenge: A Siloed Paradigm in a Multifactorial Reality	3
1.2.1	The Structural Silo	3
1.2.2	The Psychosocial Silo	4
1.2.3	The Integration Gap	4
1.3	An Integration Opportunity	5
1.3.1	Standardizing and Quantifying the Imaging Silo with AI	5
1.3.2	The Core Challenge	6
1.3.3	From Data to Decision Support	6
1.3.4	Bridging the Gap	7
1.4	Research Problem and Knowledge Gaps	7
1.5	Research Aim and Specific Objectives	9
1.6	Scope and Delimitations of the Dissertation	10
1.6.1	Scope	10
1.6.2	Delimitations	11
1.7	Dissertation Outline	12
1.8	Summary of Contributions	13
2	Literature Review	15
2.1	Clinical Foundations of Low Back Pain	15
2.1.1	Epidemiology and Classification of LBP	15
2.1.2	Spinal Anatomy and Pathoanatomy on MRI: Foundations for Quantitative Analysis	17
2.1.3	The Biopsychosocial Model of Pain	17
2.1.4	Patient-Reported Outcome Measures (PROMs) in Spine Care	18
2.2	Technological Foundations: AI in Healthcare	19
2.2.1	Deep Learning for Medical Image Analysis	19
2.2.2	The State-of-the-Art and Limitations in Automated Spine MRI Interpretation	20

2.2.3	Machine Learning for Biopsychosocial Phenotyping and Stratification	20
2.2.4	The Integration Gap: From Disparate AI Tools to Clinical Integration	22
2.2.5	Human-in-the-Loop AI and Explainability	23
2.3	Synthesis: The Integration Imperative in AI for Spine Care	23
2.3.1	The Self-Reinforcing Cycle of Clinical Fragmentation	24
2.3.2	The Core Unmet Need: A Framework for Clinical Synthesis	25
2.3.3	Conclusion	25
3	A Clinically Integrated AI Framework for Personalized Spine Care	27
3.1	Introduction	27
3.2	Clinical Imperative and Foundational Design Principles	28
3.2.1	User-Centered Design: Defining Clinical Requirements	28
3.2.2	Synthesized Stakeholder Needs and Clinical Decision Points	29
3.2.3	Derived Design Principles	31
3.2.4	Clinical Workflow Integration: Mapping the Framework to Practice	32
3.3	System Architecture Overview	35
3.3.1	Data Input Layer	35
3.3.2	Security and Governance Layer	35
3.3.3	Interoperability and Normalization Layer	36
3.3.4	Modular AI Processing Core	36
3.3.4.1	Computer Vision Module	36
3.3.4.2	Multimodal Profiling and Synthesis Module . . .	36
3.3.4.3	LLM Synthesis & RAG Module	37
3.3.5	Clinical Application Layer	38
3.3.6	MLOps and Continuous Learning Layer	38
3.4	Use Case Scenarios	38
3.5	Discussion	41
3.5.1	Principal Contributions and Integrative Novelty	42
3.5.2	Differentiation from Existing Work	42
3.5.3	Critical Limitations and Mitigation Strategies	43
3.5.4	Conclusion and Forward Path	43
3.5.5	Open research directions	44
4	LumbarAI: Clinically-Aligned AI Model for Automated Spine MRI Analysis	45

4.1	Introduction	45
4.1.1	Radiological Features in Lumbar MRI Analysis	46
4.1.2	Research Trajectory: From External Validation to Novel Architecture	46
4.2	Part I: Benchmarking the State-of-the-Art	49
4.2.1	Rationale for Benchmarking	49
4.2.2	Methods	49
4.2.3	Agreement and Reliability Metrics	51
4.2.4	Results and Identified Gaps	53
4.2.5	Derived Requirements for the Imaging Module	54
4.3	Part II: LumbarAI – A Curriculum Learning Framework	55
4.3.1	The LumbarAI Model: Rationale and Architecture	55
4.3.1.1	Core Hypothesis and Design Philosophy	55
4.3.1.2	Model Architecture: Workflow-Aligned Hybrid Architecture	56
4.3.1.3	Swin3D Transformer Backbone	56
4.3.1.4	Multi-View Feature Fusion	56
4.3.1.5	Model Implementation and Complexity	58
4.3.2	Four-Phase Curriculum Learning Strategy	58
4.3.3	Loss Functions	59
4.3.3.1	Single-view (Sagittal-only) Loss	59
4.3.3.2	Multiview Fusion Losses	59
4.3.3.3	Curriculum and Task-weighted Loss Scheduling .	60
4.3.4	Methods	60
4.3.4.1	Datasets and Data Curation	60
4.3.4.2	Multi-View Preprocessing Pipeline	61
4.3.4.3	Experimental Setup and Evaluation Strategy . .	64
4.3.4.4	Implementation Details	65
4.3.5	Results	65
4.3.5.1	Performance of the Sagittal-Only Baseline	65
4.3.5.2	Hybrid Architecture Performance	67
4.3.6	Curriculum Learning Progression Analysis	69
4.3.7	Ablation Studies: Component Contribution Analysis	71
4.3.7.1	The Necessity of Curriculum Learning	72
4.3.7.2	Impact of Individual Components	73
4.3.7.3	Synthesis of Ablation Findings	74
4.3.8	Independent Validation with Blind Annotations: LumbarAI vs. SpineNet	75

4.3.9	Qualitative Interpretability Analysis of LumbarAI	78
4.4	Discussion	83
4.4.1	Clinical Alignment and Workflow Integration	83
4.4.2	Technical Innovations and Comparison with Existing Methods	84
4.4.3	Limitations and Methodological Considerations	86
4.4.4	Future Directions	88
5	Psychosocial Determinants and Phenotypic Stratification in Spinal Pain	89
5.1	Introduction	89
5.2	Dataset and Methods	91
5.2.1	Study Design and Setting	91
5.2.2	Ethical Considerations	91
5.2.3	Participants	91
5.2.4	Measures	92
5.2.5	Data Preprocessing	93
5.2.6	Cluster Characterization	94
5.3	Results	95
5.3.1	Coping Determinants of Pain	95
5.3.1.1	Descriptive Statistics and Bivariate Associations .	95
5.3.1.2	Uni-variate Regression Analysis	97
5.3.1.3	Multiple Regression Analysis	98
5.3.1.4	Mediation Analysis	100
5.3.2	Biopsychosocial Phenotyping via Unsupervised Clustering	104
5.3.2.1	Identification of Patient Phenotypes	105
5.3.2.2	Psychosocial and Quality-of-Life Differentiators .	109
5.3.2.3	Network-Level Characteristics	110
5.3.2.4	Clinical–Radiological Decoupling	112
5.3.2.5	Predictive Modelling of Cluster Membership . . .	114
5.3.2.6	Simplified Predictive Model	116
5.4	Discussion	120
5.4.1	Integrated Interpretation and Clinical Implications	120
5.4.2	Limitations	122
5.4.3	Methodological Considerations and Strengths	123
5.5	Conclusion	124
6	Proof-of-Concept: Feasibility and Interpretability of the Integrated Workflow	125
6.1	Introduction	125

6.2	Methods for the Integrated Pilot Study	126
6.2.1	Study Design and Participants	126
6.2.2	Study Procedures	126
6.2.3	Nature of the Manual Integration	127
6.3	Summary of Validated Core Components	127
6.4	Results	127
6.4.1	Integrated Patient Profiles	127
6.4.2	Reporting Efficiency	128
6.4.3	Impact on Treatment Planning	128
6.4.4	Radiologist Feedback and Qualitative Insights	129
6.5	Discussion	130
6.5.1	Synthesis of Evidence	130
6.5.2	Limitations and the Path to Automation	130
6.5.3	Future Directions	131
6.6	Conclusion	131
7	Summary, General Discussion, and Conclusions	133
7.1	Thesis Summary	133
7.2	General Discussion	134
7.2.1	Imaging Severity and the Limits of Structural Explanation	134
7.2.2	Coping Phenotypes as Organizing Principles	135
7.2.3	AI as Clinical Decision Support	135
7.3	Future Work	135
7.4	General Conclusion	136
A	Appendix: LumbarAI Output and Draft Report	137
	Bibliography	141

List of Figures

3.1	Multidisciplinary Patient Journey Workflow	33
3.2	System Architecture Overview	37
4.1	T2-weighted MRI examples for all grading levels of each evaluated radiological feature	47
4.2	External validation workflow for SpineNetV2	50

4.3	The LumbarAI hybrid architecture	57
4.4	Multiview Preprocessing Pipeline for Level-Specific ROI Extraction.	62
4.5	Illustration of the DICOM world coordinate system.	63
4.6	Receiver Operating Characteristic (ROC) Curves for the Hybrid Model Architecture.	69
4.7	Comprehensive Explainability Analysis of the sagittal only model	79
4.8	GradCAM Visualizations for Multiview Spine Pathology Tasks	81
4.9	LumbarAI User Interface	82
5.1	Scatterplots illustrating bivariate relationships between painNRS and each CopeNVI subscale	96
5.2	Outlier and normality assessment of pain intensity (NRS) scores	98
5.3	Synthesized mediation model depicting primary pathways among coping strategies and pain intensity.	101
5.4	Cluster validation overview.	106
5.5	Psychosocial profile of borderline Cluster 2 patients relative to cluster centroids.	107
5.6	Distribution of Adjusted Rand Index (ARI) values across 1000 bootstrap iterations.	109
5.7	PCA biplot showing patient data points	110
5.8	Network graphs for each cluster	112
5.9	Correlation heatmaps for Cluster 1 (left) and Cluster 2 (right)	113
5.10	Scatter plots of key variable relationships comparing Cluster 1 (blue) and Cluster 2 (purple).	114
5.11	Decision tree for cluster classification using the 4 simplified variables.	117
5.12	Interactive clinical screening report for a Maladaptive phenotype patient.	119

List of Tables

3.1	Representative clinical use cases illustrating potential workflow improvement after modular AI framework deployment	41
4.1	Standard radiological features for degenerative lumbar spine assessment.	48

4.2	MRI acquisition protocol for the XrayService AI-Assisted Dataset (XSA).	50
4.3	Inter-rater reliability and agreement evaluation	51
4.4	Key agreement metrics results of external validation	54
4.5	Summary of dataset sources and annotation strategies	61
4.6	Sagittal-Only Baseline Performance on the XSA, RSNA and SPI- DER test sets.	66
4.7	Multiview model performance across all test sets	67
4.8	Comparison of Multi-View and Sagittal-Only Models: Task-Averaged Performance Metrics from Tables 4.6 and 4.7.	68
4.9	Curriculum Learning Phase Performance: Accuracy and F1 Score Across Phases 1-4 for Multiview Tasks (Combined Test Set, 2516 samples)	70
4.10	Curriculum Learning Phase-to-Phase Progression	72
4.11	Ablation Study: Impact of Curriculum Learning on Model Perfor- mance (Multiview Tasks)	72
4.12	Comprehensive Ablation Study: Impact of Individual Components on Multiview Task Performance	73
4.13	Performance Comparison on Blind-Annotated External Validation Dataset (XSB Consensus Dataset)	75
5.1	Means and Standard Deviations for all CopeNVI subscales and NRS pain scores	95
5.2	Correlation matrix of pain intensity (NRS) with CopeNVI subscales	96
5.3	Regression coefficients for pain intensity (NRS) predicted by CopeNVI subscales	97
5.4	Multiple regression analysis: All CopeNVI subscales predicting pain intensity (NRS)	99
5.5	Mediation analysis: Coping strategies as mediators between other coping strategies and pain intensity	102
5.6	Comparison of clustering solutions across four algorithms.	108
5.7	Descriptive statistics and between-cluster comparisons.	111
5.8	Comparison of Structural Pathology by Psychosocial Cluster	115
6.1	Summary of dominant PROMs, clinical, and imaging findings across pilot patients.	128
6.2	Reporting time efficiency: MRI-only vs. AI-assisted editing.	128

1

Introduction

This chapter presents the clinical background and research objectives for developing AI tools in low back pain management. It identifies the research gap, outlines the study's scope and limitations, and details the specific aims and contributions of this dissertation, concluding with an overview of the document structure.

1.1 The Global Burden of Low Back Pain

Low back pain (LBP) constitutes a major and escalating global health challenge. It is a condition of exceptional prevalence: in 2020, an estimated 619 million individuals were affected worldwide, with projections suggesting a rise to 843 million by 2050, driven primarily by population growth and ageing demographics [1, 2, 3]. In terms of disability, LBP is the foremost contributor to years lived with disability (YLDs) among all musculoskeletal disorders [1, 2, 3]. While age-standardised rates have shown some decline, the absolute number of cases and the associated disability burden continue to grow, indicating that demographic shifts are offsetting epidemiological gains [4, 2]. This trend underscores a critical public health paradox: despite advancements in understanding, the population-level impact of LBP is increasing.

The burden manifests profoundly at both individual and societal levels. For patients, particularly working-age adults, LBP is a leading cause of persistent functional impairment, reduced quality of life, and lost productivity [4]. Societally, LBP imposes a substantial economic toll. It ranks among the top conditions contributing to disability-adjusted life years (DALYs) in high-income countries

[5]. Direct healthcare costs are significant, though estimates vary widely across healthcare systems and methodologies. For instance, a systematic review and meta analysis performed in 2023, focusing on high-income settings, reported mean direct costs of approximately USD 9,231 per patient [6]. Critically, indirect costs related to work absenteeism and presenteeism are frequently reported to constitute the predominant component of total societal costs, amplifying the economic impact well beyond direct medical expenditure [6]. This economic profile highlights LBP not merely as a medical issue, but as a significant socioeconomic one.

Understanding this escalating burden requires examining its multifactorial risk profile. It remains a subject of ongoing research and debate. Established demographic and biomechanical risk factors include advancing age, occupational ergonomic stressors, and elevated body mass index [4, 2]. However, the role of psychosocial factors, such as depression, anxiety, and maladaptive beliefs about pain, is increasingly recognised as equally, if not more, prognostic of chronicity and disability with respect to structural findings [7]. This complex aetiology, straddling biological and psychological domains, presents a fundamental challenge for simplistic diagnostic and treatment paradigms.

Consequently, managing LBP effectively within healthcare systems is notoriously difficult. Although international guidelines advocate a stepped-care model that prioritises education, physical activity, and non-opioid analgesics, their implementation is highly inconsistent [8, 9]. Significant disparities in access to high-quality, guideline-concordant care exist across and within countries, often influenced by socioeconomic status and geographic location [4, 10]. In some regions, this manifests as under-treatment and lack of access to effective therapies; in others, as over-reliance on imaging, invasive procedures, or opioids contrary to guideline recommendations [8]. These systemic failures in translating evidence into practice directly contribute to the cycle of disability and cost described above.

In summary, LBP represents a persistent and growing global burden characterised by high prevalence, significant disability, and major economic cost. Its multifactorial and often biopsychosocial nature complicates clinical management, while systemic inequities and inconsistent guideline adoption hinder effective care delivery. This context creates an urgent imperative for innovative approaches that can improve the precision, efficiency, and integration of spine care pathways to mitigate one of the most substantial public health challenges of our time.

1.2 The Clinical Challenge: A Siloed Paradigm in a Multifactorial Reality

The substantial global burden of LBP described in Section 1.1 stems not from a simple, unitary disease, but from a complex, multifactorial condition involving structural, inflammatory, neurological, and significant psychosocial influences [11]. Effective management therefore requires a holistic, biopsychosocial understanding. However, the standard diagnostic workflow is fundamentally fragmented, operating with distinct, non-integrated streams of information (e.g. advanced structural imaging, clinical examination, and patient-reported psychosocial data). This siloed paradigm forces clinicians to perform a complex synthesis of disparate data, contributing to diagnostic uncertainty, inconsistent treatment pathways, and the persistent challenge of aligning pathological findings with patient-reported outcomes, a key driver of the disability and cost burden previously outlined.

1.2.1 The Structural Silo: The Limited Clinical Utility of Isolated MRI Findings

Magnetic resonance imaging (MRI) is a pivotal tool for ruling out serious pathology (e.g., malignancy, infection, degeneration, herniation) and planning surgical interventions for specific structural conditions [12]. Its detailed visualization of anatomy is codified through numerous grading and classification systems, including the Pfirrmann classification for disc degeneration [13], Modic changes [14], the Lee grading system for central canal and foraminal stenosis [15], and the Combined Task Force nomenclature for disc herniation and nerve root compression [16]. However, the centrality of MRI in the diagnostic mindset often leads to an over-reliance on structural findings for explaining pain, despite clinical guidelines discouraging routine MRI for non-specific LBP [17].

This over-reliance on imaging findings is problematic for two primary, evidence-based reasons. First, the clinical specificity of common degenerative findings is low. Their prevalence in asymptomatic individuals increases with age, exceeding 90% in older populations [18]. Consequently, the correlation between the severity of a specific finding and a patient's pain intensity or functional disability is weak and inconsistent [19]. Second, the interpretation of these findings is subject to considerable inter-rater variability, affecting the reliability of assessments for conditions like spinal stenosis [20]. This variability underscores a lack of standardized, objective synthesis of MRI data, where the final clinical interpretation can depend

significantly on the reader. Thus, while MRI is indispensable for specific indications, it answers “what is there?” with limited fidelity to the critical question: “is this the primary contributor to this patient’s pain?”

1.2.2 The Psychosocial Silo: Critical yet Operationally Segregated Prognostic Data

Concurrently, a robust evidence base highlights that psychosocial factors and patient-reported outcome measures (PROMs) are key determinants of the pain experience and long-term disability. These PROMs capture the multidimensional impact of pain through standardized instruments, including pain intensity scales (e.g., the Numerical Rating Scale, Visual Analog Scale [21]), disability indices (e.g., the Roland–Morris Disability Questionnaire [22], Oswestry Disability Index [23]), quality-of-life assessments (e.g., the WHOQOL-BREF [24], SF-36 [25]), and psychological evaluations (e.g., the CopeNVI [26], Pain Catastrophizing Scale [27]). Data from these instruments, often explain more variance in clinical outcomes, such as pain chronicity and functional recovery, than the presence or severity of structural findings on MRI [28, 29]. The biopsychosocial model formally integrates these factors as central, rather than peripheral, to the pain process [30].

Despite their established prognostic value, these data are operationally segregated. In clinical practice, PROMs are often collected but their quantitative results are rarely integrated with imaging data to create a unified patient profile. This segregation persists due to practical workflow barriers, where PRO collection is often seen as a disruptive, labor-intensive task that is managed in systems separate from the core electronic health record [31]. Furthermore, a fundamental methodological gap exists: there is no established clinical framework to quantitatively integrate PROMs to MRI findings [32]. As a result, these data, though critical for a holistic patient description [33], remain epistemologically and practically distinct from imaging data, limiting their actionable impact on decision-making.

1.2.3 The Integration Gap: A Systemic Barrier to Precision Management

The contemporary clinical handling of spine care therefore presents a systemic integration challenge. The clinician must reconcile a structural report detailing abnormalities of ambiguous clinical significance with a psychosocial profile indicating levels of distress and disability that may correlate poorly with those findings. This mentally demanding integration makes clinicians vulnerable to well-documented cognitive biases. For example, they may anchor on a salient

imaging finding, allowing it to disproportionately shape their diagnosis despite contradictory evidence from the patient’s profile [34, 35].

This environment creates a well-recognized clinical dilemma: patients with similar degenerative changes on MRI can experience vastly different outcomes based on their psychosocial profile [36]. The current paradigm, lacking tools for a structured approach, often forces a choice between over-attributing symptoms to structural pathology (potentially driving low-value interventions) or under-treating significant distress when imaging appears benign. Both pathways contribute to the suboptimal outcomes and economic inefficiency characterizing the LBP burden.

The core challenge is therefore not a lack of data, but a lack of integration. The standard of care endorses the biopsychosocial model in principle but lacks the practical, quantitative methodologies to implement it consistently at the point of care. This gap between multidimensional aetiology and siloed assessment establishes a clear imperative for novel frameworks capable of fusing these heterogeneous data streams into interpretable, patient-specific profiles.

1.3 An Integration Opportunity: Toward an AI Framework for Multimodal Data Synthesis

The substantial clinical burden and the persistent diagnostic fragmentation detailed in the previous sections reveal a fundamental methodological gap: the lack of a systematic approach to integrate quantitative imaging data with patient-reported functional and psychosocial states. Artificial intelligence (AI) presents a compelling paradigm to address this gap. Rather than a single solution, the opportunity lies in developing an *integrative AI framework*, a structured computational pipeline designed to synthesize currently siloed data into unified, clinically interpretable profiles that support personalized, biopsychosocial management of LBP.

1.3.1 Standardizing and Quantifying the Imaging Silo with AI

A foundational step is applying AI to standardize the interpretation of spine MRI, which is inherently limited by inter-rater variability and protocol heterogeneity. Deep learning models, such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), can automatically extract and grade pathological features from images [37, 38]. These models have demonstrated promising performance in research settings for tasks like grading disc degeneration, end plate changes

and spinal stenosis [39, 40]. Critically, this process converts subjective visual assessments into structured, quantitative data. However, achieving robust clinical application requires overcoming significant challenges, including the need for advanced harmonization techniques to manage variability across scanners and protocols, and the inherent subjectivity of the grading labels used for model training.

1.3.2 The Core Challenge: Multimodal Data Fusion

The transformative potential of an AI framework lies in its capacity for *multimodal data fusion*, the computational integration of heterogeneous data types (imaging, PROMs, clinical variables) that currently reside in separate silos. Early research demonstrates the feasibility of this approach; for example, machine learning models have begun to identify correlative and predictive relationships between specific MRI-derived features and patient-reported disability scores (e.g., ODI) [41, 42]. These relationships often involve complex, high-dimensional interactions that are difficult for clinicians to detect using conventional methods due to cognitive limits and the practical constraints of a clinical workflow. The framework’s architecture must make key methodological decisions (e.g., early vs. late fusion) to effectively combine these distinct data modalities into a coherent representation.

1.3.3 From Data to Decision Support: Synthesizing a Holistic Profile

The ultimate goal of such a framework is to translate integrated data into actionable clinical insight. This involves moving beyond correlation to generate composite phenotypes or risk stratifications that encapsulate both structural and psychosocial contributors to a patient’s condition. Emerging tools like large language models (LLMs) hint at future possibilities for automated synthesis, such as generating preliminary reports that combine imaging findings with PROM trends [43]. However, their clinical application requires careful validation to mitigate risks like hallucination. The proposed framework would therefore function not as an autonomous system, but as a *clinical decision support tool*. In this augmentative paradigm, the AI performs the computational integration of data and presents synthesized findings, allowing the clinician to apply expert judgment within a human-AI collaborative workflow [44, 45, 46].

1.3.4 Bridging the Gap: A Framework for Integrative Spine Care

The methodological void identified at the heart of the clinical paradox creates a clear imperative: a structured, computational framework capable of fusing multimodal data into a holistic patient profile. The opportunity, therefore, is to define and develop such an integrative AI framework for spine care. An effective framework would need to operationalize the biopsychosocial model through a pipeline that addresses three core challenges: 1) standardized data acquisition and harmonization of heterogeneous inputs, 2) execution of robust multimodal fusion algorithms, and 3) the presentation of synthesized outcomes through clinically interpretable interfaces that would clearly lead to decision pathways. Successfully addressing these challenges would provide the first method to quantitatively bridge anatomical observation with patient experience, offering a concrete pathway toward more precise and holistic assessment.

1.4 Research Problem and Knowledge Gaps

The preceding analysis establishes a dual challenge: the clinical burden of LBP is compounded by a fragmented diagnostic paradigm, while emerging AI technologies offer a promising but under-realized opportunity for integration. This lack of connection persists because research outputs, often focusing on isolated technical benchmarks, fail to address the heterogeneity of clinical data and the absence of integrative standards required for practical workflow adoption. Consequently, the critical gaps preventing the adoption of AI into clinical workflows are not merely technical, but systemic.

1. The Modality-Specific Focus of Current AI Research: Most AI applications in spine care are modality-specific, advancing either image analysis or PROMs evaluation in isolation. This perpetuates the clinical silos described earlier, leaving unaddressed a fundamental challenge, namely, how to computationally reconcile structural findings with patient-reported symptoms. The gap, therefore, is not just in building better unimodal tools, but in developing frameworks for their *clinically meaningful integration*, where fusion produces insights greater than the sum of its parts.
2. A foundational barrier to any integrative system is the lack of robust, generalizable AI models for spine MRI. While deep learning has shown high accuracy in controlled studies, its real-world application is hampered by significant challenges. Models often fail to generalize across different MRI scanners, imaging

protocols, and patient populations, a problem starkly revealed in external validation studies [47]. Furthermore, many systems are limited to single-view analysis (e.g., sagittal-only), which is insufficient for accurately grading some key pathologies like foraminal stenosis that require multiplanar assessment [48]. This reliability deficit means that even the most advanced integrative frameworks lack a trustworthy, automated source of standardized imaging data, an essential cornerstone for multimodal fusion.

3. **The Lack of a Clinical Orchestration Framework for AI Synthesis:** Beyond stand-alone AI tools, a critical systems-level gap exists: the absence of a clinical framework capable of orchestrating diverse AI-derived outputs, from automated imaging reads to analyzed patient-reported outcomes, into a coherent, interpretable synthesis for the physician. Current AI products operate in isolation, creating a new form of digital silo. There is no established framework that defines how these outputs should be structurally combined, how their combined evidence should be transparently presented to support clinical reasoning, or how to incorporate necessary human oversight to review, edit, and finalize AI-generated assessments. This gap prevents the translation of parallel AI insights into a unified, personalized evidence base. The challenge is not merely algorithmic fusion, but the design of a clinical workflow that meaningfully synthesizes quantitative imaging interpretations with qualitative psychosocial and clinical data to directly inform and personalize treatment recommendations at the point of care.
4. **The Translational Chasm Between Algorithm and Clinical Workflow:** Promising models often neglect the ergonomics of clinical practice. A critical gap exists in designing how an integrative system would interface with hospital infrastructures (e.g., PACS, EHRs), manage data privacy and interoperability, and, most importantly, present synthesized findings to clinicians in a way that reduces cognitive load rather than adding to it. Without solving these translational challenges, AI systems risk remaining sophisticated academic exercises with limited real-world clinical utility.

Synthesis of the Core Research Problem

These identified gaps are not isolated but form a logical sequence of barriers preventing the evolution from AI-assisted diagnosis to AI-augmented, personalized spine care. The reliability deficit in imaging AI means we lack trustworthy automated inputs. Even with such inputs, the absence of a clinical orchestration framework means we have no principled method to synthesize these inputs with complementary psychosocial data for human clinical reasoning. Finally, the

translational chasm ensures that any research prototype remains disconnected from real-world practice. Together, these gaps define a systemic failure: the field possesses advanced tools but lacks the integrating framework to make them clinically cohesive and actionable.

Therefore, the core research problem this dissertation addresses is the design and validation of the foundational components for a clinically integrated AI framework. This framework must first establish reliable automated interpretation of spinal imaging, and second, provide a structured methodology for synthesizing these interpretations with psychosocial and clinical data into an evidence-based, transparent, and editable summary to support personalized treatment planning.

To tackle this systemic problem, the research presented in this dissertation is structured to provide a proof-of-concept for each necessary layer. It progresses from establishing reliable inputs to demonstrating a method for clinical synthesis, within a proposal for a deployable system architecture. The ultimate aim is to move beyond isolated AI tools toward a coordinated system that bridges the critical divide between multimodal data and personalized clinical decision-making.

1.5 Research Aim and Specific Objectives

Given the identified systemic knowledge gaps, the aim of this dissertation is to design, and develop foundational components of an integrative, clinically relevant AI framework that integrates automated MRI interpretation with biopsychosocial patient profiling to support personalized spine care.

This aim is pursued through the following specific objectives:

1. **To propose a conceptual framework for integrating AI into spine care environments, ensuring workflow compatibility and adherence to ethical and privacy standards.** (*Addresses the translational chasm gap*): This objective defines the system-level blueprint, specifying components, data flows, and human–AI interaction points required to operationalize AI outputs in clinical practice. Evaluation focuses on feasibility, alignment with clinical decision-making, and incorporation of human oversight for reviewing and editing AI-generated outputs.

2. **To develop and validate a reliable, multi-view AI model for automated interpretation of lumbar spine MRI.**

(Addresses the reliability deficit gap): This objective provides standardized anatomical inputs through a deep learning model that grades key pathologies (e.g., disc degeneration, central canal stenosis) from sagittal and axial MRI views, ensuring robustness across imaging variations.

3. **To develop a methodology for integrating AI-derived imaging data with patient-reported psychosocial and clinical profiles.**

(Addresses the clinical orchestration gap): This objective defines a repeatable process for combining quantitative imaging outputs (Objective 2) with multidimensional patient data (e.g., pain scores, disability indices, coping strategies). The output is a unified, interpretable patient profile (e.g., data-driven phenotype or structured evidence summary) to support personalized assessment.

By fulfilling these objectives, this dissertation will deliver the foundational proof-of-concept for a new paradigm in spine care: one where AI does not merely automate isolated tasks but provides a structured, synthesizing layer that bridges multimodal data to directly support personalized, evidence-based clinical decision-making.

1.6 Scope and Delimitations of the Dissertation

This dissertation focuses on the design, development, and proof-of-concept validation of core components of an integrative AI framework for low back pain (LBP) assessment. The scope is intentionally bounded to ensure feasibility while addressing the defined research gaps and providing a foundation for future translational work.

1.6.1 Scope

The scope includes the technical, clinical, and methodological work required to achieve the objectives outlined in Section 1.5:

- **Objective 1 (conceptual framework):** The architectural blueprint in Chapter 3 defines how AI components interoperate in a clinical environment, addressing the gap between technical development and clinical deployment.
- **Objective 2 (reliable imaging):** The Computer Vision Module (LumbarAI) is developed in Chapter 4 for automated grading of lumbar spine pathologies from MRI, addressing limitations in current imaging assessment.

- **Objective 3 (clinical integration methodology):** The Multimodal Assessment Module in Chapter 5 defines a methodology for integrating imaging outputs with psychosocial PROMs to generate data-driven patient phenotypes, addressing the clinical orchestration gap.

The LLM-RAG Synthesis Module for advanced clinical summarization is excluded due to distinct computational requirements and the need for specialized clinical text corpora, and is designated as future work.

Clinical Focus: The framework targets diagnostic support and patient stratification in non-specific mechanical low back pain. This reflects the lumbar spine’s susceptibility to degenerative changes due to its load-bearing function. The system automates interpretation of key degenerative features from T2-weighted lumbar MRI and integrates these with cross-sectional PROMs (pain, disability, quality of life, coping).

Data and Validation: Model development and validation use a primary clinical dataset, with external testing on public multi-institutional datasets (e.g., RSNA LumbarDISC, SPIDER) to assess robustness across imaging protocols. Public datasets may lack annotations for certain pathologies addressed in the primary study.

1.6.2 Delimitations

The following explicit boundaries clarify what this foundational work does not encompass, ensuring a tractable project and preventing misinterpretation of its contributions.

- **Clinical Integration and Validation Depth:** This work delivers validated algorithmic components and a system blueprint. It does *not* include full-scale clinical deployment, integration into live electronic health record systems, prospective clinical trials, or regulatory approval processes, as these require multi-year, interdisciplinary efforts beyond a single dissertation.
- **Technical and Methodological Boundaries:** The Computer Vision Module is developed exclusively for **lumbar spine T2-weighted MRI**. Other spinal regions (cervical, thoracic), other imaging modalities (CT, X-ray), and other MRI sequences (e.g., T1, contrast-enhanced) have been excluded. The integration methodology treats data as static; longitudinal analysis of symptom trajectories or treatment response modeling is out of scope. Furthermore, while expert radiologist labels provide the ground truth for model training, the

inherent inter-rater variability in these labels is acknowledged as a constraint of the clinical data rather than a factor controlled within this study.

- **Clinical Application and Generalizability:** The framework is designed as a decision-support tool for non-specific LBP. It is explicitly *not* designed for, nor validated on, cases of radiculopathy with confirmed compressive etiology, post-surgical spines, high-grade instability, or traumatic/inflammatory spinal conditions. Its outputs are synthesized evidence summaries to inform personalized assessment, not autonomous treatment recommendations or surgical planning tools.

1.7 Dissertation Outline

This thesis is structured to systematically build the components of the integrative framework proposed to address the research problem (Section 1.4). The chapters progress from establishing the foundation and architecture to developing, validating, and synthesizing the core technical and methodological contributions.

Chapter 2: Literature Review establishes the clinical and technical foundation. It examines the multifactorial nature of low back pain, the role and limitations of spine MRI, the prognostic value of patient-reported outcomes, and the current state of AI in medical imaging and multimodal analysis, thereby precisely defining the research gaps this dissertation targets.

Chapter 3: A Framework for Integrative AI in Spine Care presents the architectural response to the identified gaps. It proposes a three-module conceptual blueprint designed to operationalize the biopsychosocial model, detailing the envisioned data flow and interactions between automated imaging analysis, multimodal patient assessment, and clinical synthesis.

Chapter 4: Establishing a Robust Imaging Foundation (Module 1) fulfills the first research objective. It begins with an external validation of SpineNetV2, providing an independent performance benchmark and identifying key limitations. It then introduces and validates *LumbarAI*, our novel deep learning model that uses curriculum learning to achieve reliable, multi-pathology grading from both sagittal and axial MRI views.

Chapter 5: A Methodology for Clinical Data Integration (Module 2) fulfills the second research objective. This chapter develops the integration methodology in two parts: first, by analyzing the cross-sectional relationships and mediated associations between coping strategies and clinical outcomes; second,

by applying unsupervised clustering to the fused outputs of imaging data and psychosocial PROMs, thereby identifying distinct, data-driven patient phenotypes.

Chapter 6: Proof-of-Concept: Demonstrating the Framework’s Logic demonstrates the feasibility of core integrative premise of the framework. Using the outputs from Chapter 4 (imaging findings) and Chapter 5 (phenotype profiles), this chapter presents detailed case analyses to demonstrate how their concurrent application enables a richer, more holistic patient characterization rather than the interpretation of either data stream in isolation.

Chapter 7: Summary, General Discussion, and Conclusions synthesizes the dissertation collective findings. It provides a concise summary of each chapter, examines their integrated implications for the biopsychosocial model and AI in spine care, and concludes by outlining the defined trajectory for future research and clinical translation.

1.8 Summary of Contributions

This dissertation makes the following contributions to the field of AI-enhanced spine care, each situated within the context of existing literature:

Technical Contributions

1. **A clinically-grounded architectural framework for integrative spine care:** While prior work has developed AI models for isolated tasks, such as disc-related pathology severity grading [40, 48] or pain prediction from questionnaires [28], no existing framework specifies how these components should interoperate in a clinical workflow. This dissertation proposes a novel modular architecture (Chapter 3) that explicitly defines interfaces between automated imaging analysis, patient profiling, and clinical synthesis, providing a blueprint for moving beyond siloed AI toward integrated decision support.
2. **LumbarAI: A multi-view deep learning model with curriculum learning for spine MRI grading:** Building on single-view architectures like SpineNetV2 [48], which this dissertation first externally validates on an external dataset (revealing performance drops for cross-sectional pathologies), this work develops LumbarAI (Chapter 4). The model introduces two novel elements: (a) a dual-pathway architecture that separately processes sagittal and axial views to mimic radiological reasoning, and (b) a staged curriculum learning strategy that prevents catastrophic forgetting during multi-view fusion. In cross-sectional pathologies (e.g., spinal stenosis, foraminal stenosis),

LumbarAI achieves superior performance improvement over the SpineNetV2 baseline, addressing a critical reliability gap identified in the external validation.

3. **A methodology for deriving interpretable biopsychosocial phenotypes from PROMs:** While previous analyses of patient-reported outcomes in low back pain have typically examined variables in isolation or used simple sum-scores [29?], this dissertation demonstrates a systematic approach combining network analysis with unsupervised clustering, the methodology identifies two distinct patient phenotypes, differentiated primarily by pain self-efficacy versus fear-avoidance beliefs, and quantifies the relative contribution of each psychosocial factor to pain intensity within each phenotype.

Clinical and Translational Contributions

4. **Independent benchmarking of automated spine MRI grading:** By rigorously evaluating SpineNetV2 on a local clinical dataset ($n = 350$ patients) and publicly reporting performance metrics (Section 4.3), this dissertation provides the first independent validation of a state-of-the-art spine AI model outside its development setting. The finding that performance degrades substantially for specific pathologies (notably cross-sectional pathologies such as: foraminal stenosis, central canal stenosis and herniation) establishes a critical benchmark for the field and directly motivated the multi-view architecture of LumbarAI.
5. **Proof-of-concept demonstration of integrative clinical utility:** While prior work has explored methods for incorporating patient-reported data into decision models [32] and hypothesized the value of integrating imaging with psychosocial factors [11], empirical testing of such integrated approaches with clinician users has remained limited. In a small-scale proof-of-concept study involving two radiologists and three patient cases (Chapter 6), we provide initial evidence that presenting integrated AI-generated preliminary reports and PROMs based patient phenotyping can alter treatment prioritization and reduce active reporting time by 40–85% across two radiologists.

Together, these contributions advance the conceptual role of AI in spine care from single-modality automation toward integrative decision support, providing both methodological innovations and empirical evidence for a more holistic approach to low back pain management. With this foundation established, Chapter 2 proceeds to examine the relevant literature in greater depth.

2

Literature Review

This chapter provides a comprehensive literature review establishing the multidisciplinary foundation for an integrative AI framework in spine care. It first examines the clinical context of low back pain, including its epidemiology, pathoanatomy as revealed by magnetic resonance imaging (MRI), the biopsychosocial model, and the critical role of patient-reported outcome measures (PROMs). The review then surveys the technical evolution of artificial intelligence in medicine, with a focused analysis of deep learning applications in spine MRI, machine learning for patient phenotyping, and the emerging field of multimodal data fusion. By synthesizing these clinical and technical domains, the chapter critically identifies key research gaps, specifically, the disconnect between advanced unimodal AI tools and the holistic needs of clinical practice. This synthesis directly motivates and justifies the novel integrative framework proposed in the subsequent chapter.

2.1 Clinical Foundations of Low Back Pain

2.1.1 Epidemiology and Classification of LBP

Low back pain (LBP) represents a preeminent global public health challenge, with lifetime prevalence estimates suggesting it may affect up to 80–90% of individuals, underscoring its near-ubiquitous nature [49, 11, 50]. As the foremost cause of years lived with disability (YLD) worldwide, it imposes a staggering burden on healthcare systems, economies, and individual well-being [51, 52, 2]. This burden is amplified by the inherent clinical complexity of this condition by which a clear pathoanatomical cause is often elusive.

LBP is a multifactorial syndrome rather than a single disease [53], involving complex interactions between mechanical, degenerative, inflammatory, and potent psychosocial determinants [54, 7]. This complexity manifests in the well-documented disconnect between structural imaging findings and subjective pain experience. Degenerative changes visible on magnetic resonance imaging (MRI) are highly prevalent, in a majority of asymptomatic adults and are increasing with age [18, 55]. However, their correlation with the presence, intensity, and trajectory of pain is often weak and inconsistent [19]. This symptom-imaging discordance lies at the heart of the diagnostic paradox in spine care, challenging clinicians to determine the clinical relevance of common anatomical findings.

From a clinical management perspective, LBP is often categorized temporally as acute (< 6 weeks), subacute (6–12 weeks), or chronic (> 12 weeks) [56, 57]. Although useful, these arbitrary temporal boundaries alone are insufficient to capture the condition’s heterogeneity. Chronic LBP (CLBP) is particularly complex, frequently involving a state where the nervous system becomes hypersensitive to pain signals and demonstrating variable treatment responses [58, 59]. Risk factors for its development and persistence are multifactorial, spanning biomechanical exposures (e.g., heavy lifting), lifestyle factors (e.g., obesity, smoking), and, critically, psychosocial factors such as depression, pain catastrophizing, fear-avoidance beliefs, maladaptive coping strategies, and sleep disturbances [60, 61, 62].

Consequently, evidence-based guidelines advocate a patient-centered, stratified approach [63, 64]. The initial assessment focuses on ruling out specific, serious pathology (e.g., fracture, malignancy, cauda equina syndrome) via “red flag” screening [65]. For the remaining 85–95% of cases, classified as non-specific LBP, management prioritizes conservative care (education, physical activity, psychological support), with imaging reserved for specific indications to avoid the common pitfall of misattributing pain to incidental, age-related findings on scans [64]. Although tools like the STarT Back Screening Tool incorporate psychosocial factors for risk stratification, there is still a significant precision gap in predicting individual trajectories and treatment responses [66].

This multifactorial complexity, and its inconsistent representation across isolated streams of clinical data, imaging, and patient reporting, creates a fragmented evidence base that current assessment tools, including unimodal AI solutions, fail to bridge. The limitations of existing approaches underscore the pressing need for a new paradigm: an integrative framework to combine anatomical, functional, and psychosocial information into a coherent, patient-specific assessment. This

unmet clinical need directly motivates the integrative AI framework developed in this dissertation.

2.1.2 Spinal Anatomy and Pathoanatomy on MRI: Foundations for Quantitative Analysis

The lumbar spine is a complex biomechanical structure whose detailed anatomy, comprising vertebrae, intervertebral discs, facet joints, and neural elements, is exquisitely visualized by Magnetic Resonance Imaging (MRI) [67, 68]. MRI has become the cornerstone for diagnosing degenerative pathologies such as disc herniation, spinal stenosis, and nerve root compression [69]. To standardize reporting, classification systems like the Pfirrmann (disc degeneration) and Modic (endplate changes) scales have been widely adopted [13, 70].

However, the translation of this high-resolution imaging into consistent, quantitative clinical data is fundamentally limited [20, 71]. The interpretation of key features, particularly the grading of stenosis severity, classification of disc herniations, and assessment of neural foramen, suffers from significant inter- and intra-observer variability. This subjectivity undermines diagnostic reproducibility, complicates longitudinal tracking (monitoring change over time), and introduces noise into clinical decision-making [72]. Furthermore, heterogeneity in MRI scanners, protocols, and acquisition parameters presents a major challenge for developing robust, generalizable analytical tools.

Consequently, while MRI provides essential structural insight, its operational value as a source of consistent, quantitative data for advanced analytics or personalized medicine is constrained. The current paradigm yields a qualitative, inconsistently measured output. This represents a critical bottleneck: transforming rich imaging data into reliable, standardized metrics. Overcoming this bottleneck through automated, objective interpretation is a prerequisite for building the integrative, data-driven models required for modern spine care.

2.1.3 The Biopsychosocial Model of Pain

The biopsychosocial model, established by Engel, posits that pain is a multi-dimensional experience determined by dynamic interactions between biological, psychological, and social factors [30]. This represents a necessary paradigm shift from a purely biomedical, structural model, directly addressing its core failure: the frequent discordance between tissue pathology observed on MRI and the patient’s reported pain intensity, disability, and treatment response [73, 19].

Psychological factors, such as pain catastrophizing, fear-avoidance beliefs, depression, and maladaptive coping strategies (e.g., avoidance), are consistently among the strongest predictors of pain chronicity and functional disability, often outweighing the prognostic value of specific structural findings [27, 74, 26]. **Social factors**, including occupational stress, lack of social support, and socioeconomic barriers, further modulate the pain experience and access to care [61, 4]. The model implies that effective assessment must address this multifaceted reality. It defines the necessary multidimensional space for understanding LBP, moving beyond the search for a structural “cause” to explain why patients with similar MRI findings have vastly different outcomes [75]. While conceptually foundational, the biopsychosocial model remains largely unimplemented in clinical systems. Manual synthesis of these domains is low-fidelity and heuristic-driven. **The critical gap is the lack of a framework to quantitatively integrate psychological and social components, most often captured via Patient-Reported Outcome Measures (PROMs), alongside with anatomical imaging data.**

2.1.4 Patient-Reported Outcome Measures (PROMs) in Spine Care

Patient-Reported Outcome Measures (PROMs) are standardized instruments designed to directly capture the patient’s subjective health status, providing the essential quantitative data to operationalize the psychological and social domains of the biopsychosocial model [76]. In LBP, key PROM categories include: **Pain Intensity:** Numeric Rating Scale (NRS), Visual Analog Scale (VAS) [77]. **Function and Disability:** Oswestry Disability Index (ODI), Roland-Morris Disability Questionnaire (RMDQ) [78, 79]. **Psychological and Coping Factors:** Pain Catastrophizing Scale (PCS), Coping Strategies Questionnaire (CopeNVI) [27, 26].

Despite their established validity and prognostic superiority to imaging findings for many outcomes [29], PROMs face significant challenges. In clinical practice, they are often collected in administrative silos, temporally misaligned with imaging studies, and subject to data quality issues like incomplete responses [80]. Critically, their quantitative results are rarely systematically co-registered or computationally integrated with MRI-derived data within clinical workflows.

Therefore, PROMs and MRI represent parallel, unlinked evidentiary streams. This segregation undermines a global assessment, forcing clinicians to perform a complex cognitive evaluation. PROMs thus embody a dual role: they are essential to value the patient’s experience independently of imaging, and simultaneously

a core component to the integration problem. The challenge is not the value of PROMs, but the creation of frameworks to fuse this experiential data stream with the anatomical data stream from MRI, enabling a truly biopsychosocial analysis.

2.2 Technological Foundations: AI in Healthcare

2.2.1 Deep Learning for Medical Image Analysis

Deep learning, particularly convolutional neural networks (CNNs), has become the dominant approach for medical image analysis, enabling systems from early feasibility studies [81] to comprehensive multi-pathology analysis [82]. However, successful application in complex domains like spine MRI requires overcoming challenges beyond basic model design.

A primary architectural limitation of conventional CNNs lies in their restricted receptive fields, which hinder the modeling of long-range spatial dependencies, especially critical when analyzing complex 3D anatomical relationships across multiple MRI slices in spine imaging [83]. Vision Transformers (ViTs) and their hierarchical extensions, notably the Swin Transformer [83] and its 3D counterpart Swin-Video [84], address this shortcoming by leveraging self-attention mechanisms to capture global contextual information efficiently. In the domain of volumetric medical imaging, 3D Swin-based architectures such as Swin-UNETR [85] and SwinBTS [86] have shown strong performance in 3D segmentation tasks, benefiting from hierarchical multi-scale feature learning and localized windowed self-attention. However, their application to multi-task pathological classification in spine MRI, which requires simultaneous interpretation of diverse, spatially distributed abnormalities, remains largely unexplored.

Furthermore, the field contends with a fundamental “data bottleneck”. Performance is constrained by: (1) small, heterogeneously annotated datasets [87]; (2) the high cost of expert annotation; and (3) the inherent moderate inter-rater agreement in subjective grading tasks (e.g., Pfirrmann), which sets a ceiling on model reliability [88]. This is compounded by acquisition heterogeneity across MRI scanners, introducing domain shift that degrades generalizability [89], making strategies like robust training and harmonization essential for clinical translation.

2.2.2 The State-of-the-Art and Limitations in Automated Spine MRI Interpretation

Building on these foundations, automated spine MRI analysis has progressed to grading pathologies like disc degeneration and stenosis [88]. A critical analysis reveals interconnected limitations.

The most prominent is the prevalent single-view, 2D paradigm. State-of-the-art systems like SpineNetV2 [48], a multi-task CNN validated across diverse populations [90, 91, 47, 92], are typically trained and validated on sagittal slices alone. This creates a critical “3D information bottleneck” [88]. Pathologies like foraminal stenosis and disc herniation morphology are optimally characterized on axial images [93, 94, 95], and models limited to sagittal views show degraded performance on these tasks [47].

This points to a broader multiview integration challenge. The clinical workflow seamlessly uses sagittal and axial views, yet most AI systems lack this capability. A core technical barrier is catastrophic forgetting [96], where introducing a new view degrades performance on previously learned tasks. While curriculum learning strategies offer a framework for progressive integration in other domains [97, 98], a dedicated view-level curriculum for spine MRI, designed to preserve single-view expertise while acquiring multiview competence, is missing.

In summary, while 3D Transformers can overcome CNNs’ context limitations and curriculum learning can manage integration complexity, no existing framework synthesizes them to solve the spine-specific multi-view integration problem. This gap, the need for a system that leverages volumetric attention and a progressive view-based curriculum to integrate axial data without forgetting sagittal expertise, directly motivates the LumbarAI framework presented in Chapter 4.

2.2.3 Machine Learning for Biopsychosocial Phenotyping and Stratification

Machine learning applied to patient-reported outcome measures (PROMs) offers a powerful, complementary axis to imaging AI by computationally modeling the biopsychosocial determinants of low back pain (LBP) [99]. This approach directly confronts the clinical paradox, revealing that patient subgroups can be defined by psychological and functional characteristics that are largely orthogonal to structural findings. A comprehensive 2025 study specifically in workers with chronic LBP demonstrated the utility of this method, applying hierarchical clustering to PROMs like the ODI and VAS to identify distinct biopsychosocial phenotypes with

significantly different levels of disability, absenteeism, and psychological distress [100]. This empirically validates the existence of patient subgroups not visible through imaging alone.

Unsupervised clustering techniques are central to this effort, identifying phenotypes based on patterns in pain intensity, disability (ODI, RMDQ), psychological distress (catastrophizing, fear-avoidance), and coping strategies [101, 102, 100, 103]. These methods can distinguish, for example, a “high-distress, high-disability” cluster from a “low-distress, mechanical-pain” cluster, demonstrating that symptom severity is often poorly predicted by imaging alone [104, 100, 105, 106]. The value of this approach is recognized beyond LBP; a 2024 systematic review of studies using machine learning to integrate biological, psychological, and sociological data for modeling suicide risk found that such “Whole Person Modeling” provides early empirical evidence for the superior predictive power of the biopsychosocial model over single-domain approaches [102].

However, PROMs-based phenotyping has significant limitations that constrain its clinical translation. First, phenotypes are typically static snapshots derived from cross-sectional data, failing to capture dynamic patient trajectories or treatment-response patterns. Second, the clusters themselves can be unstable, sensitive to the choice of instruments, preprocessing methods, and study population, with limited external validation in diverse spine care settings [102, 100]. A fundamental philosophical critique also exists: scholars argue that the current biopsychosocial model, while a conceptual advance, is itself not operationally prepared for the highly individual predictions generated by machine learning, lacking a framework to highlight the importance of each factor for a specific patient [107].

Most critically, these data-driven profiles exist in a conceptual and operational silo. They provide a rich description of the “biopsychosocial state” but lack a standardized ontology or data structure to be integrated and synthesized with the “anatomical state” derived from imaging. This lack of a shared framework for joint representation creates a major barrier to automated integration [102, 107]. Consequently, while phenotyping stratifies risk and imaging AI grades anatomy, clinicians are left to manually reconcile these two distinct, evidence-based representations of the same patient, a cognitive task for which they lack dedicated tools. This defines a critical gap: the absence of a framework for the integrative analysis of a patient’s unique combination of structural pathology and biopsychosocial profile, a necessary step toward actionable, personalized clinical decision support [108, 109].

2.2.4 The Integration Gap: From Disparate AI Tools to Clinical Integration

Recent AI advances have created sophisticated yet isolated tools for spine care. In spinal imaging, deep learning models demonstrate strong performance in detecting and grading pathologies such as disc degeneration and stenosis [110, 48]. However, these models largely remain confined to research environments, with limited external validation and minimal integration into clinical IT ecosystems like PACS and EHRs [88, 111]. This isolation is compounded by the “black box” nature of many models, which undermines the clinical trust necessary for adoption. While explainability methods like saliency maps are recognized as essential for high-stakes decision-making [112, 113], they are often not implemented in a way that integrates with a clinician’s workflow.

Parallel to imaging AI, the field has developed robust methodologies for incorporating the patient’s perspective. Patient-reported outcome measures (PROMs) provide critical data on pain, disability, and psychosocial factors that are poorly correlated with imaging findings [114]. Systems like PROMIS offer standardized instruments [115], and machine learning techniques can stratify patients into distinct biopsychosocial phenotypes [116]. Despite their prognostic value, these data streams are operationally siloed; they are rarely computationally fused with imaging data to create a unified patient profile.

Emerging technologies like large language models (LLMs) present a potential pathway for integration. They have shown utility in tasks such as drafting radiology reports [117] and could, in theory, synthesize multimodal data [118]. However, current applications are narrow. For instance, some frameworks use LLMs for treatment support but lack diagnostic imaging input [119], while others extract data from reports but do not perform the integrative reasoning required for diagnosis [120]. This exemplifies a broader trend: LLMs are not yet deployed in a manner that performs the core clinical task of integrative reasoning across data types.

A synthesis of these research trajectories reveals a persistent, system-level gap. Studies either advance unimodal capabilities (e.g., imaging interpretation [121, 39], PROMs analysis [116]) or develop single-module support tools [119]. As noted in surveys of the field, there is limited incorporation of LLMs and PROMs into integrated clinical workflows [122]. Consequently, while the component technologies exist, the literature demonstrates a clear absence of a framework designed for **integrative clinical synthesis**, one that merges reliable imaging analysis,

quantitative psychosocial profiling, and evidence-based narrative generation within a clinician-verified, human-in-the-loop workflow.

2.2.5 Human-in-the-Loop AI and Explainability

Human-in-the-Loop (HITL) systems integrate human expertise with computational intelligence, recognizing that clinical decision-making requires both objective analysis and contextual judgment [44, 123]. Rather than replacing clinicians, HITL approaches enable collaborative workflows where AI provides preliminary analysis that clinicians review, edit, and validate [124].

For such collaboration to be effective, explainability is essential. Clinicians must understand AI reasoning to trust and appropriately act on its outputs [37]. In medical imaging, saliency maps and attention visualizations can highlight which image regions drive predictions [112]. For multimodal systems, explaining how different modalities (imaging, PROMs) contribute to conclusions enables clinicians to validate cross-modal associations [113].

Trust calibration is equally critical: AI confidence scores must align with actual accuracy to avoid over-reliance or under-utilization [125]. Uncertainty estimation techniques can flag low-confidence cases for human review, optimizing the allocation of expert attention [126].

Despite these advances, HITL principles have not been systematically applied to integrative spine care. Existing systems offer unimodal explanations (e.g., MRI heatmaps) but lack frameworks for explaining synthesized conclusions that combine imaging and psychosocial data. This gap, the absence of explainable, human-verifiable integration, directly motivates the design of the framework proposed in Chapter 3.

2.3 Synthesis: The Integration Imperative in AI for Spine Care

The preceding review delineates significant yet parallel advancements across three domains: the automation of spine MRI interpretation, the computational modeling of biopsychosocial patient profiles, and the principles of human-centered AI design. However, a critical synthesis reveals that progress within each domain has, paradoxically, reinforced the very clinical fragmentation it sought to overcome. The identified gaps are not isolated but are causally linked, compounding facets of a singular, systemic failure: the field confusion of technical multimodal integration

with actionable clinical synthesis, and the consequent absence of a principled framework to achieve the latter.

2.3.1 The Self-Reinforcing Cycle of Clinical Fragmentation

The identified research gaps are not isolated problems but form a self-reinforcing cycle. A limitation in one domain creates a fundamental barrier to progress in the next, perpetuating a state of fragmentation where powerful tools remain disconnected.

1. The Unreliable Foundation of Imaging AI. Automated grading models are limited by their architecture (e.g., struggling to analyze both sagittal and axial views) and, more fundamentally, by the moderate agreement between expert radiologists. For example, if two radiologists disagree on whether a stenosis is ‘moderate’ or ‘severe,’ any AI trained on their labels inherits this ambiguity. If this core anatomical input is unstable, any downstream system built upon it will be fundamentally constrained.

2. The Conceptual Mismatch with Patient Experience. Advanced computational phenotyping using Patient-Reported Outcome Measures (PROMs) operates in a separate conceptual space. It identifies groups like “high pain catastrophizers” or “low-functioning patients,” but these statistical clusters lack a shared language to connect to imaging findings. There is no standard way to computationally relate “Pfirschnann Grade III degeneration at L4-L5” to a “high score on the Pain Catastrophizing Scale.” This is not a technical glitch but an ontological divide, the language of anatomy and the language of lived experience are not aligned.

3. The Missing Bridge for Clinical Decision-Making. Even if the first two gaps were solved, current human-AI interaction research offers generic solutions. For a spine specialist, an AI must do more than show a heatmap on an MRI and a bar chart of PROMs scores side-by-side. It must actively help synthesize them, for instance, by proposing a draft report statement like, “Severe central canal stenosis is present, but the patient’s extreme disability score suggests psychosocial factors are likely major contributors.” Systems are not designed from the ground up to enable this kind of cross-modal reasoning and editable synthesis, creating a final barrier to clinical trust and utility.

This cycle reveals the core challenge: the field possesses powerful, isolated tools (an imaging grader, a phenotyping algorithm) but lacks the integrative methodology to combine them into a single, clinically coherent assessment. The literature

confirms that imaging alone cannot explain symptoms, and psychosocial data alone cannot pinpoint structural causes. Therefore, the paramount need is not for better individual algorithms, but for a principled framework to synthesize them.

2.3.2 The Core Unmet Need: A Framework for Clinical Synthesis

The critical void is the lack of a clinical synthesis framework. This entails a paradigm shift from technical fusion to reasoning-enabled integration. Such a framework must be explicitly designed to:

- Bridge the Ontological Divide: Create a shared computational-representational space that enables reasoning across modalities, mapping categorical imaging grades to continuous psychosocial variables in a semantically meaningful way.
- Generate Interpretable, Actionable Narratives: Move beyond parallel outputs to produce synthesized summaries that explain why a specific combination of findings matters for a patient’s care. Interpretability here is not a usability feature but the essential mechanism allowing clinicians to validate cross-modal associations and build trust.
- Embed Human-Centric Workflow as a First Principle: Incorporate clinician verification and editing not as an add-on, but as the core feedback loop for ensuring safety, enabling continuous refinement, and maintaining diagnostic authority.

Current research addresses facets of this challenge but treats them in isolation. The field has advanced multimodal fusion (mathematical data combination) but has neglected clinical synthesis (the production of actionable reasoning). No existing approach in the spine care literature proposes a cohesive architecture that binds these elements together from first principles to computationally resolve the symptom-imaging paradox.

2.3.3 Conclusion

In summary, the trajectory of AI in spine care reveals a compelling paradox: substantial innovation within technical siloes has led to a landscape of sophisticated but disconnected tools, perpetuating the very clinical fragmentation it aimed to solve. The biopsychosocial model provides the conceptual map for holistic understanding, but the field lacks the computational infrastructure to navigate it.

The imperative that emerges is not for incremental improvements but for a foundational shift toward integrative system design, from developing better point

solutions to architecting systems for clinical reasoning. This shift requires a new kind of framework, one conceived from the outset to enable the trustworthy synthesis of multimodal data into clinician-interpretable, patient-specific insight. The next chapter responds directly to this imperative by proposing a novel architectural framework designed to operationalize this shift and bridge the long-standing divide between multimodal data and personalized clinical decision-making.

This chapter has reviewed the technical and clinical foundations for AI in low back pain (LBP) management. It established LBP's biopsychosocial complexity and critically analyzed advances in spine MRI analysis and patient phenotyping. The synthesis revealed a critical gap: despite progress in individual domains, the field lacks a framework for their clinical integration. The identified limitations directly motivate the integrative architecture detailed in Chapter 3, which serves as the blueprint for this thesis.

3

A Clinically Integrated AI Framework for Personalized Spine Care

This chapter presents the conceptual architecture for a modular AI ecosystem designed to bridge the gap between advanced computational tools and clinical decision-making in spine care. Building directly on the integration imperative identified in Chapter 2, the proposed framework is designed for clinical deployment from first principles, prioritizing explainability, interoperability with hospital IT systems, and a human-in-the-loop workflow. This chapter details the system design, clinical motivation, and proposed use cases, serving as the foundational blueprint that the subsequent empirical chapters will attempt to validate.

3.1 Introduction

As presented earlier, low back pain (LBP) represents a preeminent global health challenge characterized by staggering prevalence, substantial economic cost, and, most critically, etiological complexity that straddles structural, clinical, and psychosocial domains [127, 128]. As established in the preceding literature review, this complexity has fostered a landscape of advanced but disconnected technological solutions: artificial intelligence (AI) excels in automated imaging analysis and patient phenotyping, yet these tools operate in parallel silos, lacking the methodological framework for clinical synthesis.

Consequently, the critical barrier to personalized spine care is no longer a shortage of data or analytical power, but the absence of a system capable of integrating these elements into a coherent, actionable clinical narrative. The field possesses robust

imaging AI, validated patient-reported outcome measures (PROMs), and emerging synthesis technologies like large language models (LLMs), but no architecture exists to orchestrate them within a clinician-verified workflow.

This chapter introduces a framework designed to address this integration gap. Its development is guided by a core thesis: that effective AI for spine care must be architected not as a collection of isolated models, but as a unified system for clinical synthesis. This requires moving beyond technical data integration to enable reasoning that mirrors the clinician’s task, reconciling anatomical findings with the patient’s lived experience to inform personalized assessment. Therefore, the proposed framework is built upon foundational design principles of interoperability, explainability, and human-in-the-loop collaboration, ensuring it complements rather than disrupts clinical expertise.

The chapter proceeds as follows: Section 3.2 derives the framework’s design requirements from stakeholder needs and clinical decision points. Section 3.3 details the full system architecture, from data ingestion to the AI processing core and clinical interface. Section 3.4 illustrates its application through key use case scenarios. Finally, Section 3.5 discusses the framework contributions, positions it within the research landscape, and outlines a roadmap for the validation work presented in the remainder of this dissertation.

3.2 Clinical Imperative and Foundational Design Principles

Building upon the integration imperative established in Chapter 2, this chapter translates the identified gaps into a concrete architectural blueprint. Since clinical synthesis is ultimately a human cognitive task, the architecture must be explicitly grounded in the needs and reasoning processes of its end-users. The following section details this user-centered design methodology, synthesizes the elicited requirements, and derives the core technical principles that govern the framework structure.

3.2.1 User-Centered Design: Defining Clinical Requirements

The framework design was initiated through engagement with clinicians at the X-ray service s.r.l spine care clinic [129], which funds and hosts this research.

A series of semi-structured discussions were conducted with practicing radiologists, pain specialists, musculoskeletal specialists and rehabilitation therapists. These discussions, focused on diagnostic challenges and workflow inefficiencies, were analyzed to identify cross-specialty priorities. This qualitative, practice-informed approach complements the literature review to ensure the framework addresses genuine clinical needs. The following six synthesized priorities emerged consistently:

- **Integrate Disparate Data:** Reduce the cognitive burden of manually integrating fragmented information from imaging, clinical notes, and patient-reported outcomes (PROMs) into a unified patient profile.
- **Support Clinical Reasoning:** Aid in interpreting the interaction between structural pathology and the biopsychosocial context (e.g., how a specific disc finding relates to a patient’s coping strategies and occupational demands) to explain symptom severity and predict trajectories.
- **Facilitate Interdisciplinary Coordination:** Create a shared, multimodal evidence base to align perspectives across specialties, improving communication throughout the care pathway.
- **Enhance Patient Engagement:** Provide a coherent structure to translate complex, multi-domain findings into accessible language, supporting shared decision-making and therapeutic alliance.
- **Enable a Learning System:** Support continuous improvement through clinician feedback and outcome data, while maintaining rigorous data governance and privacy. This implies a system designed for iterative refinement.
- **Uphold Clinical Authority & Ethics:** Augment, never replace, clinical judgment. This necessitates editable outputs, clear uncertainty reporting, and mechanisms to mitigate automation bias, preserving the clinician’s ultimate diagnostic authority.

These priorities establish the foundational user requirements, highlighting that the system must solve cognitive and workflow challenges, not merely technical ones. A critical implication is the need for a shared representational layer, a computational structure to relate imaging findings to psychosocial constructs, which addresses the ontological gap identified in Chapter 2.

3.2.2 Synthesized Stakeholder Needs and Clinical Decision Points

The cross-specialty clinician priorities manifest as distinct functional requirements for key stakeholders. Translating these needs into specific system functions provides

a direct blueprint for the architectural components developed in Section 3.3.

Radiologists are the primary consumers of the imaging AI component. Their core need is to efficiently and confidently verify and finalize diagnostic findings. To support this, the system must provide explainable outputs, such as saliency map, confidence scores and feature attributions, that allow validation of the AI’s reasoning for specific gradings (e.g., stenosis severity). Furthermore, it requires editable interfaces that enable the straightforward correction of AI errors, particularly those arising from domain shift or ambiguous imaging features, all within a workflow that integrates seamlessly with existing PACS to avoid disruptive context switching. These functions collectively ensure the AI acts as a verifiable and trustworthy assistant, fully upholding the radiologist’s diagnostic authority.

Pain Specialists (e.g., anesthesiologists, interventional pain physicians, neurologists, spine surgeons) require the synthesis of multimodal data to inform high-stakes management decisions, including risk stratification, the choice between surgical and conservative pathways, and the setting of realistic patient expectations. For this group, the system must therefore generate structured, interpretable risk profiles that integrate quantitative anatomical severity from imaging with key psychosocial and functional risk factors derived from PROMs. Additionally, it should provide prognostic context by placing the patient’s current profile within a cohort of similar historical phenotypes. By accessing a (hypothetical) longitudinal registry of outcomes, it can answer a critical clinical question: “For patients with a similar profile of anatomy and psychosocial factors, what were the typical 1- and 2-year trajectories for pain, function, and probability of surgery?” This moves beyond a static snapshot to offer evidence-based insight into probable futures. This functionality directly translates the integrative biopsychosocial model into actionable clinical intelligence. It shifts the specialist’s role from manually correlating disparate data points to reviewing a synthesized, evidence-based foundation for intervention planning and nuanced patient counseling about likely outcomes.

Rehabilitation Clinicians (e.g., physiotherapists, occupational therapists, psychologists, nurses, physiatrists) focus on personalizing and monitoring therapy. Their decisions focus on customizing treatment plans and tracking functional recovery progress. To be clinically useful, the system must generate individualized treatment suggestions that thoughtfully consider structural limitations alongside

patient-specific functional goals and psychological barriers. It must also present this information through a unified clinical dashboard that clearly tracks PROM trends, such as pain and disability scores, with the specific anatomical and functional constraints identified in the imaging. This allows the clinician to see, for instance, whether a plateau in functional improvement is concurrent with a particular structural limitation, thereby supporting a holistic, biopsychosocial view of the patient’s recovery journey and enabling truly tailored, evidence-based care planning.

3.2.3 Derived Design Principles

The clinical requirements and the synthesized stakeholder needs converge into four governing design principles that serve as the architectural foundation for the proposed framework. These principles are not merely desirable features but essential constraints derived from the imperative to build a system for *clinical synthesis* rather than isolated analysis.

Modularity and Interoperability constitute the first principle, addressing the need for a system that can evolve and integrate into heterogeneous clinical environments. The framework must be composed of discrete, standards-based modules, such as dedicated components for imaging analysis, psychosocial profiling, and clinical synthesis, that can be developed, validated, and updated independently. This modular design is essential for two key reasons: 1) Research and Development Efficiency, as it allows individual modules (e.g., a new AI model for stenosis grading) to be developed, tested, and validated in isolation without overhauling the entire system; and 2) Long-term Clinical Sustainability, enabling incremental updates and the replacement of obsolete components. Crucially, these modules must interoperate seamlessly with existing hospital IT infrastructure. Adherence to established standards like Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) for clinical data and Digital Imaging and Communications in Medicine (DICOM) for imaging is non-negotiable for enabling this integration, ensuring that AI-generated insights can flow into and augment established clinical workflows without creating disruptive new silos [130, 131].

Explainability by Design forms the second, critical principle for building clinical trust and ensuring safety. For clinicians to rely on AI outputs, especially those informing high-stakes decisions, the system’s reasoning must be transparent. This necessitates that explainability mechanisms, such as feature attribution scores, attention visualizations for imaging models, confidence scores and traceable logic for multimodal synthesis, are embedded as core functionalities within each AI

module, not added as post-hoc justifications. This transparency allows clinicians to validate findings, detect potential biases, and understand the relative contribution of different data modalities to a final recommendation. Beyond being a usability concern, this principle is increasingly a regulatory mandate, as frameworks like the European Union’s AI Act classify medical AI systems as high-risk, explicitly requiring adequate transparency and human oversight measures [112, 132, 133].

Human-in-the-Loop (HITL) Collaboration is the third principle, operationalizing the ethical imperative to augment rather than replace clinical expertise. The architecture must be fundamentally structured to support, not bypass, clinician oversight. This means designing interfaces and data flows that allow, and indeed require, clinicians to review, correct, and provide final approval for all key AI-generated outputs, from a graded imaging finding to a synthesized treatment summary. This collaborative approach mitigates risks like automation bias and ensures the clinician retains ultimate diagnostic authority [133, 134]. Furthermore, this interaction creates a valuable feedback loop; clinician corrections become curated data that can be used to continuously refine and improve the underlying AI models, fostering a learning health system [135].

Workflow-Centric Integration is the final principle, ensuring the framework delivers value in practice, not just in theory. The system must be designed to augment, not disrupt, the existing sequence of clinical activities, from initial patient assessment and data acquisition through specialist interpretation and collaborative treatment planning. As illustrated in Figure 3.1, the framework components should act as intelligent aids at specific, high-value touchpoints within this pathway. This requires a deep understanding of clinical rhythms and a flexible architecture that can adapt to variations in care pathways across different healthcare settings, always aiming to reduce cognitive load and administrative burden rather than adding to it. The following subsection applies this principle through a detailed analysis of the representative spine care pathway (Figure 3.1), mapping the framework’s components to specific clinical stages to demonstrate this augmentation in practice.

3.2.4 Clinical Workflow Integration: Mapping the Framework to Practice

For the proposed framework to be viable, it must integrate into the real-world pathways of spine care. While specific sequences can vary by healthcare system, with differences in gatekeeping, specialty referral, and the depth of psychosocial assessment, a generalized care trajectory can be abstracted to reveal common,

high-value touchpoints for AI augmentation. Figure 3.1 illustrates one such representative pathway. The framework is designed to interface with this trajectory not by imposing a rigid new process, but by providing intelligent support at its core stages, each of which generates or requires specific data that the system must organize, interpret, and present in a clinically meaningful way.

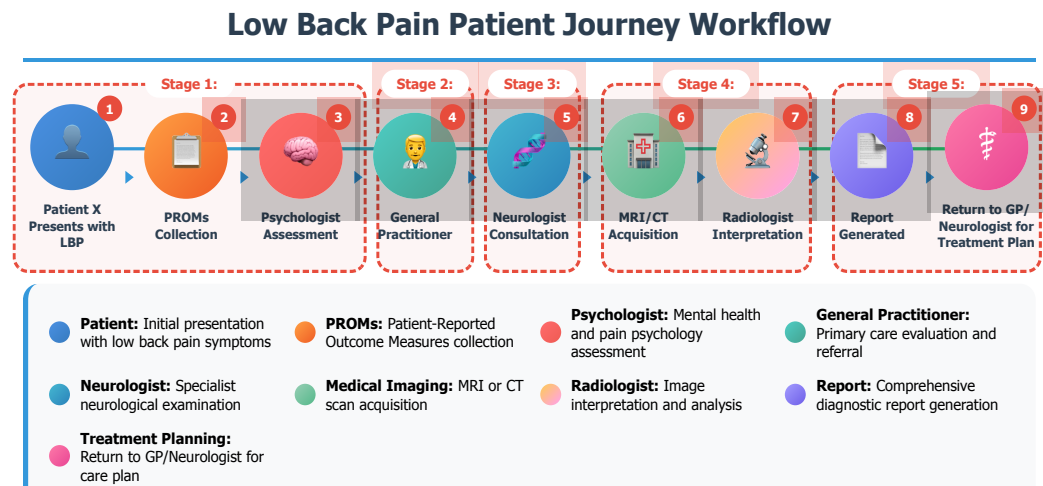


Figure 3.1: *Multidisciplinary Patient Journey Workflow for Low Back Pain Assessment.* A comprehensive 9-step sequential workflow illustrating the patient care pathway from initial presentation through treatment planning. Each step is color-coded by healthcare specialty: patient presentation (blue), PROMs collection (orange), psychological assessment (red), primary care evaluation (teal), neurological consultation (blue), medical imaging acquisition (green), radiological interpretation (yellow-pink), report generation (purple), and treatment planning (pink).

Stage 1: Patient Presentation, Psychological Assessment, and PROMs Collection This initial stage generates the foundational psychosocial and experiential data layer. It involves the collection of patient-reported outcome measures (PROMs) that quantify pain, disability, quality of life, and coping strategies, alongside potential formal psychological evaluation. This stage is critical as it captures dimensions of the pain experience independently from imaging. The framework data ingestion layer must be flexible enough to accommodate the variability in this stage, where the depth of psychosocial assessment may differ based on clinical context and resources.

Stage 2: Primary Care (GP) Evaluation Here, a general practitioner or primary care specialist performs a comprehensive clinical assessment, systematically documenting medical history, comorbidities, lifestyle factors, and demographic

data. This stage produces a structured clinical context profile. The data from this stage provides essential risk stratification variables and comorbid flags that must be fused with both the PROMs from Stage 1 and the anatomical findings from later stages.

Stage 3: Specialist Neurological Consultation This stage represents a key clinical decision point. A neurologist or other spine specialist conducts a focused examination to assess neurological deficits, determine the likely pain generator, and establish the specific indication for advanced imaging. The framework must respect and be triggered by this clinical reasoning, ensuring that subsequent automated imaging analysis is performed in direct response to a justified, specialist-driven diagnostic hypothesis.

Stage 4: Imaging Acquisition and Radiological Interpretation This is the core anatomical data generation stage, involving the acquisition and interpretation of spine MRI or CT scans. Traditional AI tools operate in isolation here. The framework *Computer Vision Module* engages at this point with a crucial architectural distinction: it functions as a structured data provider. Its outputs are not merely a visual aid for the radiologist but are formatted as standardized, graded anatomical descriptors designed to be immediately consumable by downstream multimodal synthesis modules.

Stage 5: Comprehensive Reporting and Integrated Treatment Planning This culminating stage is where the synthesis gap is most apparent and where the framework delivers its core value. It requires the integration of all prior data streams, psychosocial (*Stage 1*), clinical (*Stage 2*), contextual (*Stage 3*), and anatomical (*Stage 4*), into a coherent narrative for diagnosis and management. The framework targets this gap directly: its *Multimodal Profiling Module* fuses these heterogeneous data types, and its *LLM Synthesis Module* generates a preliminary, evidence-anchored summary. This output supports the interdisciplinary discussion essential for finalizing a personalized treatment plan, thereby transforming fragmented data into a unified basis for a shared decision-making.

In summary, this five-stage workflow analysis provides the definitive blueprint for the framework modular architecture. Each stage corresponds to a primary data type or clinical function addressed by a specific system component. The overarching architecture is designed to ensure these components work in concert, creating a continuous thread of integrated reasoning that bridges the transitions between

stages, from initial symptom presentation to a comprehensive, personalized care plan.

3.3 System Architecture Overview

The clinical requirements and design principles derived in Section 3.2 require an architecture that goes beyond standard clinical decision support systems. While such systems often integrate isolated AI tools, they rarely address the foundational challenge of biopsychosocial synthesis. Rather than contributing another single-purpose algorithm, this work proposes a modular, interoperable AI ecosystem whose core novelty lies in its modular framework design for integrative clinical reasoning: a pipeline that integrates multimodal data into a unified patient profile and embeds a retrieval-augmented language layer to translate it into evidence-linked, clinician-facing narratives. As illustrated in Figure 3.2, this is realized through a secure orchestration of specialized modules governed by continuous learning loops. *Throughout this chapter, a unified patient profile refers to a structured multimodal representation that jointly encodes anatomical, clinical, and psychosocial information for downstream reasoning.*

3.3.1 Data Input Layer

The system is anchored in the routine data sources of clinical practice to ensure practical viability. A dedicated ingestion layer connects to standard hospital infrastructure: Picture Archiving and Communication Systems (PACS) for DICOM-format spine imaging, and Electronic Health Records (EHR) for structured clinical data. Critically, it also interfaces with Patient-Reported Outcome Measures (PROMs) collection portals. This layer assembles the full multimodal substrate, structural, clinical, and psychosocial, while accommodating the variable availability of PROMs across institutions.

3.3.2 Security and Governance Layer

Prior to any computation, a dedicated security and governance layer mediates all data and model access. This is not merely a compliance checkpoint but a foundational design choice to embed ethical use. It enforces authentication, authorization, and audit logging for both data queries and AI inference requests, while implementing necessary encryption and anonymization. By ensuring compliance with standards like HIPAA [136], GDPR [137] and EU AI act [138] from the

outset, this layer institutionalizes trust and control, a prerequisite for deploying any AI within sensitive clinical environments.

3.3.3 Interoperability and Normalization Layer

With data access secured, the next barrier is heterogeneity. Raw DICOM pixels, HL7 messages, and PROM JSON cannot be directly consumed by unified AI models. Therefore, an interoperability layer acts as the system’s semantic integration fabric. It employs standardized adapters and terminology services (e.g., SNOMED CT for clinical findings [139], LOINC for laboratory observations [140], RadLex for imaging features [141]) to extract, transform, and map data into a harmonized, normalized format. This process resolves syntactic and semantic inconsistencies, creating a coherent staging area of quality-controlled data with preserved provenance, ready for multimodal analysis.

3.3.4 Modular AI Processing Core

The analytical core of the framework is a suite of three synergistic modules. Their sequential design, from unimodal analysis to multimodal profiling to narrative synthesis, constitutes the architectural response to the integration gap. A cross-cutting requirement for these modules is explainability, implemented in module-specific ways to ensure trust at each step.

3.3.4.1 Computer Vision Module

This module provides reliable, automated interpretation of spine MRI, directly addressing the generalizability and inter-rater variability limitations of current tools. It performs disc localization, segmentation, and pathology grading (e.g., Pfirrmann classification). Its output is a set of structured, quantitative imaging descriptors (e.g., “L4-L5: central canal stenosis, Grade 2”) paired with visual evidence (segmentation overlays, saliency maps) and confidence estimates. This implementation represents a core empirical component of this dissertation, developed and validated in Chapter 4.

3.3.4.2 Multimodal Profiling and Synthesis Module

This module is the conceptual and methodological centerpiece of the integrative framework. It ingests the structured imaging descriptors alongside normalized PROMs and clinical data to perform patient profiling. Methodologically, this involves encoding PROMs into normalized numerical features and applying unsupervised or contrastive learning techniques to derive data-driven biopsychosocial

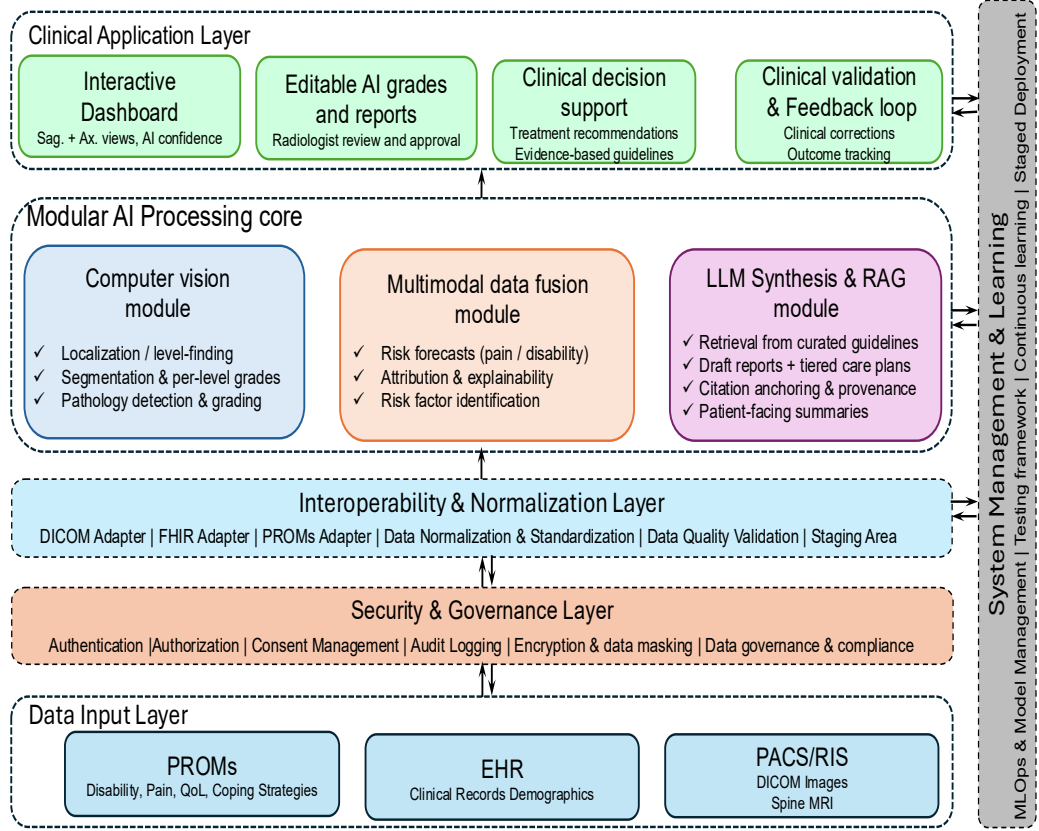


Figure 3.2: *System Architecture Overview: a modular, interoperable AI ecosystem for spine image analysis integrating multimodal data, human governance, and continuous learning.*

phenotypes or unified latent representations. This output, a coherent, quantitative patient profile, is the essential bridge between isolated data silos and is validated through the phenotyping work presented in Chapter 5.

3.3.4.3 LLM Synthesis & RAG Module

This module translates the unified patient profile into actionable clinical language. Importantly, it never operates on raw clinical data; it only consumes the structured, clinician-verifiable profile. To ensure safety and accuracy, it employs a Retrieval-Augmented Generation (RAG) architecture [142], constraining the generative process to a curated corpus of clinical guidelines (e.g., from bodies like NASS [143] or NICE [144]) and institutional protocols. Using structured prompt templates, it generates preliminary, versioned, and editable narratives, where edits generate structured provenance metadata, that explicitly cite their sources in the patient’s data profile and the external knowledge base. This forward-looking RAG-driven approach offers a practical balance between generative flexibility and safety, representing a visionary component for future work.

3.3.5 Clinical Application Layer

This layer operationalizes the human-in-the-loop paradigm by embedding the AI core into clinician workflow. It provides an interactive dashboard for clinicians to verify outputs, editable interfaces for all AI-generated content, and decision-support panels presenting synthesized narratives. Crucially, every clinician interaction and edit is logged, creating the structured feedback essential for system learning and maintaining an audit trail for accountability.

3.3.6 MLOps and Continuous Learning Layer

To ensure the framework evolution, a dedicated MLOps and continuous learning layer manages the full model lifecycle. This includes version control for models and prompts, performance monitoring, staged deployment, and, most importantly, a secure pipeline to use clinician corrections and outcome data for feedback-driven retraining. This introduces a structured mechanism through which clinician interactions serve as auditable feedback for traceable learning, closing the loop for controlled, continuous improvement.

In summary, this architecture provides the technical blueprint for a new class of clinical AI system. Its novelty lies not in its individual layers, many of which use established informatics patterns, but in their specific orchestration to solve the biopsychosocial integration problem. The empirical contributions of this dissertation focus on developing and validating the core AI processing modules (Chapters 4 & 5), thereby demonstrating that the most technically challenging parts of this integrative architecture are feasible. This supports the broader claim that biopsychosocial integration is achievable through a principled, modular design, with the full realization of the integrated ecosystem representing a compelling direction for future multidisciplinary research and development.

3.4 Use Case Scenarios

The following vignettes concretize the proposed framework operation in representative clinical contexts. They serve a dual purpose: first, to illustrate the tangible differences between isolated imaging workflows and AI-augmented multimodal synthesis; and second, to demonstrate how the architecture core components, particularly the LLM/RAG layer and the human-in-the-loop paradigm, translate conceptual design into clinical utility. These scenarios are hypothetical but

grounded in common clinical presentations, intended to clarify the framework’s anticipated behavior and to guide subsequent validation efforts.

Scenario 1: Chronic Low Back Pain in an Office Worker A 45-year-old office worker presents with chronic low back pain. The spine MRI shows moderate L4–L5 disc degeneration and mild central canal stenosis.

- **Imaging-Only Workflow:** An automated grading tool outputs structured findings (e.g., “Pfarrmann Grade IV, mild canal stenosis”). The radiologist must manually contextualize these findings, a process vulnerable to inter-rater variability and increased cognitive load from reconciling data across disparate systems (PACS, EHR).
- **Multimodal Framework Workflow:** The system ingests the imaging data alongside PROMs (showing moderate disability) and EHR data (indicating sedentary lifestyle). The Multimodal Profiling Module synthesizes these inputs into a unified profile indicating a *low-to-moderate risk phenotype*. The LLM Synthesis Module, guided by RAG, generates an evidence-anchored draft report favoring conservative, exercise-based rehabilitation. The radiologist reviews the draft alongside visual AI overlays, edits the language for precision, and finalizes the report. This workflow reduces unwarranted diagnostic escalation, standardizes reporting language, and shortens the time from imaging acquisition to a personalized care recommendation.

Scenario 2: Acute-on-Chronic Pain in a Manual Laborer with Complex Biopsychosocial Factors A 62-year-old manual laborer presents with acute exacerbation of chronic pain. MRI reveals multilevel degeneration and Modic type II changes ¹.

- **Imaging-Only Workflow:** The pronounced structural pathology may disproportionately bias clinical reasoning toward early surgical referral, potentially overlooking modifiable non-structural factors.
- **Multimodal Framework Workflow:** The framework integrates the severe imaging findings with PROMs revealing high disability, maladaptive coping strategies, poor psychological quality of life, and EHR-documented comorbidities (obesity, diabetes). The synthesized profile highlights a *high-risk, complex biopsychosocial phenotype*. The LLM module surfaces guideline-based

¹Modic changes are specific vertebral endplate and bone marrow signal alterations on MRI associated with disc degeneration. Type II is characterized by fatty replacement (hyperintensity on T1-weighted images), indicating a chronic, stable degenerative state [70].

recommendations for pre-habilitation: metabolic optimization, targeted physiotherapy, and psychological interventions to address coping and well-being. This equips the surgeon to make a more informed, person-centered decision, whether to proceed with surgery immediately or to optimize the patient’s condition first, demonstrating how multimodal integration expands, rather than restricts, therapeutic options.

Scenario 3: Safety-Critical Surgical Escalation A 38-year-old patient presents with acute bilateral lower limb weakness and severe radicular pain, suggestive of cauda equina syndrome.

- **Conventional Workflow:** Relies on the clinician to recognize red flags, manually prioritize the MRI read, and initiate urgent referrals, a process susceptible to delays under high clinical load.
- **Multimodal Framework Workflow:** The case is flagged by the system’s triage logic. The Computer Vision Module processes the MRI with high priority, identifying a high-confidence compressive lesion. Simultaneously, the LLM module, constrained by RAG to acute care guidelines, generates an urgent alert with prepopulated referral orders for neurosurgical review. The clinician receives an immediate notification; all actions (acknowledge, modify, act) are captured in an immutable provenance log. This scenario demonstrates the framework capacity to accelerate time-sensitive care while preserving clinician authority and ensuring full decision traceability for safety auditing and quality improvement.

Evaluation Metrics and Forward Path To support evaluation and prospective deployment, each vignette suggests measurable endpoints for future validation: changes in time-to-report and time-to-referral, shifts in referral appropriateness or downstream resource utilization, clinician acceptance and edit rates, safety metrics for urgent alerts (for example time-to-escalation and false-alarm rates), and the completeness of provenance metadata (fraction of drafts with full guideline anchors). Together, these scenarios are intended to make the system anticipated behaviors concrete and to guide subsequent validation work that would be necessary before clinical implementation.

Collectively, these vignettes transition the discussion from architectural theory to potential clinical impact. They provide a concrete basis for designing the validation studies necessary to advance the framework from a conceptual blueprint toward future clinical implementation.

The preceding vignettes are summarized in Table 3.1, which contrasts the current practice with the proposed AI-enhanced workflow for each case synthesized for clarity.

Table 3.1: Representative clinical use cases illustrating potential workflow improvement after modular AI framework deployment

Use Case	Current Clinical Practice	AI-Enabled Enhancement
Case 1: Chronic Low Back Pain	<i>Imaging-driven reports; limited context integration.</i> <ul style="list-style-type: none"> Structured image grades require manual contextualization. High inter-reader variability and time-intensive reporting. 	<i>Context-aware multimodal synthesis.</i> <ul style="list-style-type: none"> Fuses spine imaging, EHR, and PROMs for risk-graded, evidence-linked drafts. Expected: ↓ time-to-report, ↑ consistency, ↑ efficiency.
<i>Metrics: time-to-report, edit frequency, report concordance</i>		
Case 2: Surgical Bias Mitigation	<i>Structure-dominant decision pathways.</i> <ul style="list-style-type: none"> Imaging severity outweighs psychosocial context. Limited support for pre-operative optimization. 	<i>Risk-adjusted, evidence-grounded recommendations.</i> <ul style="list-style-type: none"> Multimodal analytics highlight comorbidities and PROM-derived psychosocial risk. Expected: ↓ inappropriate referrals, ↑ patient optimization.
<i>Metrics: referral appropriateness, optimization rate, clinician acceptance</i>		
Case 3: Urgent Escalation	<i>Manual escalation; inconsistent traceability.</i> <ul style="list-style-type: none"> Relies on clinician vigilance for urgent routing. Limited auditability of decisions and timing. 	<i>Automated, auditable escalation.</i> <ul style="list-style-type: none"> Prioritized processing triggers guideline-aligned alerts with prepopulated orders. Immutable provenance log tracks clinician response. Expected: ↓ time-to-escalation, ↑ safety, ↑ accountability.
<i>Metrics: time-to-escalation, false-alarm rate, provenance completeness</i>		

3.5 Discussion

This chapter has presented a modular, integrative framework designed to translate AI research into clinically trustworthy tools for spine care. The proposed

architecture, spanning a standards-first data middleware, a multimodal AI core, an evidence-anchored LLM/RAG layer, and human-centered governance, provides a concrete blueprint for building safe, auditable, and incrementally deployable clinical decision support. The following discussion synthesizes its principal contributions, positions it within the research landscape, addresses critical limitations, and outlines a path toward validation.

3.5.1 Principal Contributions and Integrative Novelty

This work makes five key contributions to the field of clinical AI for spine care:

- (1) **Multimodal Integration Paradigm:** It advances a model where PROMs, EHR data, and imaging are treated as co-equal inputs, enabling risk profiles that reflect the true multifactorial nature of low back pain.
- (2) **Evidence-Anchored Language Layer:** It proposes a practical LLM/RAG architecture to generate citation-tethered draft reports and recommendations, explicitly mitigating the safety risks of general-purpose generative AI.
- (3) **Human-in-the-Loop Governance:** It embeds clinician review, editable outputs, and structured feedback as core design principles, ensuring augmentation rather than replacement of clinical expertise.
- (4) **Deployable System Architecture:** It specifies a model-management and middleware architecture that foregrounds reproducibility, staged deployment (shadow → pilot → production), and regulatory traceability.
- (5) **Validation Blueprint:** It outlines a forward-looking evaluation strategy, aligning technical, workflow, and safety metrics with each phase of a future system rollout.

The primary innovation lies not in these individual components, which align with established best practices, but in their integration into a cohesive, modular framework. This unification of multimodal fusion, evidence-anchored communication, and governed human-AI collaboration within a single architectural blueprint represents the core conceptual advance of this chapter.

3.5.2 Differentiation from Existing Work

As established in Chapter 2, prior AI research in spine care has largely produced valuable but isolated models for imaging, prediction, or PROMs analysis. The framework proposed here builds on that foundation to envision an interoperable ecosystem for clinical synthesis. It moves beyond single-purpose algorithms toward a standard-first design that fuses multimodal data and uses a retrieval-augmented

language layer to generate evidence-linked clinical narratives. Furthermore, it introduces a structured mechanism where clinician interactions provide auditable feedback for continuous, traceable system refinement. The proposal thus extends the landscape from isolated analytical tools toward a cohesive, accountable system designed for operational integration across the full diagnostic and treatment workflow.

3.5.3 Critical Limitations and Mitigation Strategies

While the proposed architecture provides a comprehensive blueprint, translating it into a robust, production-ready system faces several nontrivial challenges that must be acknowledged. Data quality, heterogeneity, and inconsistent data completeness in real-world settings pose a persistent threat to model generalizability and reliability. Mitigating this requires more than technical solutions; it demands disciplined data governance, site-specific harmonization protocols, and the implementation of conservative abstention rules where inputs are incomplete or of low confidence. The ambition of a continuously learning system, while desirable, introduces significant operational and regulatory complexity. A practical path forward involves predetermined change-control plans, validation gates tied to performance metrics, and staged rollouts that mirror regulatory expectations for Software as a Medical Device (SaMD). Perhaps the most scrutinized component will be the LLM-based synthesis layer. Even with a Retrieval-Augmented Generation (RAG) architecture constraining outputs to a curated evidence base, risks of hallucination, subtle bias, and over-reliance remain. These demand a multi-layered safety strategy incorporating retrieval-confidence thresholds, rule-based refusal policies for high-risk recommendations, mandatory human verification pipelines, and systematic, scenario-based testing. Finally, the infrastructure and computational costs of deploying such an integrated ecosystem raise valid concerns about equity and accessibility. Strategies such as federated learning for model development, offering lightweight model variants, and designing tiered deployment packages can help reduce barriers for resource-limited clinical sites, ensuring the benefits of such systems are broadly adopted.

3.5.4 Conclusion and Forward Path

This chapter articulated a comprehensive architectural vision to bridge the integration gap in spine care. By deriving design principles from clinical needs and proposing a modular system for multimodal synthesis, it provides a foundational blueprint for human-centered AI decision support. The framework’s core value

lies in unifying anatomical, clinical, and psychosocial data, traditionally analyzed in silos, into a coherent, actionable narrative for personalized assessment.

To assess initial feasibility, a limited proof-of-concept study implemented key framework components in a controlled setting, evaluating workflow integration, module interoperability, and clinician interaction with prototype outputs. These findings, while not formal validation, provided crucial insights into practical integration hurdles and informed architectural refinements and the proposed evaluation roadmap.

The staged roadmap, progressing from technical validation and workflow simulation to pilot implementation and prospective multicenter studies, outlines a responsible path for future development, systematically addressing the limitations noted above. The following chapters build upon this blueprint. Chapters 4 and 5 develop and validate the core Computer Vision and Multimodal Profiling modules. Chapter 6 presents the methodology and results of the integrated proof-of-concept study, offering a preliminary examination of the framework’s operational feasibility and impact on simulated clinical workflows.

3.5.5 Open research directions

Several research and engineering avenues remain open. Federated and privacy-preserving learning strategies could improve generalizability while protecting local data sovereignty. Robust, domain-specific evaluation frameworks for LLM outputs in clinical contexts are needed to quantify fidelity, hallucination risk, and clinical utility. Integrating longitudinal and real-world outcome signals (wearables, remote PROMs) could extend the system toward prognostic and long-term care planning. Finally, sociotechnical research into clinician acceptance, human-AI collaboration ergonomics, and equitable deployment will be central to adoption.

This chapter presented a modular AI framework for personalized spine care, integrating three core modules for comprehensive clinical decision support. By presenting this framework early in the thesis, we aim to deliver the complete vision before diving into implementation details. Chapter 4 now establishes the baseline for Module 1 through external validation of SpineNetV2, identifying critical gaps that Module 1 must address, particularly the need for multiview integration to accurately assess cross-sectional pathologies.

4

LumbarAI: Clinically-Aligned AI Model for Automated Spine MRI Analysis

This chapter establishes and develops **Module 1 (Computer Vision Module)** of the framework presented in Chapter 3. The chapter follows a two-part structure: first, we establish a performance baseline by externally validating SpineNetV2 [48], the current state-of-the-art open-source AI model for automated spine MRI grading, identifying specific gaps that Module 1 must address. We then introduce and comprehensively evaluate LumbarAI, a novel curriculum learning framework designed to address these exact gaps through multiview integration and clinical workflow alignment.

4.1 Introduction

The integrated AI framework introduced in Chapter 3 relies on a robust Computer Vision Module capable of automated, reproducible interpretation of spine MRI. This module must provide structured and clinically meaningful anatomical descriptors to support the framework’s core multimodal synthesis capabilities. Its reliability must extend beyond controlled research environments to the heterogeneous imaging conditions, label inconsistencies, and acquisition variability inherent to real-world clinical practice.

4.1.1 Radiological Features in Lumbar MRI Analysis

The development of a clinically viable Computer Vision Module requires a precise definition of its target outputs. This work focuses on automating the grading of a core set of radiological features that are fundamental to the clinical assessment of degenerative lumbar spine disease. These features were selected for their established diagnostic relevance, consistent observability on standard MRI sequences, and direct utility in forming the structured anatomical profile required for downstream multimodal synthesis within the integrative framework (Chapter 3).

The targeted features encompass:

- **Disc Degeneration and Morphology:** Pfirrmann grading (1–5) [13] and disc narrowing (1–4).
- **Canal and Neural Foramen Compromise:** Central canal stenosis (1–4) and left/right foraminal stenosis (0–1).
- **Vertebral Body and Endplate Pathology:** Spondylolisthesis (0–1), endplate defects (0–1), and Modic-type marrow changes (0–1).
- **Herniation:** Disc herniation (0–1).

Accurate grading of these 11 features across lumbar levels (L1–L2 to L5–S1) provides the essential structured anatomical data layer. Table 4.1 details the clinical definitions and grading scales for each pathology. Figure 4.1(a) provides illustrative T2-weighted sagittal slices for pathologies assessable from the sagittal view alone (Pfirrmann grading, narrowing, spondylolisthesis, upper endplate defects, lower endplate defects, upper marrow changes, lower marrow changes). Figure 4.1(b) presents illustrative T2-weighted images for pathologies requiring both sagittal and axial views for comprehensive assessment (central canal stenosis, foraminal stenosis left, foraminal stenosis right, herniation). Together, these figures establish the visual lexicon for the automated analysis tasks undertaken in this chapter.

4.1.2 Research Trajectory: From External Validation to Novel Architecture

To determine whether existing state-of-the-art approaches could fulfil the role of the Computer Vision Module, we first conducted a rigorous external validation of SpineNetV2 [48], a model designed to grade the features described above using sagittal T2w MRI. Beyond prior results obtained on curated datasets, we performed a full-feature evaluation on a real-world clinical dataset collected at our

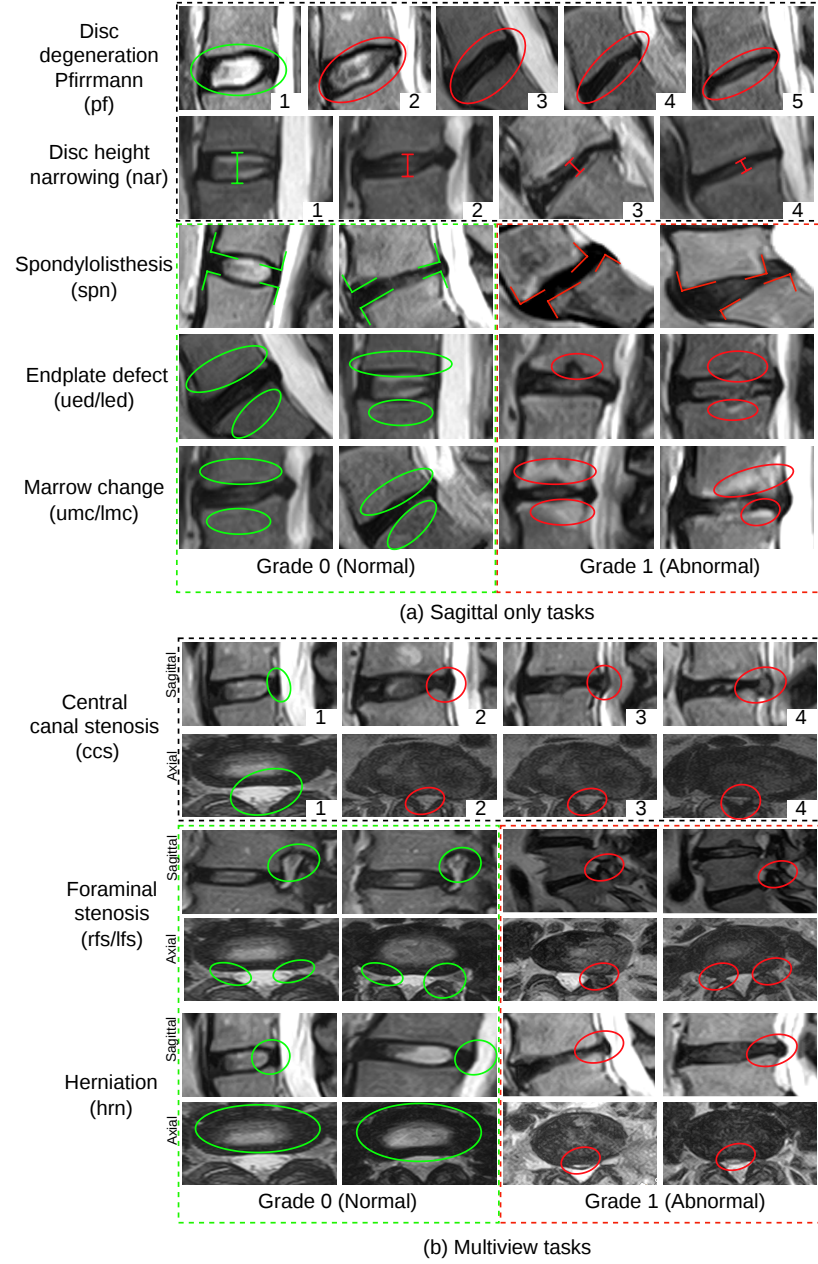


Figure 4.1: *T2-weighted MRI examples for all grading levels of each evaluated radiological feature. (a) Sagittal only tasks (features graded from sagittal views): Pfirrmann grade (1-5), disc height narrowing (1-4), and binary assessments (0/1) for spondylolisthesis, endplate defect, and marrow change. (b) Multiview tasks (features graded using combined sagittal and axial views): central canal stenosis (1-4) and binary assessments for foraminal stenosis and herniation. Green = Normal (0), Red = Pathological/Positive*

Table 4.1: *Standard radiological features for degenerative lumbar spine assessment.*

Feature	Clinical/Radiological Definition	Grading Scale
Pfirrmann	Disc degeneration classification based on nucleus signal, structural homogeneity, and disc height.	1 (normal) to 5 (severe)
Disc Narrowing	Reduction in disc height relative to adjacent reference levels.	1 (normal) to 4 (extreme)
Central Canal Stenosis	Constriction of the central spinal canal with potential thecal sac or cauda equina compression.	1 (none) to 4 (severe)
Spondylolisthesis	Anterior/posterior displacement of a vertebra (Meyerding grades I–V).	0 (absent), 1 (present)
Endplate Defect	Localized irregularity, erosion, or fracture of the vertebral endplate.	0 (absent), 1 (present)
Marrow Change	Vertebral marrow signal alterations adjacent to the endplate.	0 (absent), 1 (present)
Foraminal Stenosis	Narrowing of the neural foramen affecting the exiting nerve root.	0 (absent), 1 (mild–severe)
Disc Herniation	Displacement of disc material beyond the intervertebral disc space.	0 (absent), 1 (present)

local clinic, *X-Ray Service s.r.l.*, hereafter referred to as the **XrayService AI-Assisted Dataset (XSA)**. This evaluation served two purposes: (1) establishing a realistic performance baseline under genuine deployment conditions, and (2) identifying which anatomical features can be reliably inferred from sagittal views and which require multi-view analysis.

The insights from this validation directly informed the second phase of our work: the development of a new, multi-view transformer-based architecture designed to meet the Computer Vision Module’s requirements. This development followed a deliberate, stepwise methodology. First, we reproduced a sagittal-only baseline mirroring SpineNetV2’s functional scope using a modern transformer backbone, ensuring compatibility with this dissertation’s broader architectural principles. This baseline served as both a technical reference and a pedagogical “teacher” for subsequent learning stages.

Building on this foundation, we systematically integrated axial information through a complementary axial-view expert. To ensure stable multi-view fusion and avoid catastrophic forgetting, we employed a staged curriculum learning strategy: early phases prioritized robust view-specific expertise via teacher-student guidance, followed by a controlled transition to end-to-end joint optimization where sagittal and axial signals were seamlessly integrated. The resulting system, *LumbarAI*, embodies a clinically aligned pipeline that reflects standard radiological practice, evaluating longitudinal structures in sagittal planes and cross-sectional pathology in axial views.

This chapter therefore presents two major contributions, (1) the external validation of SpineNetV2 and (2) the design, training, and evaluation of the multi-view LumbarAI system, as sequential steps in a unified research trajectory. Together, they demonstrate both the empirical necessity and the technical feasibility of an automated MRI analysis system capable of supporting the integrative framework proposed in Chapter 3. Part I details the validation study and its identified limitations, while Part II introduces the LumbarAI architecture, experimental setup, results, and discussion.

4.2 Part I: Benchmarking the State-of-the-Art model and Identifying Gaps (External Validation of SpineNetV2)

4.2.1 Rationale for Benchmarking

To define the performance baseline and precise design goals for the framework’s Computer Vision Module, we began with a rigorous external validation of the state-of-the-art. This first step was essential to move beyond theoretical limitations and quantify, using our own clinical data, which aspects of automated MRI interpretation were already robust and which required a novel architectural solution. We selected SpineNetV2 [48] as the benchmark, a leading sagittal-only deep learning model for multi-feature lumbar MRI grading.

While two prior external validations exist [91, 90], their scope was limited to subsets of SpineNetV2’s 11 radiological features. Our study provides the first comprehensive validation on its complete pathology set. This approach allowed us to empirically distinguish between pathologies well-served by sagittal-only analysis and those demonstrably requiring multi-view reasoning, thereby establishing the empirical foundation for the LumbarAI architecture developed in Part II.

4.2.2 Methods

Our validation protocol (Figure 4.2) was designed to mirror a realistic deployment scenario, prioritizing scientific rigor and clinical relevance.

Dataset and Preprocessing. We used the **XrayService AI-Assisted Dataset (XSA)** [129], comprising 1,615 spine MRI scans. To create a cohort relevant to SpineNetV2’s lumbar focus, we applied the following exclusion criteria: thoracic and cervical scans (removing non-lumbar studies), pediatric patients (age < 14

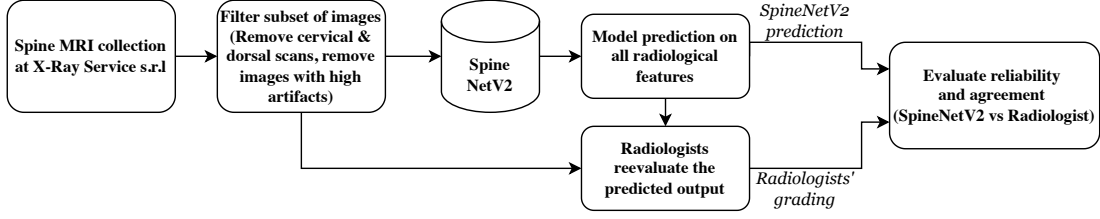


Figure 4.2: *External validation workflow for SpineNetV2. This process was designed to produce a reliable performance baseline under conditions mimicking real-world clinical integration.*

years), and scans with significant image artifacts. After exclusions, the final cohort consisted of 1,747 lumbosacral discs from 353 patients (mean age 54 ± 15.4 years; 44.5% female). All data were fully anonymized in compliance with GDPR, ensuring patient privacy while maintaining reproducibility. The acquisition protocol is detailed in Table 4.2.

Table 4.2: *MRI acquisition protocol for the XrayService AI-Assisted Dataset (XSA).*

Parameter	Sagittal	Axial
Scanner	HITACHI Airis Vento (0.3T)	
Sequence	2D Spin Echo	
Repetition Time (ms)	3135	5000
Echo Time (ms)	108	110
Slice Spacing (mm)	4.7	5.0

AI-Human Collaborative Annotation Protocol. To establish a reliable reference standard, we employed an AI-assisted radiologist review process. Two expert radiologists (17 and 38 years of experience) first independently annotated 60 cases to establish inter-rater reliability, demonstrating strong agreement (Cohen’s $\kappa > 0.73$ for all features; see Table 4.3). Given this high concordance, the remaining cases were reviewed by a single radiologist who corrected SpineNetV2’s predictions. This protocol balanced rigor with efficiency, leveraging AI to pre-populate gradings while ensuring final labels reflected expert clinical judgment.

Evaluation. We evaluated SpineNetV2’s predictions against the radiologist-corrected gold standard using a comprehensive suite of metrics appropriate for multi-class and binary tasks: Accuracy, Balanced Accuracy, Precision, Recall, F1-score, Brier Score Loss (BSL), Cohen’s Kappa (κ), Lin’s Concordance Correlation Coefficient (LCCC), and Matthews Correlation Coefficient (MCC). This multi-faceted approach provided a nuanced view of performance, distinguishing between

Table 4.3: *Inter-rater reliability and agreement evaluation of two expert radiologists on 289 discs acquired from 60 patients.*

Pathology	BSL	κ	LCCC	MCC
Pfirrmann	0.055	0.823	0.973	0.823
Narrowing	0.055	0.838	0.966	0.839
Central Canal Stenosis	0.0138	0.851	1.00	0.851
Spondylolisthesis	0.007	0.930	1.00	0.930
Upper Endplate Defect	0.031	0.835	0.908	0.835
Lower Endplate Defect	0.059	0.734	0.828	0.736
Upper Marrow	0.066	0.769	0.842	0.770
Lower Marrow	0.083	0.795	0.797	0.695
Left Foraminal Stenosis	0.038	0.890	0.990	0.890
Right Foraminal Stenosis	0.035	0.901	0.980	0.901
Herniation	0.045	0.900	0.962	0.900

Abbreviations: BSL: Brier Score Loss, κ : Cohen’s kappa, LCCC: Lin’s Concordance Correlation Coefficient, MCC: Matthews Correlation Coefficient

overall correctness (Accuracy), class imbalance robustness (Balanced Accuracy, MCC), and agreement accounting for chance (κ , LCCC).

4.2.3 Agreement and Reliability Metrics

In assessing the external validation of a deep learning model, inter-rater agreement and reliability metrics [145] play a pivotal role, offering a comprehensive insight into the concordance between the model predictions and radiologists’ annotations. This study employs the following metrics to evaluate the level of agreement and assess the model’s performance:

1. **Accuracy** measures the overall correctness of predictions:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}; \quad (4.1)$$

2. **Precision** evaluates the proportion of true positives among all predicted positives:

$$\text{Precision} = \frac{TP}{(TP + FP)}; \quad (4.2)$$

where TP are True Positives, and FP are False Positives.

3. **Recall** evaluates the proportion of true positives among all actual positive samples in the dataset. Recall can also be called *Sensitivity*, or *True positive rate*:

$$\text{Recall} = \frac{TP}{(TP + FN)}; \quad (4.3)$$

where FN are False Negatives.

4. **F1 score** is the harmonic mean of precision and recall:

$$F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot TP}{2TP + FP + FN}; \quad (4.4)$$

5. **Balanced accuracy** adjusts accuracy by averaging the sensitivity and specificity of the model:

$$\text{Balanced accuracy} = \frac{(\text{Sensitivity} + \text{Specificity})}{2} = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right), \quad (4.5)$$

where: TN refers to True Negatives, *Sensitivity* refers to the ‘true positive rate’ which represents the percentage of positive cases that the model correctly identifies. *Specificity*, on the other hand, denotes the ‘true negative rate’ indicating the percentage of negative cases that the model correctly identifies.

6. **Matthew’s Correlation Coefficient (MCC)** captures the correlation between predicted and observed binary classifications:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}; \quad (4.6)$$

7. **Lin’s concordance correlation coefficient (CCC)** is the concordance between a model prediction (Y) and a gold standard test or measurement (X):

$$CCC = \frac{2C_{XY}}{C_{XX} + C_{YY} + (\mu_x - \mu_y)^2}; \quad (4.7)$$

where:

- C_{XY} is the sample covariance between predicted outcomes (Y) and actual values (X);
- C_{XX} and C_{YY} are the sample variances of X and Y, respectively;
- μ_x and μ_y are the means of the actual values and predicted outcomes, respectively.

8. **Brier Score Loss** is a measure of the mean square difference between the predicted and actual probability:

$$\text{Brier Score Loss} = \frac{1}{N} \sum_{i=1}^N (P_i - O_i)^2, \quad (4.8)$$

where N is the number of observations, P_i is the predicted probability, and O_i is the observed outcome (0 or 1).

9. **Cohen’s Kappa** is a measure of inter-rater reliability or the agreement between two dependent categorical samples:

$$\kappa = \frac{(P_o - P_e)}{1 - P_e}, \quad (4.9)$$

where:

- P_o is the Observed Agreement: $P_o = \frac{TP+TN}{N}$;
- P_e is the Expected Agreement: $P_e = \frac{(TP+FP)*(TP+FN)+(TN+FN)*(TN+FP)}{N^2}$,
on a total of N observations.

4.2.4 Results and Identified Gaps

The external validation yielded a definitive and clinically interpretable result: the performance of a sagittal-only model is intrinsically bounded by pathology type. A clear dichotomy emerged between features assessable from the sagittal plane and those requiring axial visualization, directly quantifying the limitation that motivates a multi-view solution.

Competent Performance on Sagittal-Visible Pathologies Consistent with clinical reasoning and the model’s design, SpineNetV2 achieved strong agreement with radiologist labels for pathologies where sagittal views are diagnostically primary:

- **High Agreement:** Robust performance was observed for Pfirrmann grading ($\kappa = 0.74$), disc narrowing ($\kappa = 0.80$), central canal stenosis ($\kappa = 0.75$), and endplate/marrow changes ($\kappa = 0.73$ – 0.76).
- **Interpretation:** This result establishes a performance ceiling for sagittal-only inference on degenerative and alignment-related features, confirming that such an architecture can be a competent component for these specific tasks.

Substantial Degradation on Axial-Dependent Pathologies Critically, performance deteriorated for pathologies where axial sequences are the clinical standard for definitive assessment:

- **Weak Agreement:** Foraminal stenosis showed only moderate-to-weak agreement (Left: $\kappa = 0.47$, Right: $\kappa = 0.46$). Herniation agreement was moderate ($\kappa = 0.55$).
- **Interpretation:** These suboptimal kappa values quantitatively demonstrate the fundamental insufficiency of sagittal-only data for evaluating cross-sectional anatomy. The model cannot replicate the radiologist’s essential step of correlating sagittal findings with axial confirmation, particularly for neural foraminal and nerve root assessment.

The contrast is summarized in Table 4.4. This empirical dichotomy, strong sagittal versus weak axial performance, constituted the pivotal insight from Phase

I, providing the concrete evidence necessary to justify and guide the architectural innovations of Phase II.

Note: Throughout this chapter, the following pathology abbreviations are used: **pf** (Pfirrmann grading), **nar** (disc narrowing), **ccs** (central canal stenosis), **spn** (spondylolisthesis), **ued** (upper endplate defect), **led** (lower endplate defect), **umc** (upper marrow change), **lmc** (lower marrow change), **fsl** (left foraminal stenosis), **fsr** (right foraminal stenosis), and **hrn** (herniation).

Table 4.4: Key agreement metrics (Cohen’s κ , Balanced Accuracy) from the SpineNetV2 validation on the XSA dataset.

Feature	pf	nar	ccs	spn	ued	led	umc	lmc	fsl	fsr	hrn
κ	0.74	0.80	0.75	0.71	0.73	0.75	0.76	0.73	<i>0.47</i>	<i>0.46</i>	<i>0.55</i>
Bal. Acc.	0.79	0.85	0.78	0.99	0.85	0.86	0.90	0.89	<i>0.70</i>	<i>0.69</i>	<i>0.76</i>

Note: **Bold** denotes highest performance; *italic* highlights critical gaps.

Note on comprehensive analysis: The detailed methodology, extended results including confusion matrices, distribution analyses, and a full discussion of clinical implications are presented in the standalone publication resulting from this work [47]. This chapter provides the synthesized findings essential for motivating and framing the development of LumbarAI.

4.2.5 Derived Requirements for the Imaging Module

The quantitative findings from Part I were translated into a set of explicit design requirements for the framework’s Computer Vision module. Each requirement is directly grounded in an empirically observed limitation and its clinical interpretation.

1. Multi-View Capability.

Evidence: Moderate agreement for foraminal stenosis ($\kappa \approx 0.46$) and disc herniation ($\kappa \approx 0.55$), both of which are cross-sectional pathologies.

Requirement: The model must incorporate axial MRI information in addition to sagittal sequences to adequately capture transverse anatomical detail.

2. Task-View Alignment.

Evidence: High agreement for sagittal-visible features contrasted with reduced performance for axial-dependent findings.

Requirement: The architecture should explicitly align diagnostic tasks with their anatomically appropriate imaging planes, rather than treating all views as interchangeable inputs.

3. Preservation of Sagittal Performance.

Evidence: Strong agreement for sagittal-based assessments (e.g., Pfirrmann

grading; $\kappa = 0.71\text{--}0.80$).

Requirement: Any multi-view extension must preserve sagittal-task performance as a prerequisite for clinical credibility.

4. **Stable Multi-View Integration.**

Evidence: Known risks of performance degradation when introducing additional modalities without structured learning strategies [96].

Requirement: Multi-view fusion must be guided and staged to ensure stable, synergistic learning rather than naïve feature concatenation.

Together, these requirements define the functional specification for **LumbarAI**. They motivated a modular architecture with view-specialized components and a controlled fusion strategy, detailed in Section 4.3.

4.3 Part II: LumbarAI – A Curriculum Learning Framework

4.3.1 The LumbarAI Model: Rationale and Architecture

4.3.1.1 Core Hypothesis and Design Philosophy

The limitations identified in Part I reflect more than a data or performance issue; they expose a conceptual mismatch between conventional deep learning architectures and the cognitive process of radiological interpretation. Sagittal-only models implicitly assume that diagnostic information is uniformly accessible within a single projection, an assumption that fails for pathologies whose definition is inherently cross-sectional.

Radiologists, by contrast, engage in a structured, sequential reasoning process when interpreting spine MRI. Sagittal images are first used to establish global context, disc degeneration, alignment, and level localization, while axial slices are subsequently interrogated to assess focal compromise of neural structures [146, 147]. This view-dependent strategy is not incidental but foundational to expert interpretation. Accordingly, we seek to design an AI model that more closely reflects the sequential, view-dependent workflow used in radiological interpretation, in which sagittal images provide global anatomical context and axial views are subsequently examined for focal cross-sectional detail. Rather than collapsing information from multiple planes into a single representation, LumbarAI is structured to process and integrate complementary views in a staged manner that mirrors this diagnostic sequence. Under this design philosophy, LumbarAI

is intended to incorporate elements of radiological reasoning into the model architecture itself. By aligning task structure and view-specific processing with established clinical practice, the framework aims to support clinically coherent interpretation while remaining compatible with existing reporting workflows.

4.3.1.2 Model Architecture: Workflow-Aligned Hybrid Architecture

Design Rationale and Task Assignment: LumbarAI implements a hybrid architecture that computationally mirrors the standard radiological workflow (Figure 4.3), with task assignment following clinical practice. The Sagittal-Only Expert processes sagittal views for longitudinal pathologies (7 tasks: Pfirrmann grading, disc narrowing, spondylolisthesis, endplate/marrow changes (upper & lower)) [13, 148]. The Multi-View Expert processes both sagittal and axial views for cross-sectional pathologies (4 tasks: central canal stenosis, foraminal stenosis (left & right), disc herniation) [93, 94]. This dual-pathway design directly addresses the workflow misalignment identified in Part I.

Input Specifications: Both experts operate on 3D disc-level stacks extracted from preprocessed MRI series. The Sagittal-Only Expert receives sagittal stacks of dimensions $[B, 1, 9, 112, 224]$ (batch size, channels, slices, height, width), while the Multi-View Expert receives paired sagittal stacks alongside the corresponding axial stacks of dimensions $[B, 1, 4, 100, 100]$. Complete details of the data preprocessing pipeline are provided in Section 4.3.4.2.

4.3.1.3 Swin3D Transformer Backbone

Both expert models are built upon the Swin3D Transformer architecture [84], an extension of the 2D Swin Transformer model [83] to 3D volumetric data. For a per-disc stack of sagittal slices, 3D patch embedding tokenizes the volume. The model employs windowed self-attention within non-overlapping 3D windows, with a shifted-window scheme between blocks to enable cross-window context aggregation. This design efficiently captures both local anatomical detail and longer-range contextual patterns across the slice stack, which is crucial for pathologies like stenosis or herniation that manifest over multiple contiguous slices.

4.3.1.4 Multi-View Feature Fusion

The Multi-View Expert employs a gated late-fusion mechanism (Figure 4.3(e)–(g)). Normalized feature vectors from the sagittal ($\mathbf{f}_{sag}$) and axial ($\mathbf{f}_{axial}$) encoders are first projected to a shared dimensionality of 256 via separate linear layers with LayerNorm, GELU activation, and dropout (0.1). These projected features are

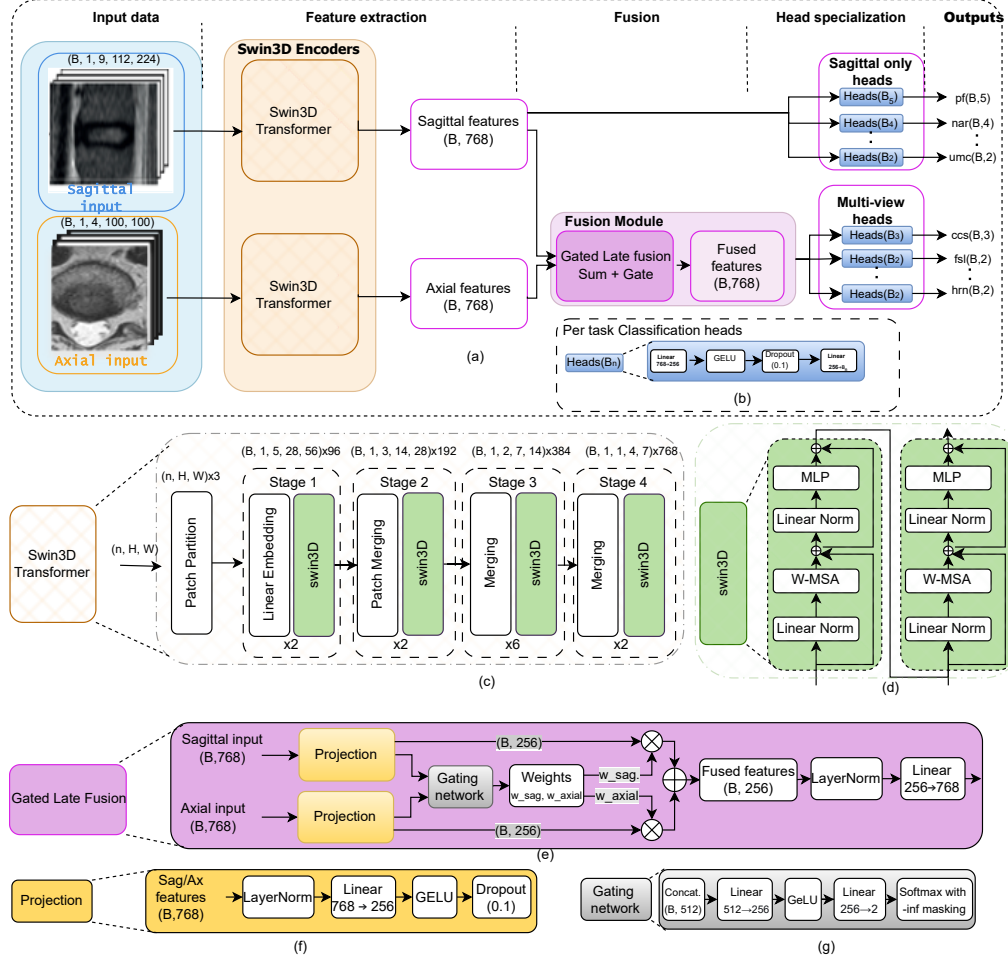


Figure 4.3: Comprehensive overview of the LumbarAI hybrid architecture. (a) High-level end-to-end architecture employing dual specialized pathways: a Sagittal-Only Expert optimized for longitudinal pathology assessment (Pfirschmann grade, disc narrowing, spondylolisthesis, endplate defects, and marrow changes) and a Multi-View Expert that integrates sagittal and axial features through gated fusion for cross-sectional analysis (central canal stenosis, foraminal stenosis, and disc herniation). (b) Task-specific classification heads applied to all pathology tasks, with output dimensions N corresponding to severity classes (binary: $N = 2$; multi-class: $N > 2$). (c–d) Detailed Swin3D transformer backbone architecture adapted from [84] for volumetric feature extraction. (e) Gated late fusion network enabling adaptive weighting of sagittal and axial feature contributions. (f) View-specific projection networks within the fusion module. (g) Attention-based gating mechanism that computes softmax-normalized weights for view importance, with $-\infty$ masking for robust handling of missing views.

concatenated and passed through a two-layer gating network ($512 \rightarrow 256 \rightarrow 2$) with softmax activation. The resulting attention weights $\mathbf{w} = [w_{sag}, w_{axial}]$ are used for weighted combination:

$$\mathbf{f}_{\text{fused}} = w_{sag} \odot \mathbf{f}_{sag} + w_{axial} \odot \mathbf{f}_{axial}.$$

The softmax layer incorporates *-inf* masking to handle missing axial views gracefully, allowing the model to dynamically learn the relative importance of each view for specific pathologies. It is worth noting that the gated fusion mechanism described above can be interpreted as a form of attention weighting, where the learned weights w_{sag} and w_{axial} represent the relative importance of each view for the current prediction. This differs from conventional attention-based fusion in two key respects: (1) the weights are computed from the concatenated features rather than through a query-key-value mechanism, and (2) the gating network produces view-level weights rather than spatial attention maps. In ablation studies (Section 4.3.7), we compare this approach against alternative fusion strategies including sum, concatenation, and a true attention-based mechanism where features from each view are weighted by spatial attention maps before fusion.

4.3.1.5 Model Implementation and Complexity

Both expert models use identical Swin3D-Tiny backbone configurations. The Sagittal-Only Expert utilizes a single encoder, while the Multi-View Expert employs two separate encoders followed by the fusion module. Detailed parameter counts, training times, and hardware specifications are provided in Section 4.3.4.4.

4.3.2 Four-Phase Curriculum Learning Strategy

Our training strategy begins by establishing a robust sagittal-only foundation: the Sagittal-Only Expert was first pre-trained independently using class-weighted cross-entropy loss on all sagittal-view tasks. With this foundation established, we then face the challenge of multi-view integration, which presents a classic stability-plasticity dilemma [96]: how to incorporate complementary axial information without disrupting the learned sagittal representations. To address this, we employ a four-phase curriculum learning strategy for the Multi-View Expert, inspired by principles of infants educational psychology [149] and as discussed in [150]. This strategy gradually increases task complexity through sequential phases as follow:

- **Phase 1: Axial Branch Learning** – Introduce axial processing while keeping sagittal backbone frozen. Trainable: axial encoder, fusion module, multi-view heads.
- **Phase 2: Fusion Unfreezing** – Align sagittal/axial spaces. Unfreeze: fusion module, axial encoder, sagittal projections (LR: 1×10^{-6}). Add contrastive consistency loss.
- **Phase 3: Joint Fine-Tuning** – Refine aligned system. Unfreeze both encoders (LRs: Sag 1×10^{-6} , Ax 5×10^{-5} , Fus 1×10^{-4}). Enable task-specific gating.

- **Phase 4: Head Specialization** – Calibrate final layers. Freeze backbone/fusion; train only classification heads. Weighted loss prioritizes critical pathologies.

This progressive approach allows the model to solidify foundational sagittal knowledge before introducing the axial domain, culminating in specialized multi-view task optimization.

4.3.3 Loss Functions

4.3.3.1 Single-view (Sagittal-only) Loss

For sagittal-only tasks, the loss function employs standard cross-entropy classification:

$$\mathcal{L}_{\text{sag}}(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_{i=1}^B \sum_{c=1}^C y_{i,c} \log(\hat{y}_{i,c}) \quad (4.10)$$

where B is the batch size, C is the number of classes, $\mathbf{y} \in \mathbb{R}^{B \times C}$ are the ground truth one-hot encoded labels, and $\hat{\mathbf{y}} \in \mathbb{R}^{B \times C}$ are the predicted class probabilities.

4.3.3.2 Multiview Fusion Losses

The multiview fusion incorporates three loss components:

- (1) **Contrastive Loss** (temperature $\tau = 0.07$):

$$\mathcal{L}_{\text{contrastive}} = - \log \frac{\exp(\text{sim}(\mathbf{f}_{\text{sag}}, \mathbf{f}_{\text{ax}})/\tau)}{\sum_j \exp(\text{sim}(\mathbf{f}_{\text{sag}}, \mathbf{f}_{\text{ax}}^{(j)})/\tau)} \quad (4.11)$$

where $\mathbf{f}_{\text{sag}}, \mathbf{f}_{\text{ax}} \in \mathbb{R}^d$ are feature embeddings from sagittal and axial views of the same disc level, $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} / (|\mathbf{u}| |\mathbf{v}|)$ denotes cosine similarity, B is the batch size, and τ is the temperature scaling parameter. This loss encourages alignment between corresponding sagittal and axial features.

- (2) **Consistency Loss**:

$$\mathcal{L}_{\text{consistency}} = \|\mathbf{f}_{\text{sag}} - \mathbf{f}_{\text{ax}}\|_2^2 \quad (4.12)$$

ensuring L2-norm feature consistency between views.

- (3) **Teacher Distillation Loss** (temperature $T = 3.0$):

$$\mathcal{L}_{\text{distill}} = \text{KL}(\sigma(\mathbf{z}_{\text{student}}/T) \parallel \sigma(\mathbf{z}_{\text{teacher}}/T)) \quad (4.13)$$

where $\mathbf{z}_{\text{student}}, \mathbf{z}_{\text{teacher}} \in \mathbb{R}^C$ are logits from student (axial) and teacher (sagittal) models, $\sigma(\cdot)$ denotes softmax activation, $\text{KL}(\cdot \parallel \cdot)$ is Kullback-Leibler divergence,

and T is the distillation temperature. This prevents performance collapse during Phase 2 by distilling knowledge from the pre-trained sagittal-only teacher model.

4.3.3.3 Curriculum and Task-weighted Loss Scheduling

The curriculum learning approach employs phase-specific loss combinations:

Phase 1 (Sagittal backbone frozen):

$$\mathcal{L}_{\text{phase1}} = \sum_{t \in \mathcal{T}_{\text{sag}}} \mathcal{L}_{\text{CE}}^{(t)} + 1.2 \sum_{t \in \mathcal{T}_{\text{multi}}} \mathcal{L}_{\text{CE}}^{(t)} \quad (4.14)$$

Phase 2 (Distillation):

$$\mathcal{L}_{\text{phase2}} = \mathcal{L}_{\text{phase1}} + 0.5\mathcal{L}_{\text{distill}} + 0.05(\mathcal{L}_{\text{contrastive}} + \mathcal{L}_{\text{consistency}}) \quad (4.15)$$

Phase 3 (Axial regularization):

$$\mathcal{L}_{\text{phase3}} = \mathcal{L}_{\text{sag_only}} + 1.2\mathcal{L}_{\text{multiview}} + 0.05(\mathcal{L}_{\text{contrastive}} + \mathcal{L}_{\text{consistency}}) \quad (4.16)$$

where

$$\mathcal{L}_{\text{sag_only}} = \sum_{t \in \mathcal{T}_{\text{sag}}} \mathcal{L}_{\text{CE}}^{(t)} + 0.01 \sum_{t \in \mathcal{T}_{\text{sag}}} |\mathbf{f}_{\text{ax}}|_2 \quad (4.17)$$

and the axial regularization term penalizes axial feature usage for sagittal-only tasks.

Phase 4 (Task-weighted fine-tuning):

$$\mathcal{L}_{\text{phase4}} = \sum_{t \in \mathcal{T}} w_t \mathcal{L}_{\text{CE}}^{(t)} \quad (4.18)$$

with task-specific weights: $w_{\text{hrn}} = 1.4$, $w_{\text{ccs}} = 1.3$, $w_{\text{pf}} = w_{\text{fsl}} = w_{\text{fsr}} = 1.2$, $w_{\text{nar}} = w_{\text{umc}} = w_{\text{lmc}} = 1.1$, $w_{\text{spn}} = w_{\text{ued}} = w_{\text{led}} = 1.0$. During this phase, all backbone and fusion parameters are frozen.

Here, $\mathcal{T}_{\text{sag}}$ denotes sagittal-only tasks, $\mathcal{T}_{\text{multi}}$ denotes multiview tasks, and $\mathcal{T} = \mathcal{T}_{\text{sag}} \cup \mathcal{T}_{\text{multi}}$ represents all tasks.

4.3.4 Methods

4.3.4.1 Datasets and Data Curation

To ensure robustness and generalizability, LumbarAI was trained and evaluated on a multi-source aggregated dataset. A critical aspect of curation was managing annotation quality and origin, distinguishing between *gold-standard* (expert-annotated) and *silver-standard* (SpineNetV2-predicted) labels. The composition is detailed in Table 4.5.

Table 4.5: Summary of dataset sources and annotation strategies

Dataset (Country)	#Scans (Patients) (Tr/Val/test)	Views		Pathologies with Gold Standard GT	Pathologies with Silver Standard GT
		Sag.	Axial		
XrayService-AI-Assisted Dataset (XSA) (Italy)	350 (350) (245/52/53)	✓	✓	All ¹	—
RSNA2024 (USA, Canada, Brazil, Turkey, Australia, Bosnia, Ireland)	8,593 (1,972) (6015/1289/1289)	✓	✓	ccs, fsl, fsr	pf, nar, spn, ued, led, umc, lmc, hrn
SPIDER (The Netherlands)	206 (206) (144/31/31)	✓	✗	pf, spn, ued, led, hrn	nar, cc, umc, lmc, fsl, fsr
Mendeley (UK, Indonesia)	512 (512) (512/0/0)	✓	✗	—	All
XrayService-Blinded Consensus Dataset (XSBC) (Italy)	19 (19) (0/0/19)	✓	✓	All ²	—

¹Expert-verified annotations obtained via correction of SpineNetV2-generated silver-standard predictions across *all* targeted pathologies. As these annotations were produced using a model-in-the-loop workflow, they may introduce circularity bias. This dataset (XSA) was used for train, validation, and internal test only, including ablation studies and model selection [47].

²Blinded expert consensus annotations independently derived by three board-certified radiologists across *all* targeted pathologies. Two readers annotated each case independently, with disagreements adjudicated by a third reader to form a final consensus. This dataset (XSBC) was used exclusively for unbiased comparison against state-of-the-art method.

4.3.4.2 Multi-View Preprocessing Pipeline

Sagittal Stack Extraction: We employed SpineNetV2 [48] to automatically detect vertebral bodies and localize intervertebral discs (IVDs) in T2-weighted sagittal scans. Using the predicted bounding boxes, we extracted a stack of patches for each disc of interest (L1–L2 to L5–S1), resulting in per-disc stacks of dimensions $9 \times 112 \times 224$ (slices $\times$ height $\times$ width). This process, illustrated in Figure 4.4(a), enabled precise disc-level analysis by isolating relevant anatomy while preserving essential contextual information from adjacent slices.

Axial–Sagittal Geometric Alignment: To enable consistent multi-view fusion,

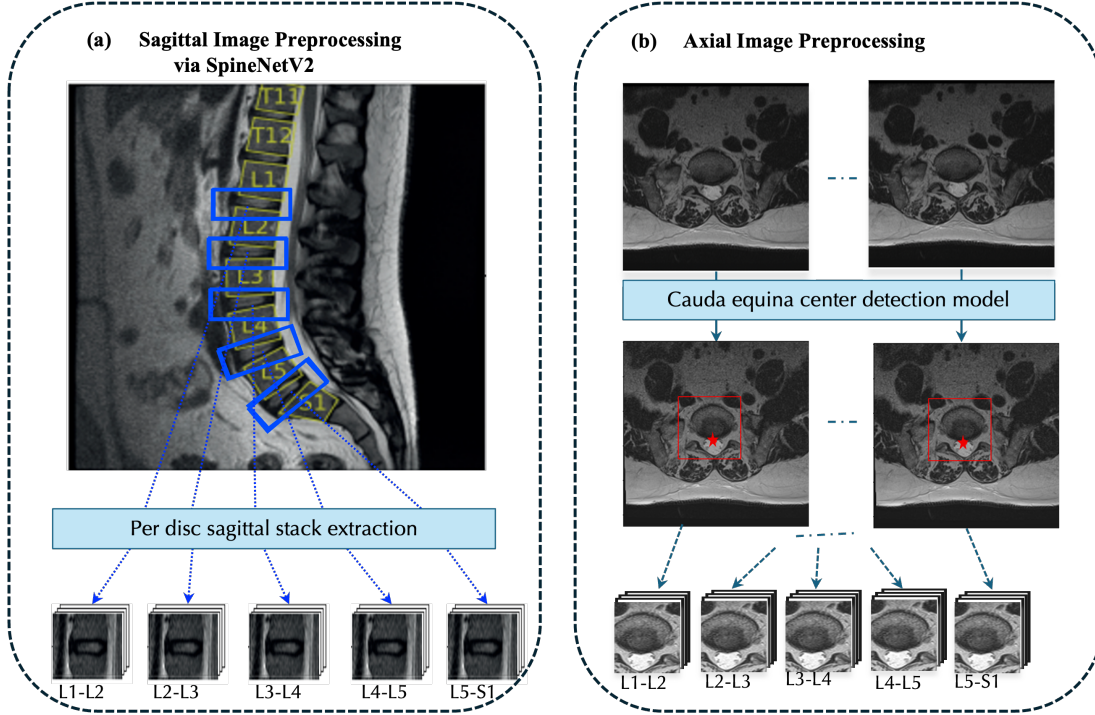


Figure 4.4: Multiview Preprocessing Pipeline for Level-Specific ROI Extraction. This figure illustrates the two-stage preprocessing pipeline for generating level-specific volumes from sagittal and axial T2-weighted MRI series. **(a) Sagittal Stack Extraction:** A mid-sagittal slice with SpineNetV2 detections (vertebral bodies in yellow, IVDs bounding boxes in blue) is used to extract five per-disc sagittal stacks (L1-L2 to L5-S1). Each stack is a 3D volume of contiguous slices cropped around its disc. **(b) Axial ROI Extraction:** For a given disc level, the axial sequence is processed by a Cauda Equina Center Localization model (ResNeXt-based) to predict the disc center ($\star$). A fixed-size (100×100) Region of Interest (ROI), adjusted by empirical offsets ($\Delta x = -20, \Delta y = -5$), is cropped around this center. The final output is an axial volume stack for each level, formed by aggregating these ROIs across slices.

axial slices were geometrically aligned to their corresponding sagittal disc levels using DICOM metadata. Both sagittal and axial series were transformed into a shared 3D patient coordinate system via the standard DICOM affine transform, which maps each pixel (i, j) to world coordinates $\mathbf{P}(x, y, z)$:

$$\mathbf{P}(x, y, z) = \mathbf{S} + i \cdot \Delta_i \mathbf{X} + j \cdot \Delta_j \mathbf{Y},$$

where $\mathbf{S} = (S_x, S_y, S_z)$ is the slice origin (ImagePositionPatient), $\mathbf{X}$ and $\mathbf{Y}$ are row and column direction cosines (ImageOrientationPatient), and Δ_i, Δ_j are pixel spacings acquired from DICOM metadata (Figure 4.5). This affine formulation enables the alignment of axial and sagittal MRI volumes into a common world coordinate system for precise multiview fusion.

Once both modalities were in the common coordinate frame, disc centroids from sagittal IVD annotations served as anatomical anchors. Each axial slice was

assigned to the nearest disc level by computing Euclidean distance between the axial slice center and each sagittal disc centroid. Axial slices physically intersecting the posterior boundary of a sagittal IVD region were grouped into that disc-level volume, ensuring anatomically consistent cross-plane pairing.

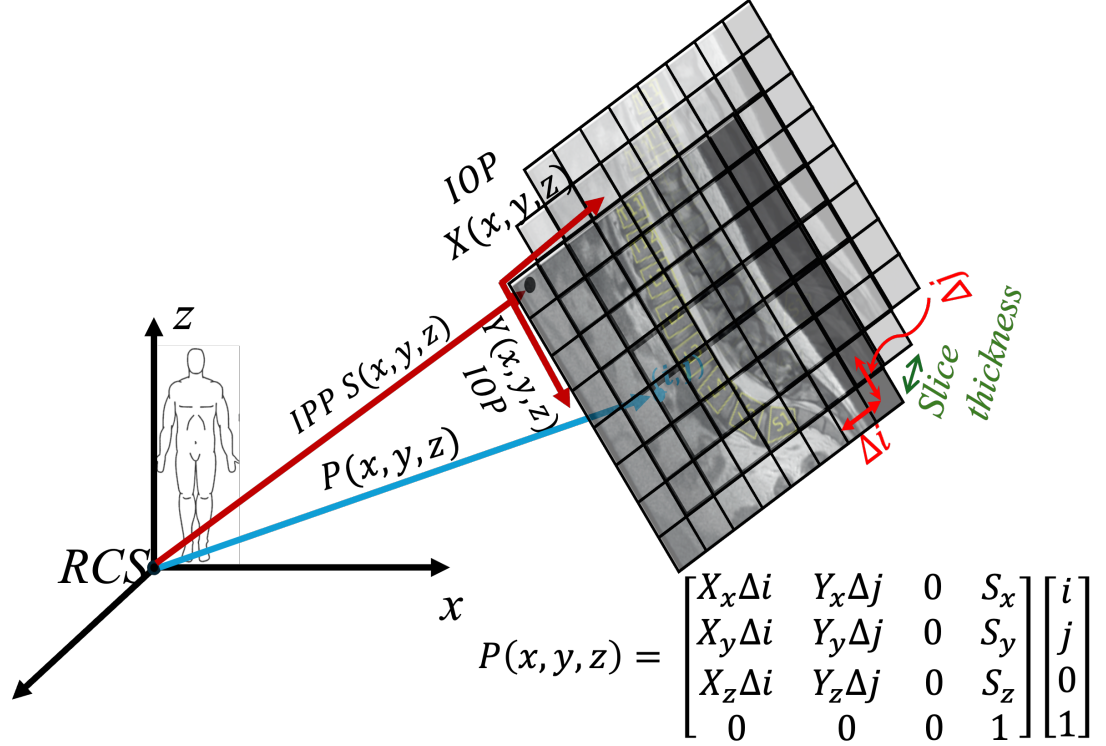


Figure 4.5: Illustration of the DICOM world coordinate system. Each pixel (i, j) is mapped to 3D patient space using the ImagePositionPatient (IPP) and ImageOrientationPatient (IOP) fields, combined with pixel spacing and slice thickness. This affine model enables consistent geometric alignment of sagittal and axial MRI sequences.

Axial ROI Extraction: Following axial–sagittal disc-level assignment, we extracted anatomically focused regions from each axial slice to construct level-specific stacks. A ResNeXt-based regression model [151] was trained to localize the cauda equina center in the axial plane, predicting coordinates (x, y) for each slice. Ground-truth centers were defined as the midpoint between left and right subarticular landmarks from the RSNA2024 LumbarDISC dataset [152]. The model was trained on 3,494 training, 873 validation, and 900 test samples, using only slices with complete bilateral subarticular annotations for label consistency. All images were normalized to $[0, 255]$ and resized to 256×256 pixels. Optimization used L1 loss over 50 epochs:

$$\mathcal{L}_1(\mathbf{p}, \mathbf{t}) = \frac{1}{n} \sum_{i=1}^n |p_i - t_i|.$$

The localization model demonstrated high precision on the test set, achieving a Mean Absolute Error (MAE) of 0.52 pixels and a Root Mean Square Error

(RMSE) of 0.73 pixels. Around the predicted center, a fixed-size (100×100) pixel ROI was cropped, with empirical offsets ($\Delta x = -20, \Delta y = -5$) to better align with the posterior disc margin. This process is illustrated in Figure 4.4(b). For each disc level, corresponding ROIs were stacked to form 3D axial volumes of dimensions $4 \times 100 \times 100$, that focus exclusively on the intervertebral space, thereby reducing computational noise from peripheral structures.

4.3.4.3 Experimental Setup and Evaluation Strategy

Dataset Splits: Patient-level 70:15:15 splits were applied to the XSA, RSNA, and SPIDER datasets, ensuring all slices from a given patient remained within the same split to prevent data leakage. This created distinct training, validation, and test sets. All Mendeley data were added exclusively to the training set to increase anatomical variability.

Preprocessing & Augmentation: All data underwent a uniform preprocessing pipeline including disc localization, multi-view geometric alignment, and ROI extraction as described in Section 4.3.4.2. During training, on-the-fly data augmentation (random flips, affine transformations, intensity perturbations) was applied using TorchIO [153]. **Evaluation Metrics:** We report Accuracy, Balanced Accuracy, F1-score (which jointly captures precision and recall), and AUC-ROC (where applicable). For multi-class tasks, metrics are macro-averaged across classes. **Threshold Selection:** For all classification tasks, decision thresholds were optimized individually for each model and each pathology class using the validation set. For binary tasks (spondylolisthesis, endplate defects, marrow changes, foraminal stenosis, herniation), thresholds were selected by maximizing the F1 score on the validation set. For multi-class ordinal tasks (Pfarrmann grading, disc narrowing, central canal stenosis), we applied a one-vs-rest sigmoid approach with a default threshold of 0.5 for each class, as these tasks require mutually exclusive predictions. To ensure fair comparison across ablation variations, thresholds were re-optimized independently for each model variant using its respective validation set, rather than transferring thresholds from the full model.

Statistical Analysis: Statistical significance of performance differences was assessed using paired Wilcoxon signed-rank tests [154] (chosen for non-parametric analysis without assuming normally distributed differences). Given the exploratory nature of comparisons and our focus on effect sizes within predefined task groups (sagittal vs. multi-view), we report uncorrected p-values alongside 95% confidence intervals computed via bias-corrected and accelerated (BCa) bootstrap with 1000 iterations [155].

Blind External Validation: For unbiased benchmarking, a completely unseen dataset of 19 MRIs (95 disc-level samples) was independently acquired from X-Ray Service s.r.l [129]. Two radiologists, blinded to any model predictions, annotated this external dataset, which was subsequently used to compare LumbarAI against SpineNetV2.

4.3.4.4 Implementation Details

The LumbarAI framework was implemented in Python 3.8 using PyTorch 1.12.0 and the Transformers library. The Swin3D backbone was adapted from the VideoSwin implementation, with modifications for 3D medical imaging. All experiments were conducted on NVIDIA A100 SXM4 GPUs with 40 GB VRAM. Training time varied by phase: Phase 1 (Axial Branch Learning) required 37 minutes, Phase 2 (Gradual Fusion Unfreezing) required 2 hours 1 minute, Phase 3 (Joint Fine-tuning) required 59 minutes, and Phase 4 (Head Specialization) required 30 minutes, for a total curriculum learning training time of 4 hours 7 minutes. The initial sagittal-only model training required 3 hours 21 minutes. Inference time per disc-level sample was approximately 0.15 seconds on a single GPU. The model contains approximately 87 million parameters, with the Swin3D backbone accounting for 82 million parameters and the task-specific heads and fusion module accounting for the remaining 5 million parameters.

4.3.5 Results

4.3.5.1 Performance of the Sagittal-Only Baseline

We established a sagittal-only baseline using a Swin3D Transformer architecture, trained end-to-end on all 11 pathological tasks from sagittal T2-weighted MRI sequences. This baseline provided a reference point for assessing the added value of multi-view integration. Model performance was evaluated on three independent test sets: the XSA, the RSNA dataset, and the SPIDER dataset. For RSNA and SPIDER, we report only tasks with corresponding gold-standard ground truth: ccs, fsl/fsr for RSNA; and pf, spn, led/ued, and hrn for SPIDER. The complete results are summarized in Table 4.6.

Analysis of Sagittal-Visible Task Performance The baseline model demonstrated strong performance across most spinal pathology tasks. For multi-class grading tasks, disc narrowing (nar, 4 classes) achieved robust results (0.783 accuracy, 0.777 F1), while Pfirrmann grading (pf, 5 classes) showed moderate performance (0.724 accuracy, 0.729 F1), reflecting the fine-grained distinctions

Table 4.6: *Sagittal-Only Baseline Performance on the XSA, RSNA and SPIDER test sets.*

	Sagittal Only Tasks							Multiview Tasks			
	pf	nar	spn	ued	led	umc	lmc	ccs	fsl	fsr	hrn
XSA											
Accuracy	0.724	0.783	0.986	0.863	0.905	0.926	0.863	0.881	0.783	0.754	0.666
F1 Score	0.729	0.777	0.984	0.830	0.895	0.928	0.822	0.872	0.780	0.745	0.589
RSNA											
Accuracy	-	-	-	-	-	-	-	0.828	0.772	0.799	0.725
F1 Score	-	-	-	-	-	-	-	0.860	0.783	0.795	0.784
SPIDER											
Accuracy	0.510	-	0.975	0.740	0.735	-	-	-	-	-	0.725
F1 Score	0.469	-	0.976	0.741	0.729	-	-	-	-	-	0.784

and inherent subjectivity of this ordinal classification. Endplate defects (ued, led) and vertebral marrow changes (umc, lmc) showed excellent performance across all metrics (accuracy: 0.86–0.926, F1: 0.82–0.928), confirming that these degenerative findings are well captured on sagittal views. Spondylolisthesis (spn) yielded near-perfect detection (0.986 accuracy, 0.984 F1). The results are noteworthy given the significant class imbalance present in the multi-class grading tasks.

Limitations on Cross-Sectional Pathologies In contrast, significant performance gaps emerged for pathologies where radiological assessment typically relies on cross-sectional imaging. Foraminal stenosis detection was notably affected: left foraminal stenosis (fsl) achieved 0.783 accuracy (0.754 F1) on XSA dataset and 0.772 accuracy (0.783 F1) on RSNA; right foraminal stenosis (fsr) showed similar limitations with 0.754 accuracy (0.745 F1) on XSA and 0.799 accuracy (0.795 F1) on RSNA. Herniation (hrn) performance was inconsistent, with moderate results on XSA (0.666 accuracy, 0.589 F1) but limited generalizability to SPIDER (0.725 accuracy, 0.784 F1). Central canal stenosis (ccs) was a notable exception, performing strongly both on XSA (0.881 accuracy, 0.872 F1) and on RSNA (0.828 accuracy, 0.860 F1), likely because its morphological signature is often visible on sagittal images. In clinical consultation, radiologists confirmed that while these pathologies can be assessed from sagittal views, axial cross-sectional images provide crucial complementary information for easier and more practical diagnosis, directly motivating the need for multiview integration.

These empirical findings and clinical insights align with established radiological practice where foraminal stenosis and herniation evaluation typically requires axial or oblique imaging to accurately localize cross-sectional compromise and neural

element compression [93, 94]. The performance dichotomy between sagittal-visible and cross-sectional pathologies validates the clinical motivation for our hybrid architecture.

4.3.5.2 Hybrid Architecture Performance

The hybrid architecture demonstrated clear task-specific specialization, with the Sagittal-Only Expert maintaining excellence on longitudinal pathologies while the Multi-View Expert achieved substantial improvements on cross-sectional assessments.

Table 4.7: *Multiview model performance across all test sets*

Dataset	Metric	ccs	fsl	fsr	hrn
XSA	Accuracy	0.923	0.843	0.851	0.786
	F1 Score	0.895	0.816	0.830	0.781
RSNA	Accuracy	0.883	0.800	0.803	–
	F1 Score	0.877	0.788	0.792	–
SPIDER	Accuracy	–	–	–	0.910
	F1 Score	–	–	–	0.891

Multi-View Task Improvements

The Multi-View model demonstrated clinically significant gains for cross-sectional pathologies across evaluation datasets (Table 4.7). Table 4.8 presents a detailed comparison between sagittal-only and multi-view performance, averaging results across available datasets for each task. Unless otherwise specified, all percentage improvements reported below represent relative percentage change relative to the sagittal-only baseline.

Foraminal Stenosis Performance Gains For left foraminal stenosis (fsl), the Multi-View model achieved a **4.6%** relative improvement in F1 score (from 0.781 to 0.816) with a **7.7%** relative improvement in accuracy (0.783 to 0.843) on the XSA dataset. External validation on RSNA showed more modest gains: **0.6%** relative F1 improvement (0.783 to 0.788) and **3.6%** relative accuracy improvement (0.772 to 0.800).

For right foraminal stenosis (fsr), improvements were more substantial: **11.4%** relative F1 gain (0.745 to 0.830) and **12.9%** relative accuracy improvement (0.754 to 0.851) on XSA. On RSNA, gains were **-0.4%** relative F1 (0.795 to 0.792) and **0.5%** relative accuracy (0.799 to 0.803).

Table 4.8: Comparison of Multi-View and Sagittal-Only Models: Task-Averaged Performance Metrics from Tables 4.6 and 4.7.

Model	fsl		fsr		hrn		ccs	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Sagittal-Only	0.7775	0.781	0.776	0.770	0.695	0.686	0.8545	0.866
Multi-View	0.8415	0.802	0.827	0.811	0.848	0.836	0.903	0.866

Herniation Assessment Enhancement Herniation (hrn) detection showed the most dramatic improvement: **32.6%** F1 improvement (0.589 to 0.781, $p < 0.001$) with **18.0%** accuracy gain (0.666 to 0.786) on the XSA dataset. External validation on SPIDER showed more modest but still meaningful gains: **13.6%** F1 improvement (0.784 to 0.891) and **25.5%** accuracy improvement (0.725 to 0.910). This pattern suggests the multi-view approach particularly enhances difficult case detection, indicating better identification of true herniation cases that were missed by the sagittal-only model.

Central Canal Stenosis Analysis Central canal stenosis (ccs) also showed consistent but smaller improvements. On the XSA dataset, the Multi-View model achieved **2.6%** F1 improvement (0.872 to 0.895) and **4.8%** accuracy gain (0.881 to 0.923). External validation on RSNA showed **2.0%** F1 improvement (0.860 to 0.877) and **6.6%** accuracy gain (0.828 to 0.883). These modest improvements suggest that while ccs is reasonably assessable from sagittal views alone, the multi-view approach still provides complementary value, particularly for complex cases requiring detailed cross-sectional assessment of lateral recesses and cauda equina areas.

To further validate the discriminative performance of our hybrid architecture, we evaluated the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for both expert models on the combined test set (RSNA, SPIDER, and XSA dataset), as shown in Figure 4.6. The sagittal-only expert model demonstrated excellent discriminative capability for sagittal-visible pathologies, with AUC values ranging from 0.84 to 0.96 across all seven tasks. Performance was highest for upper and lower marrow changes (umc: 0.96, lmc: 0.95), followed by disc narrowing (nar: 0.94), spondylolisthesis (spn: 0.93), and Pfirrmann grading (pf: 0.89). Conversely, the multiview expert model showed strong discriminative performance for the cross-sectional pathologies, with AUC values between 0.80 and 0.88. Central canal stenosis (CCS) achieved the highest score (0.88), followed by herniation (hrn: 0.84), and foraminal stenosis tasks (fsr: 0.81, fsl: 0.80). These results confirm a

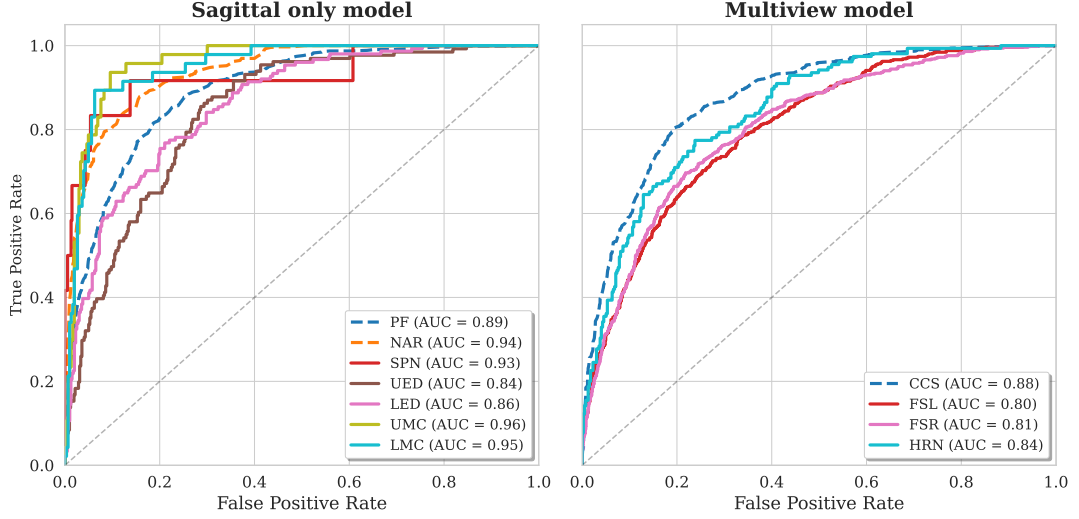


Figure 4.6: Receiver Operating Characteristic (ROC) Curves for the Hybrid Model Architecture. Dashed lines represent macro-averaged ROC curves for ordinal, multi-class tasks (*ccs*: 3-class, *nar*: 4-class, *pf*: 5-class).

clear performance dichotomy aligned with the models’ designed expertise. The sagittal-only model excels at longitudinal pathologies, while the multiview model provides robust, clinically meaningful discrimination for pathologies requiring axial-plane assessment. This distinct and complementary performance profile strongly validates our hybrid architectural design, where each expert model is optimized for its specific anatomical and diagnostic context.

4.3.6 Curriculum Learning Progression Analysis

To quantitatively assess the contribution of each stage within the proposed multiview curriculum learning strategy, the model was evaluated after each of the four sequential training phases using the combined test set (2,516 samples across all datasets). The performance metrics (Accuracy and F1 Score) for the four multiview tasks across all phases are presented in Table 4.9. The phase-to-phase progression, detailing the incremental changes from one phase to the next and highlighting the best-performing phase for each task, is shown in Table 4.10. The following analysis examines the results of each phase in sequence.

Phase 1: Axial Branch Learning. Phase 1 establishes the baseline multiview performance by introducing an axial-view processing branch while preserving the pre-trained sagittal representations. To prevent catastrophic forgetting, the sagittal backbone remains completely frozen; only the new axial encoder, the fusion module, and the multiview task heads are trainable. This conservative approach ensures robust sagittal performance is maintained while the model gradually learns

Table 4.9: *Curriculum Learning Phase Performance: Accuracy and F1 Score Across Phases 1-4 for Multiview Tasks (Combined Test Set, 2516 samples)*

Phase	ccs		fsl		fsr		hrn	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Phase 1	0.883	0.848	0.782	0.762	0.798	0.777	0.777	0.772
Phase 2	0.883	0.854	0.785	0.777	0.800	0.789	0.787	0.791
Phase 3	0.888	0.857	0.782	0.777	0.797	0.786	0.781	0.786
Phase 4	0.903	0.866	0.841	0.802	0.827	0.811	0.848	0.836
Average	0.889	0.856	0.798	0.780	0.806	0.791	0.798	0.796

Abbreviations: Acc, Accuracy; F1, F1-Score.

to incorporate axial information. As shown in Table 4.9, Phase 1 achieved an average accuracy of 0.810 and an F1 score of 0.790 across all tasks. Central canal stenosis (ccs) demonstrated the strongest initial performance (accuracy: 0.883, F1: 0.848), indicating the axial branch effectively captures relevant cross-sectional anatomy. Herniation (hrn) and foraminal stenosis tasks (fsl, fsr) showed moderate but promising performance, confirming the frozen sagittal backbone provides a stable foundation for subsequent learning phases.

Phase 2: Gradual Fusion Unfreezing with Contrastive Consistency.

Phase 2 introduces teacher-student distillation and contrastive learning between sagittal and axial feature representations. This mechanism encourages the learning of view-invariant features while preserving view-specific discriminative information. This phase yielded the largest F1 score improvement in the curriculum, with the average F1 increasing from 0.790 to 0.803 (+1.28% relative). As detailed in Table 4.10, herniation (hrn) benefited the most (F1: +1.92%), followed by left foraminal stenosis (fsl: +1.50%) and right foraminal stenosis (fsr: +1.18%). The contrastive consistency loss appears particularly effective for tasks like herniation that require precise cross-view alignment for accurate assessment. The average accuracy also improved from 0.810 to 0.814 (+0.39%), confirming the fusion mechanism successfully integrates complementary information from both views.

Phase 3: Joint Fine-tuning. Phase 3 enables joint, end-to-end optimization of both view encoders, employing carefully balanced learning rates (sagittal: 1×10^{-6} , axial: 5×10^{-5} , fusion: 1×10^{-4}) determined through hyperparameter tuning. While this phase allowed for the refinement of cross-view feature interactions, it produced mixed results: average accuracy and F1 score showed slight decreases of 0.22% and 0.13%, respectively (Table 4.10). Notably, central canal stenosis (ccs) continued to improve, with accuracy reaching 0.888 and F1 reaching 0.857. The minor fluctuations observed in other tasks, such as foraminal stenosis and

herniation, may reflect competing gradient dynamics and varying sensitivity of each pathology to fine-grained parameter updates. Tasks that rely on high-precision spatial alignment of lateral anatomical cues (e.g., foraminal stenosis) may be more susceptible to gradient interference during joint optimization. Conversely, more structurally defined tasks like ccs, which benefit from the consolidated axial-sagittal context, show continued improvement.

Phase 4: Head Specialization. The final phase focuses on expert-level optimization of the task-specific prediction heads while freezing all backbone and fusion parameters. A weighted loss function emphasizes clinically critical pathologies (herniation weight: 1.4, central canal stenosis weight: 1.3). This phase achieved the highest overall performance, with an average accuracy of 0.854 and an F1 score of 0.828, representing substantial improvements over previous phases. According to Table 4.9, Phase 4 delivered the best performance for all four tasks. Herniation (hrn) showed dramatic gains, with accuracy reaching 0.848 and F1 reaching 0.836. Similarly, left foraminal stenosis (fsl) achieved an accuracy of 0.841, and central canal stenosis (ccs) reached 0.903 accuracy. This phase demonstrates that targeted, task-specific refinement after robust feature learning enables significant performance gains, validating the head specialization strategy.

Cumulative Improvement (Phase 1 to Phase 4). The complete curriculum learning strategy achieved substantial and reasonable improvements from the initial to the final phase across all tasks. Herniation (hrn) showed the largest total gains (+7.1% accuracy, +6.4% F1), followed by left foraminal stenosis (fsl: +5.9% accuracy, +4.0% F1) and right foraminal stenosis (fsr: +2.9% accuracy, +3.4% F1). The average improvement across all tasks was +4.4% in accuracy and +3.8% in F1 score. This progressive improvement pattern, with Phase 2 providing the largest F1 boost through contrastive learning and Phase 4 delivering final, significant refinement via head specialization, validates the effectiveness of the phased curriculum. Each stage contributes meaningfully to building a robust multiview model that preserves foundational sagittal knowledge while learning to leverage complementary axial information.

4.3.7 Ablation Studies: Component Contribution Analysis

To validate that each component of the proposed curriculum learning framework contributes meaningfully to the final model performance, a series of comprehensive ablation studies was conducted. These experiments systematically remove or modify individual architectural and training components, quantifying their impact on the four core multiview tasks. The results provide empirical justification for

Table 4.10: *Curriculum Learning Phase-to-Phase Progression: Performance Changes Across Phases 1-4 for Multiview Tasks (Combined Test Set, 2516 samples)*

Phase / Change	ccs		fsl		fsr		hrn	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Phase 1	0.883	0.848	0.782	0.762	0.798	0.777	0.777	0.772
P1 → P2	+0.000	+0.006	+0.003	+0.015	+0.002	+0.012	+0.010	+0.019
Phase 2	0.883	0.854	0.785	0.777	0.800	0.789	0.787	0.791
P2 → P3	+0.005	+0.003	-0.003	+0.000	-0.003	-0.003	-0.006	-0.005
Phase 3	0.888	0.857	0.782	0.777	0.797	0.786	0.781	0.786
P3 → P4	+0.015	+0.009	+0.059	+0.025	+0.030	+0.025	+0.067	+0.050
Phase 4	0.903	0.866	0.841	0.802	0.827	0.811	0.848	0.836
Total Δ (P1 → P4)	+0.020	+0.018	+0.059	+0.040	+0.029	+0.034	+0.071	+0.064

Bold values indicate the best performing phase for each task and metric. P1→P2, P2→P3, and P3→P4 indicate phase-to-phase changes (improvements shown as positive values, degradations as negative values). Total Δ represents cumulative improvement from Phase 1 to Phase 4.

the key design choices and demonstrate that the superior performance achieved in Phase 4 is attributable to the integrated framework, rather than to any single component.

4.3.7.1 The Necessity of Curriculum Learning

The most fundamental ablation removes the entire curriculum learning strategy, training the multiview model directly from the sagittal-only foundation in a conventional end-to-end manner. This baseline evaluates the necessity of the phased, progressive learning approach. Table 4.11 compares this ablated model against the complete curriculum learning model (Phase 4).

Table 4.11: *Ablation Study: Impact of Curriculum Learning on Model Performance (Multiview Tasks)*

Model	ccs		fsl		fsr		hrn	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Without Curriculum	0.812	0.842	0.737	0.758	0.740	0.759	0.794	0.798
Full Curriculum (Phase 4)	0.903	0.866	0.841	0.802	0.827	0.811	0.848	0.836
Improvement (%)	+9.1%	+2.4%	+10.4%	+4.4%	+8.7%	+5.2%	+5.4%	+3.8%

Abbreviation: Acc = Accuracy, F1 = F1-score.

The results demonstrate that curriculum learning provides substantial and consistent improvements across all tasks, with an average gain of +8.3% in accuracy and +3.9% in F1 score. The largest accuracy improvements are observed for the foraminal stenosis tasks (fsl: +10.4%, fsr: +8.7%), which are highly dependent on the precise integration of sagittal and axial anatomical contexts, a capability progressively built through the curriculum. Notably, herniation (hrn) also shows a

significant +5.4% accuracy improvement, highlighting the benefit of the final head specialization phase. This ablation validates a progressive, phased curriculum is essential for the model to effectively learn cross-view representations without catastrophic forgetting, leading to the superior performance documented in the final phase.

4.3.7.2 Impact of Individual Components

To further isolate the contribution of specific mechanisms within the curriculum, a comprehensive ablation study was performed, as detailed in Table 4.12. The table includes the ablated “Without Curriculum” model for direct comparison with other component removals.

Table 4.12: *Comprehensive Ablation Study: Impact of Individual Components on Multi-view Task Performance*

Component	ccs		fsl		fsr		hrn	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Axial Only	0.860	0.484	0.496	0.454	0.606	0.567	0.386	0.368
Fusion: Sum	0.881	0.433	0.795	0.650	0.800	0.652	0.785	0.710
Fusion: Concatenation	0.472	0.258	0.493	0.481	0.614	0.512	0.314	0.304
Fusion: Attention	0.880	0.312	0.775	0.437	0.780	0.438	0.687	0.407
Without Curriculum (end-to-end)	0.812	0.842	0.737	0.758	0.740	0.759	0.794	0.798
Without Contrastive Loss	0.884	0.428	0.810	0.664	0.812	0.665	0.801	0.763
Full Curriculum (Phase 4)	0.903	0.866	0.841	0.802	0.827	0.811	0.848	0.836

Abbreviation: Acc = Accuracy, F1 = F1-score.

Contrastive Consistency Loss. Removing the contrastive learning mechanism (introduced in Phase 2) results in a dramatic collapse of F1 scores across all tasks, most notably for ccs (F1: 0.866 $\rightarrow$ 0.428) and the foraminal stenosis tasks, despite relatively preserved accuracy. This dichotomy confirms the critical role of this component in learning *view-invariant representations* that are robust to class imbalance. The contrastive loss ensures that semantically similar features from different views are aligned in the embedding space, a process evidently crucial for generating well-calibrated probabilistic outputs (reflected in F1) even when raw accuracy remains high. This ablation directly justifies its inclusion in Phase 2, which previously showed the curriculum’s largest F1 gain.

Multiview Integration (Axial Only). Training with only axial views, removing the foundational sagittal backbone, leads to catastrophic performance degradation, particularly for herniation (hrn accuracy: 0.848 $\rightarrow$ 0.386) and left foraminal stenosis (fsl accuracy: 0.841 $\rightarrow$ 0.496). This result underscores the non-negotiable importance of the proposed *multiview* approach and validates the initial design choice of Phase 1, which conservatively introduced the axial branch while keeping

the sagittal encoder frozen. The sagittal view provides essential longitudinal anatomical context that the axial view alone cannot supply.

Fusion Strategy. Alternative feature fusion strategies were evaluated against the baseline weighted average fusion. While *Sum* fusion maintains reasonable accuracy, it suffers significant F1 degradation (e.g., ccs F1: 0.866 $\rightarrow$ 0.433), indicating poor handling of class distributions. *Concatenation* and *Attention*-based fusion fail catastrophically, with accuracy often falling near random-chance levels. This validates the selection of the simple, learnable weighted average fusion used throughout the curriculum, which provides the stability and task-specific adaptability necessary for effective cross-view information integration across all phases.

4.3.7.3 Synthesis of Ablation Findings

The ablation studies collectively demonstrate a clear hierarchy of component importance, providing strong empirical validation for the integrated design of the proposed curriculum learning framework. It is important to note that while the hyperparameters for the full curriculum phases (1-4) were rigorously optimized, the ablation experiments were conducted with a consistent, controlled methodology to ensure fair comparison: each component was removed or replaced individually while keeping all other training parameters (learning rates, batch size, optimizer) identical to those used in the final Phase 4 model. For the “without curriculum” ablation, the total training budget (number of epochs) was set to the sum of epochs across all four curriculum phases, ensuring equivalent computational resources. This approach isolates the effect of curriculum structure rather than tuning differences.

The findings reveal that:

Multiview Integration is fundamental; performance collapses catastrophically when using only axial views, confirming that neither view alone is sufficient for accurate diagnosis of complex spinal pathologies.

Curriculum Learning is essential; end-to-end training with equivalent epochs fails to achieve the same performance, demonstrating that the progressive, phased approach enables more stable and effective learning of cross-view representations.

Contrastive Consistency Loss is critical for handling class imbalance; its removal causes a dramatic collapse in F1 scores despite preserved accuracy, highlighting its role in learning robust, view-invariant feature representations.

Weighted Average Fusion provides an optimal balance; simpler (sum) or more complex (attention, concatenation) fusion strategies consistently underperform, validating this design choice for stable feature integration.

These results robustly justify the integrated design of the proposed framework. The substantial performance advantage of the full model (Phase 4) is not attributable to any single component but emerges from their synergistic integration within the curriculum learning paradigm. The ablation methodology, maintaining consistent training budgets and hyperparameters, ensures that the observed differences genuinely reflect the value of each architectural and training component rather than optimization artifacts.

4.3.8 Independent Validation with Blind Annotations: LumbarAI vs. SpineNet

To conclusively address methodological concerns regarding potential annotation bias, an independent validation was conducted using the XSBC (XrayService-Blinded Consensus Dataset) described in Table 4.5. This evaluation utilized a dataset of 19 previously unseen MRI scans (comprising 95 disc-level samples across 11 diagnostic tasks), with consensus ground truth established by three expert radiologists following a blinded adjudication protocol. This framework eliminates the systematic circularity bias present in the XSA dataset.

Table 4.13: *Performance Comparison on Blind-Annotated External Validation Dataset (XSB Consensus Dataset)*

Metric/Model	Sagittal Only Tasks							Multiview Tasks			
	pf	nar	spn	ued	led	umc	lmc	ccs	fsl	fsr	hrn
Accuracy											
LumbarAI Acc	0.421 (0.916 [†])	0.505 (0.905 [†])	0.978	0.863	0.912	0.920	0.863	0.947	0.810	0.789	0.663
SpineNet Acc	0.442 (0.958 [†])	0.568 (0.937 [†])	0.947	0.905	0.915	0.863	0.842	0.863	0.800	0.778	0.684
Inter-Observer Acc	0.484 (0.916 [†])	0.589 (0.937 [†])	0.937	0.884	0.863	0.937	0.926	0.916	0.821	0.831	0.779
AUC											
LumbarAI AUC	0.771	0.720	0.990	0.799	0.790	0.922	0.890	0.903	0.859	0.763	0.787
SpineNet AUC	0.816	0.805	0.989	0.891	0.948	0.885	0.911	0.407	0.775	0.761	0.783

[†] Values in parentheses denote ± 1 grade accuracy for multiclass ordinal tasks (pf: Pfirrmann grading, nar: disc height narrowing). This metric is clinically more meaningful given the inherent subjectivity in these grading schemes.

Color coding for Inter-Observer Agreement vs LumbarAI Accuracy: **Green** = inter-observer agreement > LumbarAI accuracy; **Orange** = inter-observer agreement $\approx$ LumbarAI accuracy; **Red** = LumbarAI accuracy > inter-rater agreement.

Note: **Bold** values indicate best performance per task within each metric section (Accuracy or AUC) for the two comparable models (LumbarAI and SpineNet).

The comprehensive performance comparison between the proposed LumbarAI model and the SpineNetV2 baseline on this blind-annotated dataset is presented in Table 4.13. Both accuracy and area under the ROC curve (AUC) are reported

to provide complementary perspectives: accuracy reflects performance at the optimal decision threshold, while AUC captures the model’s ranking ability across all thresholds. A detailed analysis yields the following key findings:

Comparative Model Performance (Accuracy): Based on strict accuracy, the LumbarAI model demonstrated superior performance on **6 out of 11 tasks** (ccs, spn, lmc, umc, fsl, fsr), while SpineNetV2 outperformed on **4 out of 11 tasks** (pf, nar, ued, hrn), with comparable performance observed on the **led** task (0.912 vs. 0.915). This result provides strong empirical evidence that the proposed multiview curriculum learning approach in LumbarAI generalizes more effectively to unseen, independently annotated data, particularly for tasks requiring spatial integration.

Comparative Model Performance (AUC): The AUC comparison reveals a more nuanced competitive landscape. LumbarAI achieved superior AUC on **6 out of 11 tasks** (ccs: 0.903 vs. 0.407; fsl: 0.859 vs. 0.775; fsr: 0.763 vs. 0.761; hrn: 0.787 vs. 0.783; spn: 0.990 vs. 0.989; umc: 0.922 vs. 0.885), while SpineNetV2 led on **5 out of 11 tasks** (pf: 0.816 vs. 0.771; nar: 0.805 vs. 0.720; led: 0.948 vs. 0.790; ued: 0.891 vs. 0.799; lmc: 0.911 vs. 0.890). Notably, despite SpineNet’s higher accuracy on **led** and **ued**, LumbarAI’s competitive AUC on these tasks (0.790 and 0.799, respectively) suggests its probability estimates remain reasonably well-calibrated. The most striking discrepancy occurs for **ccs**, where LumbarAI’s AUC (0.903) far exceeds SpineNetV2’s (0.407), indicating that SpineNetV2’s probability rankings for this critical diagnosis are near-random despite its moderate accuracy (0.863) on this specific test set.

Task-Specific Analysis: **LumbarAI** showed pronounced advantages in diagnosing critical structural spinal pathologies. It achieved superior accuracy and AUC in detecting *Central Canal Stenosis* (ccs) (acc: 0.947 vs. 0.863; AUC: 0.903 vs. 0.407), *Spondylolisthesis* (spn) (acc: 0.978 vs. 0.947; AUC: 0.990 vs. 0.989), and bilateral *Foraminal Stenosis* (fsl, fsr). This indicates that the curriculum learning framework is particularly effective for tasks requiring the integration of complex anatomical relationships and spatial contexts to identify neural compression.

SpineNetV2 maintained a consistent advantage in *Pfirrmann Grading* (pf) and *disc height narrowing* (nar) based on both strict accuracy (pf: 0.442 vs. 0.421; nar: 0.568 vs. 0.505) and AUC (pf: 0.816 vs. 0.771; nar: 0.805 vs. 0.720). However,

the clinically relevant ± 1 grade accuracy was nearly equivalent between models (LumbarAI: 0.916/0.905; SpineNet: 0.958/0.937; Inter-Observer: 0.916/0.937), confirming both models’ utility in capturing the ordinal grading continuum. SpineNetV2 also demonstrated superior performance on *Upper Endplate Defects (ued)* (acc: 0.905 vs. 0.863; AUC: 0.891 vs. 0.799) and *Herniation (hrn)* (acc: 0.684 vs. 0.663; AUC: 0.783 vs. 0.747), suggesting heightened sensitivity to specific disc morphology features.

Benchmarking Against Human Agreement: A critical benchmark for clinical AI is performance relative to inter-observer variability among expert radiologists. The consensus ground truth established by the three-radiologist panel (two independent annotators plus senior adjudication) provides a robust reference standard. The analysis reveals a spectrum of model capability:

- **Exceeding Expert Consensus:** LumbarAI’s accuracy for *ccs* (0.947 vs. human 0.916), *spn* (0.978 vs. human 0.937), and *led* (0.912 vs. human 0.863) exceeded the pairwise inter-observer agreement, demonstrating superhuman reliability for these specific diagnoses.
- **Near Expert Consensus:** For the ordinal grading tasks (*pf*, *nar*), both models’ ± 1 grade accuracies closely approached the expert agreement level. Additionally, LumbarAI’s accuracy for *lmc* (0.863 vs. human 0.926), *umc* (0.926 vs. human 0.937), *fsl* (0.810 vs. human 0.821), and *fsr* (0.789 vs. human 0.831) approached consensus, with human agreement providing a clear but narrow performance ceiling.
- **Below Expert Consensus:** For *ued* and *hrn*, both models performed below the inter-observer agreement (*ued*: LumbarAI 0.863, SpineNet 0.905, human 0.884; *hrn*: LumbarAI 0.663, SpineNet 0.684, human 0.779). This establishes a clear performance ceiling and delineates priority areas for future model refinement, particularly for herniation detection where the human-level agreement itself is modest (0.779).

Clinical Implications: The results of this independent, blind-annotated validation confirm the clinical robustness and generalizability of the proposed LumbarAI framework. Specifically, its demonstrated strength in identifying critical structural pathologies, central and foraminal stenosis, spondylolisthesis, and endplate defects, which have direct implications for surgical triage and treatment planning, substantiates its potential as an effective clinical decision-support tool. Crucially, the model’s performance exceeding or approaching inter-rater agreement levels for

several key tasks suggests that it may serve as a reliable first-level decision-support tool within the clinical workflow, assisting clinicians in routine assessment while remaining subject to expert oversight.

As envisioned in the proposed framework, this validated model constitutes the core of the Computer Vision Module (Module 1). Its primary role is not autonomous diagnosis, but to generate a preliminary draft report of findings, thereby reducing the initial reporting burden on radiologists. This design intentionally follows a human-in-the-loop principle, where the AI-generated draft is subsequently reviewed, corrected, and refined by the expert radiologist. This collaborative workflow mitigates the risk of over-reliance on the model while leveraging its efficiency to draft comprehensive assessments, including subtle findings that might otherwise be overlooked under time pressure.

4.3.9 Qualitative Interpretability Analysis of LumbarAI

To critically evaluate the internal reasoning of the trained expert model, we performed a multi-modal explainability analysis across 11 spinal pathology tasks using six complementary techniques. Our analysis is summarized in Figure 4.7, which presents a composite visualization for direct comparison: each row corresponds to an explainability method (a: original MRI, b: attention rollout, c: saliency, d: guided backpropagation, e: signed LIME, f: Grad-CAM), and each column represents one of the 11 spinal pathology tasks. This structured presentation allows us to assess what different methods reveal, or obscure, about the model’s anatomical focus and decision-making logic, particularly for fine-grained anatomical discrimination.

Attention rollout visualizations (Figure 4.7(b)) served a dual purpose in our interpretability analysis. Primarily, they provided a high-level validation that the model’s attention is correctly anchored to the intervertebral disc region across all tasks, confirming its focus on relevant anatomy rather than irrelevant background. However, this global coherence came at the cost of diagnostic spatial precision. The heatmaps consistently failed to localize critical anatomical distinctions; for tasks requiring discrimination, such as between upper vs. lower endplate defects or left vs. right foraminal stenosis, attention activations remained symmetrically centered and nearly identical. This indicates that while the transformer’s self-attention mechanism successfully aggregates contextual information for scene understanding, it operates at a resolution too coarse to provide the fine-grained, task-specific localization needed for definitive clinical diagnoses in its current form.

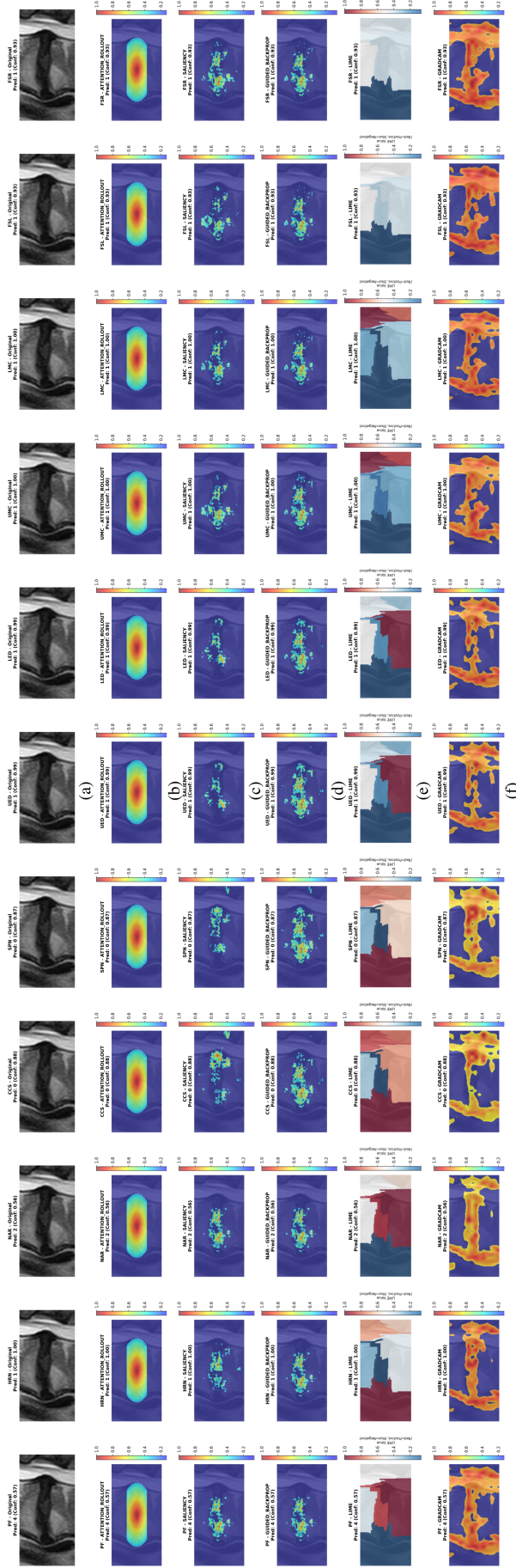


Figure 4.7: Comprehensive explainability analysis of the sagittal only model across all spinal pathology tasks. The composite visualization compares six explainability techniques applied to the model’s predictions for 11 diagnostic tasks. Each row corresponds to a specific method: (a) original mid-sagittal T2-weighted MRI slice for anatomical reference; (b) Attention Rollout heatmap; (c) Saliency Map; (d) Guided Backpropagation map; (e) signed LIME explanation (red: positive contribution, blue: negative contribution); (f) Grad-CAM activation. Each column represents one of the 11 spinal pathology tasks.

In contrast, gradient-based methods, saliency mapping (Figure 4.7(c)) and guided backpropagation (d), demonstrated high spatial fidelity aligned with radiological intuition. For Pfirrmann grading, activation concentrated on the nucleus pulposus; for herniation, it highlighted posterior disc extrusion; and for central canal stenosis, it focused on the ligamentum flavum. Crucially, these methods successfully lateralized activation for foraminal stenosis, a capability absent in attention rollout. This strong agreement with known pathological landmarks validates that the model’s convolutional backbone learns to focus on clinically relevant regions, building essential trust in its core feature extraction.

The signed LIME explanations (Figure 4.7(e)) provided unique causal insight by showing both positive (red) and negative (blue) feature contributions. For herniation, it highlighted red posterior regions (extruded material) alongside blue anterior regions (preserved disc), revealing a nuanced logic. For foraminal stenosis, it explicitly showed asymmetric red/blue patterns that lateralized to the affected side. This method excels at illustrating how the model uses the absence of a finding in one region as evidence, offering a deeper, region-based understanding of its decision-making logic.

Finally, Grad-CAM visualizations (Figure 4.7 (f)) produced structured, anatomically coherent heatmaps that often outlined bone or joint spaces. It successfully highlighted task-relevant regions, such as the disc space for Pfirrmann grading. However, for tasks requiring fine discrimination (e.g., ued vs. led), its heatmaps were less differentiated than those from gradient-based methods, suggesting it may not fully leverage the task-specific components of the architecture for the highest level of spatial distinction.

Extending this analysis, we also examined Grad-CAM activations from the multi-view Swin3D model (see Figure 4.8) to assess how integrating axial information changes the model’s focus. The multiview model revealed complementary spatial insights not apparent in sagittal-only views. In sagittal views, activation patterns highlighted longitudinal features like disc morphology and posterior extrusions. In the corresponding axial views, activation concentrated on cross-sectional anatomy, such as the central disc integrity versus peripheral herniation. This view complementarity demonstrates that the multiview architecture successfully integrates different anatomical perspectives, mirroring the radiologist’s multi-planar assessment protocol and providing more comprehensive spatial reasoning.

Our multi-method analysis reveals a clear interpretability hierarchy. For precise localization and lateralization, gradient-based methods are most reliable. For

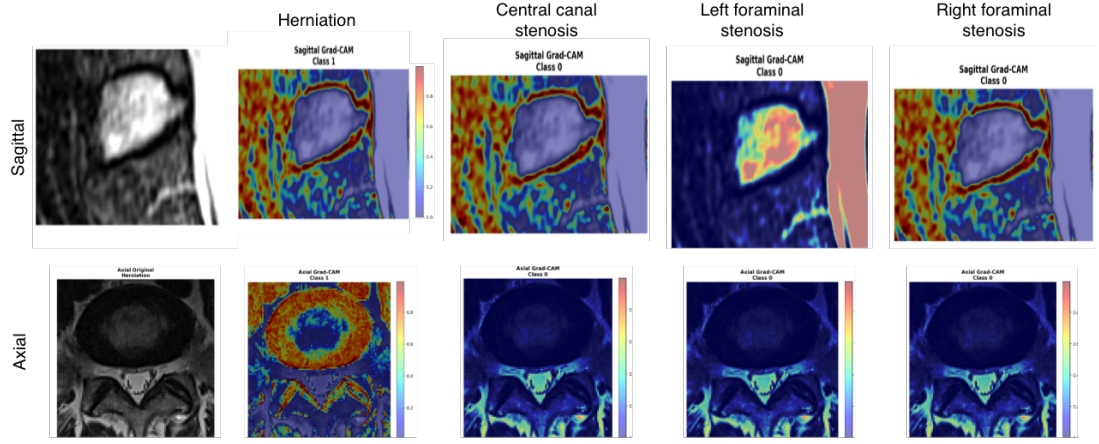


Figure 4.8: GradCAM attention visualizations demonstrating the anatomical focus of the multiview Swin3D model on key lumbar spine pathologies. The visualization is organized as follows: The first column shows the input T2-weighted MRI for reference (L4-L5 disc, sagittal and axial planes). Subsequent columns present the model’s attention heatmaps (red/yellow: high attention, blue: low attention) for specific tasks. **Herniation (hrn)** visualizations localize activation to sites of disc material extrusion. **Central Canal Stenosis (ccs)** maps highlight the central spinal canal region of neural compression. **Left and Right Foraminal Stenosis (fsl, fsr)** activations focus on the respective neural foramina. The heatmaps confirm that the model, trained with a curriculum learning strategy, successfully attends to clinically correct anatomical regions for cross-sectional pathology assessment, validating the clinical alignment and view-integration capability of the proposed multiview architecture.

understanding causal, region-based reasoning, signed LIME is uniquely valuable. Grad-CAM provides structured, intermediate-level feature understanding, while attention rollout, despite computational efficiency, offers only a high-level context that lacks diagnostic spatial specificity. This hierarchy underscores the necessity of using complementary explainability techniques to obtain a complete picture of model behavior.

The findings provide complementary insights into the model’s reasoning at different scales. The robust performance of gradient-based methods confirms the strength of the convolutional backbone in learning precise, discriminative features for localized pathologies. Meanwhile, the attention rollout, by consistently highlighting the intervertebral disc space across tasks, serves a different but crucial function: it provides a high-level validation that the model’s primary attention is correctly anchored to the disc material rather than irrelevant background anatomy. This global anatomical grounding is a necessary foundation for the subsequent fine-grained analysis performed by downstream network layers.

In summary, this analysis serves as a multi-faceted diagnostic tool. It confirms that the model learns clinically sound features through its convolutional foundation, critically identifies a weakness in the transformer’s attention for fine-grained

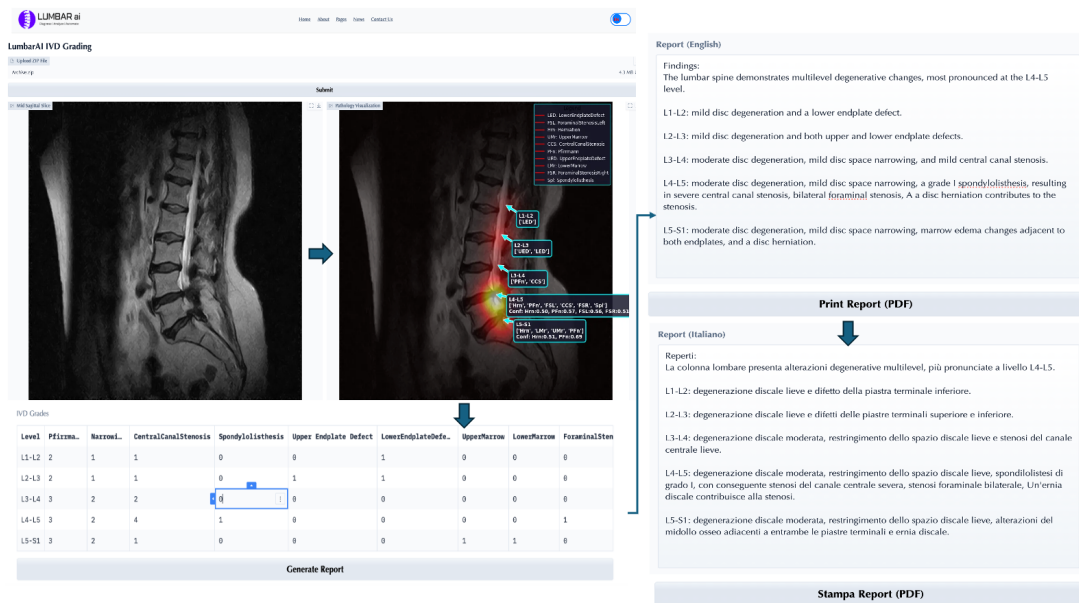


Figure 4.9: *LumbarAI User Interface: A comprehensive Human-in-the-Loop (HITL) workflow for automated intervertebral disc grading. The interface displays: (1) file upload section for DICOM ZIP files, (2) synchronized visualization panels showing mid-sagittal MRI slice (left) and pathology visualization with annotations and confidence scores (right), (3) interactive editable DataFrame containing IVD grades for all five lumbar levels, and (4) bilingual report generation (English and Italian) with PDF export functionality. The reactive visualization system automatically updates when physicians modify grades in the DataFrame, ensuring visual annotations remain synchronized with clinician assessments.*

tasks, and demonstrates the interpretability benefits of a multiview design. This commitment to transparency was extended through the development of a custom patient-level explainability module within the LumbarAI Human-in-the-Loop interface (Figure 4.9). This module synthesizes the model’s 55 individual predictions (11 tasks across 5 discs) into an integrated clinical visualization. It calculates a disc-level severity score by aggregating severe findings (e.g., assigning points for high-grade narrowing, Pfirrmann degeneration, or stenosis). This score dynamically controls a heatmap overlaid on the mid-sagittal MRI, with color intensity indicating cumulative pathological burden at each disc level. Concurrently, short abbreviations for the severe tasks present (e.g., “ccs”, “pf”, “nar”) are labeled with arrows pointing to the affected disc, and a legend displays all tasks with their confidence scores. This design translates complex multi-task predictions into an instant, intuitive anatomical overview, allowing a clinician to quickly grasp the location and composition of severe pathology directly within the diagnostic interface. Furthermore, the interface automatically translates these numeric severity scores into a structured narrative draft report. Both the foundational tabular scores and the generated narrative remain fully editable, embodying the

core Human-in-the-Loop principle.

This interface serves as a preliminary demonstration of how the LumbarAI curriculum learning framework can be operationalize within a real clinical workflow. It aims that this tool should act as a collaborative draft-generator, augmenting rather than replacing radiologist expertise, and provides a functional prototype for subsequent prospective clinical studies.

The primary contribution is therefore fivefold: it (1) establishes a validated multi-method protocol for interpreting spine AI, (2) provides actionable evidence for refining hybrid architectures, (3) demonstrates the necessity of complementary explainability techniques, (4) reveals how multiview integration enhances interpretability, and (5) introduces a novel patient-level explainability synthesis for clinical workflow integration. This work underscores that for high-stakes medical applications, a suite of explainability methods, from granular feature attribution to synthesized clinical summaries, is essential to fully understand, trust, and deploy AI systems effectively.

4.4 Discussion

This chapter presents LumbarAI, a novel curriculum learning framework that addresses a critical limitation in automated lumbar spine MRI analysis: the effective integration of multiview information while maintaining expertise on sagittal-visible tasks. Our results demonstrate that a hybrid architecture, employing separate Transformer-based models for sagittal-only and multiview tasks within a structured four-phase curriculum, achieves clinically significant improvements for cross-sectional pathologies while preserving superior performance for longitudinal pathologies. This approach mirrors and enhances the systematic workflow of expert radiologists, bridging the gap between computational efficiency and clinical intuition.

4.4.1 Clinical Alignment and Workflow Integration

The substantial performance improvements observed for cross-sectional pathologies, particularly foraminal stenosis (29.3%-32.2% F1 improvement) and herniation, validate our framework’s clinical alignment. These results are consistent with established radiological literature emphasizing the necessity of both sagittal and axial views for accurate assessment of neural compression [93, 94, 95, 156]. The human-level performance achieved on central canal stenosis (0.916 accuracy, matching inter-observer agreement) on the independent blind-annotated dataset

represents a particularly meaningful advancement, as this pathology directly influences surgical decision-making [157].

Our hybrid architecture consciously emulates the radiologist’s diagnostic workflow: initial screening of sagittal sequences for alignment and disc-space abnormalities followed by targeted axial evaluation of neural foramina and central canal [158]. This deliberate design choice enhances clinical adoption potential by presenting findings in a familiar cognitive framework. Furthermore, the framework’s capability to generate comprehensive preliminary drafts addresses a pressing clinical need, reducing radiologist burnout while potentially decreasing oversight of subtle findings [159, 160].

The validation on the blinded dataset provides compelling evidence of generalizability beyond the training distribution, a critical consideration for clinical deployment [161]. The observed performance approaching inter-rater agreement for several tasks suggests LumbarAI could function effectively as a collaborative tool rather than a replacement, aligning with emerging consensus on optimal human-AI interaction in medical imaging [161].

4.4.2 Technical Innovations and Comparison with Existing Methods

LumbarAI introduces a structured curriculum learning paradigm for spine MRI that departs from conventional end-to-end training. Its four-phase progression, from axial branch integration to head specialization, incrementally builds multiview competence while limiting catastrophic forgetting, a persistent issue in medical transfer learning [162]. Ablation analyses demonstrate that the complete curriculum improves mean accuracy by 8.3% over direct joint training (from 0.771 to 0.854) and mean F1 score by 3.9% (from 0.789 to 0.828), with contrastive learning playing a decisive role in stabilizing multiclass performance under class imbalance.

External evaluation against SpineNetV2 [48] on the independently annotated XSBC dataset provides a differentiated performance profile. LumbarAI outperformed SpineNetV2 on 6 of 11 tasks by accuracy, including central canal stenosis (ccs: 0.947 vs. 0.863), spondylolisthesis (spn: 0.978 vs. 0.947), and bilateral foraminal stenosis (fsl: 0.810 vs. 0.800; fsr: 0.789 vs. 0.778). In contrast, SpineNetV2 retained advantages in Pfirrmann grading (pf: 0.442 vs. 0.421), disc height narrowing (nar: 0.568 vs. 0.505), upper endplate defects (ued: 0.905 vs. 0.863), and herniation (hrn: 0.684 vs. 0.663). This pattern suggests that LumbarAI’s multi-

view curriculum learning is particularly effective for pathologies requiring spatial integration across sagittal and axial planes, whereas SpineNetV2’s architecture may be better suited for isolated disc morphology assessment. Rather than a categorical superiority, results indicate complementary architectural strengths.

However, to fully recognize LumbarAI’s contribution, it must be situated within the broader landscape of deep learning applications in spine MRI. Recent literature confirms that deep learning achieves expert-level accuracy for isolated spinal tasks. For Pfirrmann grading, Minin et al. [163] introduced SpineScan, a dedicated model achieving strong performance on disc degeneration assessment. Liu et al. [164] developed a comprehensive system evaluated on the RSNA 2024 challenge dataset, while Zheng et al. [165] achieved high-quantitation accuracy specifically for intervertebral disc degeneration. Zhang et al. [166] focused exclusively on lumbar disc herniation detection and classification.

These contributions collectively demonstrate that deep learning can achieve expert-level performance on carefully circumscribed diagnostic tasks. However, they also reveal a fundamental limitation of the current research paradigm: each study addresses a single pathology in isolation, producing highly specialized but narrowly focused tools. From a clinical deployment perspective, this fragmentation presents several insurmountable challenges. First, the sheer number of potential spinal pathologies: disc degeneration, herniation, spinal stenosis (central, lateral recess, and foraminal), endplate changes, Modic changes, spondylolisthesis, and others, would require dozens of independently trained models for comprehensive assessment. Second, each such model demands its own computational resources for training and inference, making simultaneous deployment computationally infeasible in clinical environments where hardware is often constrained. Third, workflow integration becomes complex: a radiologist reviewing a single case would need to sequentially consult multiple AI tools, each generating separate outputs in different formats, defeating the purpose of workflow efficiency.

LumbarAI addresses this fragmentation through a dual-expert, multi-task architecture comprising a Sagittal-Only Expert (seven longitudinal pathologies) and a Multi-View Expert (four cross-sectional pathologies). This design reduces model count (two instead of eleven), enables shared feature representations for cross-pathology reasoning, provides a unified output space, and aligns with clinical reading patterns (sagittal screening followed by axial refinement). Although single-task models may achieve marginally higher ceiling performance, such gains come at the cost of scalability and integration. LumbarAI explicitly prioritizes deployability and resource efficiency, with residual errors mitigated through a human-in-the-loop

framework, where LumbarAI serves as a first-level draft generator and a radiologist subsequently corrects or confirms the proposed drafts.

Our results compare favorably with recent multi-view approaches in spine imaging as well. Park et al. [167] recently proposed a transformer-based model that encodes MRI volumes from multiple sequences and anatomical views, demonstrating that joint multi-view processing outperforms task-specific baselines trained on individual views for spinal stenosis grading. Similarly, Batra et al. [168] introduced M-SCAN, a multistage framework employing multi-view cross attention that achieved an AUROC of 0.971 for spinal canal stenosis grading, surpassing other state-of-the-art methods. For herniation, foraminal and central canal stenosis, LumbarAI achieves F1 scores of 0.80-0.86, which is comparable to the performance reported in these recent multi-view studies. Our curriculum learning strategy extends this work by explicitly addressing the catastrophic forgetting problem [96] that prior studies have noted but not systematically resolved [169].

Building on these comparisons, the trajectory of AI in spine care may benefit from evolving toward integrated assessment systems rather than remaining focused on isolated pathology detectors. The single-pathology models discussed above [163, 164, 165, 166] have advanced the field by demonstrating feasibility and establishing valuable performance benchmarks. Building on this foundation, the next frontier, one that LumbarAI seeks to explore, is the synthesis of these capabilities into unified frameworks that can better support the complexity of real-world clinical decision-making. Emerging multi-view architectures [167, 168] further suggest the promise of this direction, and LumbarAI’s curriculum learning approach offers a potential solution to the stability-plasticity challenge inherent in multi-view integration.

4.4.3 Limitations and Methodological Considerations

Several limitations must be acknowledged to contextualize our findings appropriately. First, the fundamental design choice of employing multi-task, multi-class models inherently involves a trade-off. While training dedicated single-task models for each of the 11 pathologies would likely yield marginal gains in individual task accuracy, such an approach is clinically impractical due to prohibitive computational resource demands and workflow complexity. Our philosophy prioritizes clinical feasibility and operational efficiency, creating a minimal number of expert models that generalize well across the diagnostic spectrum. This pragmatic compromise accepts a potential ceiling on per-task performance in exchange for a unified, resource-efficient system that aligns with real-world clinical deployment

constraints. This limitation is planned to be actively mitigated by the Human-in-the-Loop (HITL) framework, where the AI serves as a proficient draft generator, and the radiologist’s expertise corrects any residual inaccuracies, ensuring final diagnostic quality.

Second, the anatomical scope of this work is intentionally focused. The LumbarAI model was developed and validated exclusively for the lumbar spine (L1-L2 to L5-S1), which is anatomically the most vulnerable to biomechanical stress and degenerative changes, representing the highest clinical burden of degenerative spinal disease. While this focus aligns with the greatest clinical need and provides a robust proof-of-concept, it inherently limits the model’s immediate applicability to the cervical and thoracic (dorsal) spinal regions. Future studies would benefit from extending this curriculum learning framework to incorporate these areas, which present distinct anatomical contexts and pathological spectra.

Methodologically, the scope of our study is defined by both its input data and diagnostic tasks. This work focused exclusively on T2-weighted sagittal and axial sequences. The incorporation of other standard MRI sequences, such as coronal views, T1-weighted scans, and fluid-sensitive sequences like STIR, was not explored. These sequences provide complementary information, for example, T1-weighted images offer superior anatomical detail for marrow evaluation, while STIR sequences are highly sensitive for detecting bone marrow edema, and their integration could potentially enhance diagnostic accuracy and model robustness for specific pathologies. Future work could explore a truly multi-parametric input framework to leverage this richer diagnostic context.

Similarly, the model’s diagnostic capability is currently bounded by the 11 specific radiological tasks on which it was trained and validated. While this set covers the most common and clinically impactful degenerative findings, the spectrum of lumbar spine pathology is vast. Incorporating additional conditions, such as facet joint arthropathy, ligamentous hypertrophy, or rare entities like infection or tumor, would require a substantial, expert-annotated dataset for each new task. We acknowledge this limited but focused task scope; extending the model’s diagnostic scope remains a significant future challenge contingent upon the availability of high-quality annotated data.

Furthermore, the use of silver-standard labels from SpineNet for training may also constrain models to learn within SpineNet’s decision boundaries, potentially limiting the discovery of novel diagnostic features. Future work with comprehensively annotated multiview and multi-sequence datasets would further validate

and extend our approach.

4.4.4 Future Directions

The pathway for advancing this research is directly outlined by the limitations acknowledged in this work. The most immediate future directions, including extending the model to cervical and thoracic anatomy, incorporating additional MRI sequences (T1-weighted, STIR), expanding the spectrum of detectable pathologies, and pursuing broader multi-center external validation, have been detailed in the preceding section.

Beyond these extensions, the paramount next step is to transition from technical validation to prospective clinical impact assessment. This necessitates the design and execution of rigorous real-world clinical trials to evaluate the LumbarAI system’s effect on diagnostic workflow efficiency, report consistency, and its ultimate influence on clinical decision-making pathways and patient outcomes. Such studies are essential to generate the evidence required for regulatory approval and successful integration into standard clinical practice.

5

Psychosocial Determinants and Phenotypic Stratification in Spinal Pain

5.1 Introduction

The previous chapter introduced LumbarAI as a workflow-aligned Computer Vision Module, fulfilling the first requirement of the integrative framework. Yet anatomical findings alone are insufficient for personalized spine care. The clinical imperative, derived from stakeholder needs in Section 3.2, is to integrate imaging data with the patient’s lived experience—their pain, disability, quality of life, and coping strategies—into a complete biopsychosocial profile. This chapter develops the methodological foundation for that synthesis, constituting Module 2 of the framework: the **Multimodal Profiling Module**.

In the context of chronic pain, coping refers to the cognitive and behavioral strategies individuals employ to manage pain-related demands and their emotional, functional, and social consequences. These strategies do not reflect pain intensity itself, but rather how individuals respond to and regulate their experience of pain. Despite their routine clinical use, a critical gap identified in both the literature and clinical practice is the lack of a systematic, data-driven understanding of how these psychosocial constructs interact with each other and with anatomical pathology. Current practice often treats Patient Reported Outcome Measures (PROMs) as isolated scores, thereby overlooking the complex patterns through which coping strategies co-occur and define a patient’s overall adaptive or maladaptive phenotype [32, 36].

Addressing these gaps is essential for the development of a functional Multimodal Profiling Module. Accordingly, this chapter focuses on two sequential methodological questions:

1. **Which coping strategies are most strongly associated with pain intensity?** To address this question, we examine how individual coping strategies relate to pain severity using multivariable regression, followed by interaction and mediation analyses to clarify whether these associations are independent, additive, or primarily indirect.
2. **Can we identify meaningful subgroups?** We then employ unsupervised learning to ask whether patients naturally cluster into distinct, clinically interpretable **psychosocial phenotypes**. This moves beyond predicting single variables to identifying holistic patient profiles (e.g., “*low pain, low disability, adaptive coping, and high quality of life*” versus “*high pain, high disability, maladaptive coping, and low quality of life*”), which may reflect shared risk trajectories and distinct treatment needs. This phenotypic approach directly operationalizes the clinical need for “Supporting Clinical Reasoning” (Section 3.2) by providing a shared, quantitative taxonomy for interdisciplinary discussion.

Objectives and Chapter Structure

The objective of this chapter is to establish and critically evaluate the analytical foundations of the Multimodal Profiling Module through the integration of real-world patient-reported outcomes and imaging-derived information. Specifically, this chapter aims to:

- Investigate the role of coping strategies as determinants of pain intensity using supervised statistical modelling.
- Extend the analytical scope to the full PROMs dataset to derive biopsychosocial patient phenotypes through unsupervised clustering.
- Examine the clinical relevance of these phenotypes by comparing them against baseline characteristics, including structural findings extracted from spine MRI examinations.

The chapter is organized to reflect this progression. Section 5.2 introduces the dataset, describes the PROMs instruments, and outlines the supervised and unsupervised analytical approaches. Section 5.3 presents the empirical findings, beginning with the determinants of pain and subsequently detailing the identification and characterization of psychosocial phenotypes. Finally, Section 5.4

synthesizes these results, discusses their clinical implications for stratified spinal pain management, and clarifies how the derived phenotypes interface with the Computer Vision Module (Chapter 4) to enable the multimodal integration central to the framework.

5.2 Dataset and Methods

5.2.1 Study Design and Setting

This chapter reports two complementary cross-sectional studies conducted on a single clinical cohort, integrating (1) the analysis of coping strategies as determinants of pain intensity, and (2) the identification of biopsychosocial phenotypes using unsupervised clustering. Both components used a cross-sectional design and shared the same patient population.

Data were collected consecutively from adults with spine-related pain in a multidisciplinary clinic located in Brescia, Italy: X-Ray Service s.r.l. [129]. The recruitment took place between August 2024 and September 2025. All assessments, questionnaires, and radiological data (when available) were collected during routine clinical visits by trained personnel following standardized protocols.

5.2.2 Ethical Considerations

All procedures were conducted in full compliance with the Declaration of Helsinki and the European General Data Protection Regulation (GDPR). Written informed consent was obtained from all participants prior to inclusion. Ethical approval was granted through institutional review processes, including a formal joint data controllership agreement between the University of Brescia and X-Ray Service s.r.l. (Protocol UNIBS 253697, 27/09/2024), and a Data Protection Impact Assessment (DPIA) approved on 01/08/2025. Patient data were pseudonymized by the clinical sites before transfer for analysis.

5.2.3 Participants

A total of 142 adults with spine-related pain affecting the cervical, thoracic, lumbar, or sacrococcygeal regions were included. Eligibility criteria required individuals to be (1) 18 years or older, (2) diagnosed with spine-related pain, and (3) able to understand and complete the study questionnaires independently. Exclusion criteria included severe neurologic disorders, acute psychiatric crises, or inability to comprehend study materials.

The final cohort consisted of 85 males (59.9%) and 57 females (40.1%), with a mean age of 51.8 years (SD = 12.3). For analyses involving radiology data, a subset of 129 participants was retained due to missing radiological information in 13 cases.

5.2.4 Measures

Pain intensity

Pain intensity was measured using the Numerical Rating Scale (NRS), ranging from 0 (“no pain”) to 10 (“worst imaginable pain”) [170, 171]. The location (cervical, thoracic, lumbar, sacrococcygeal, or combinations thereof) and duration (categorized as < 4 weeks, 4–12 weeks, > 12 weeks, or as specifically provided) of pain were recorded.

Functional Disability

Functional disability was assessed using the Italian version of the Roland–Morris Disability Questionnaire [22], a validated adaptation of the original instrument [172]. Scores range from 0 to 24, with higher scores indicating greater disability.

Quality of Life

Quality of life was measured using the Italian WHOQOL-BREF questionnaire [173], a validated Italian version of the World Health Organization Quality of Life-BREF instrument [174]. This instrument assesses four core domains, Physical Health, Psychological Health, Social Relationships, and Environment, and provides overall QoL and health satisfaction scores. All scores range from 0 to 100, with higher scores indicating better quality of life.

Coping Strategies

Coping strategies were assessed with the CopeNVI questionnaire [26], the validated Italian short form of the COPE instrument [175, 176]. Its use for pain populations is well-established in several studies [177, 178, 179]. The CopeNVI evaluates five domains: Social Support, Avoidance, Positive Attitude, Problem Orientation, and Transcendent Orientation. Responses for each domain were recorded on a four-point Likert scale (from 1: *I usually don't do it* to 4: *I almost always do it*), with higher scores indicating greater use of the respective strategy. Using this standardized Italian version ensured cultural relevance for our exclusively Italian-speaking cohort.

5.2.5 Data Preprocessing

Several preprocessing steps were performed to ensure data quality and suitability for subsequent analysis. All raw questionnaire responses were converted into processed numeric variables following their respective standardized manuals to generate continuous, psychometrically valid variables.

Specifically, for the **WHOQOL-BREF**, the instrument’s 26 items were processed according to official WHO guidelines to construct four core domain scores (Physical Health, Psychological Health, Social Relationships, Environment) and two overall perception scores. Each item, originally on a 1–5 Likert scale, contributed to its respective domain. Domain raw scores were then transformed to a standardized 0–100 scale using the prescribed formula (Eq. 5.1) to ensure comparability with the full WHOQOL-100. Consequently, higher scores on this scale indicate a better quality of life.

$$\text{Transformed Score} = \frac{(\text{Actual Raw Score} - \text{Lowest Possible Raw Score})}{\text{Possible Raw Score Range}} \times 100 \quad (5.1)$$

For the **CopeNVI (25-item version)**, responses for each of the five subscales (Social Support, Avoidance, Positive Attitude, Problem Orientation, Transcendent Orientation) were recorded on the instrument’s 4-point Likert scale (1 = *I usually don’t do it* to 4 = *I almost always do it*). For analysis, scores for each subscale were averaged across its constituent items, yielding five continuous subscale scores. Higher average scores indicate a greater reported use of that specific coping strategy.

Functional disability was quantified as the total score from the Roland–Morris Disability Questionnaire, calculated by summing the number of ‘YES’ responses across its 24 items, resulting in a range of 0–24. Finally, **pain intensity** was represented directly by the patient’s self-reported Numeric Rating Scale (NRS) score (0–10). These transformations produced a final dataset of clinically interpretable, continuous variables ready for statistical and clustering analyses.

The clinical dataset was merged with radiological variables to form the full biopsychosocial dataset. No missing values were present among the final clustering features. Feature selection for clustering was theoretically driven by the biopsychosocial model, prioritizing modifiable clinical and psychosocial variables. Twelve continuous variables were retained: pain intensity (NRS), functional disability (Roland score), the five WHOQoL-BREF domains, and the five CopeNVI coping

dimensions. Demographic, categorical (pain location/duration), and radiological variables were excluded from clustering but retained for subsequent cluster characterization. All continuous variables were standardized using z-score normalization (mean = 0, SD = 1). While Principal Component Analysis (PCA) was used for visualization, clustering algorithms operated on the full set of standardized variables to preserve clinical interpretability.

Coping Determinants

Associations between pain intensity and coping strategies were examined using Pearson and Spearman correlation coefficients [180]. Multiple linear regression models were constructed to estimate the independent contributions of each coping dimension to pain intensity. Interaction effects were tested to evaluate whether the association between avoidance and pain varied across levels of positive attitude. Mediation analyses followed the classical Baron and Kenny approach [181], with indirect effects tested using Sobel statistics.

Unsupervised Biopsychosocial Phenotyping

To identify data-driven patient phenotypes, unsupervised clustering techniques were applied using both K-means clustering [182] and hierarchical agglomerative clustering with Ward's linkage [183]. Cluster quality and internal validity were assessed through multiple metrics, including the silhouette coefficient, within-cluster sum of squares (WCSS), and the Davies–Bouldin index [184]. Convergence and agreement between the two clustering approaches were quantified using the adjusted Rand index [185], ensuring methodological consistency. To further evaluate the robustness of the derived cluster solution, a bootstrap stability analysis (1,000 iterations) was conducted, providing an estimate of the reproducibility of cluster assignments under resampling [185].

5.2.6 Cluster Characterization

Cluster profiles were described using means, medians, standard deviations, and frequency distributions. Group comparisons for continuous variables employed Shapiro–Wilk tests for normality [186] and Levene's tests for homogeneity of variances [187]. Independent-samples t-tests, Welch's t-tests, or Mann–Whitney U tests were applied as appropriate [188, 189]. Categorical variables were compared using chi-square tests or Fisher's exact tests [190, 191]. When analyzing multiple radiological variables, both Bonferroni [192] and Benjamini–Hochberg false discovery rate corrections [193] were applied. Effect sizes were computed using Cohen's

d for continuous variables [194] and Cramér’s V for categorical variables [195].

5.3 Results

5.3.1 Coping Determinants of Pain

The initial analysis focused on delineating the relationships between coping strategies and pain intensity, progressing from simple associations to complex multivariate models to establish the primary determinants.

5.3.1.1 Descriptive Statistics and Bivariate Associations Between Coping Strategies and Pain Intensity

Descriptive statistics (the means and standard deviations) for all CopeNVI subscales and Pain NRS scores across the cohort are reported in Table 5.1. Participants reported a mean pain intensity of 5.89 (SD = 2.77) on the 0–10 Numerical Rating Scale, with scores distributed across the full measurement range and no evident floor or ceiling effects.

Table 5.1: Means and Standard Deviations for all CopeNVI subscales and NRS pain scores

Variable	Mean (SD)	N
Social Support	2.47 (0.79)	142
Avoidance	1.74 (0.92)	142
Positive Attitude	2.72 (1.04)	142
Problem Orientation	2.88 (0.81)	142
Transcendent	1.64 (0.78)	142
Pain NRS	5.89 (2.77)	142

Bivariate correlation analyses revealed several statistically significant associations between pain intensity and CopeNVI subscales (Table 5.2). Avoidance coping demonstrated a strong positive association with pain intensity ($r = 0.601$, $p < 0.001$), indicating higher pain levels among individuals who more frequently engaged in avoidant strategies. In contrast, Positive Attitude exhibited a strong inverse relationship with pain intensity ($r = -0.541$, $p < 0.001$), suggesting a potential protective role of adaptive cognitive coping. Social Support was moderately and negatively correlated with pain intensity ($r = -0.323$, $p < 0.001$). Problem Orientation and Transcendent coping showed weak associations that did not reach statistical significance under the predefined threshold, indicating a

more limited or indirect relationship with pain intensity in this sample. Figure 5.1 illustrates these bivariate relationships through scatterplots with regression lines and coefficients, visually confirming the linear patterns, particularly the strong association between Avoidance and higher pain intensity.

Table 5.2: *Correlation matrix of pain intensity (NRS) with CopeNVI subscales*

CopeNVI Subscale	Pearson's r	p-value	N
Social Support	-0.323	<0.001	142
Avoidance	0.601	<0.001	142
Positive Attitude	-0.541	<0.001	142
Problem Orientation	-0.183	0.050	142
Transcendent	-0.049	0.050	142

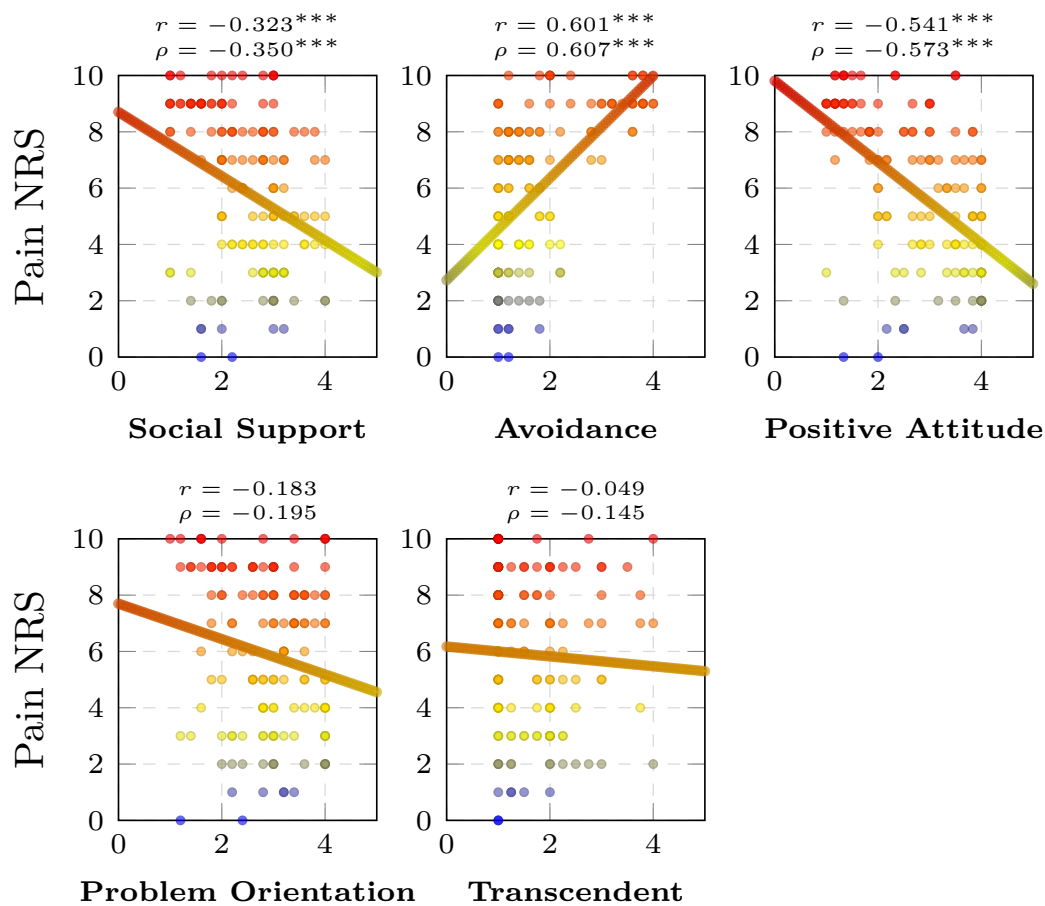


Figure 5.1: *Scatterplots illustrating bivariate relationships between pain intensity (NRS) and each CopeNVI subscale. Regression lines and correlation coefficients (r) are displayed. *** indicates $p < 0.001$.*

5.3.1.2 Uni-variate Regression Analysis

To quantify the strength of association between individual coping strategies and pain intensity, a series of univariable linear regression models were estimated with pain intensity (NRS) as the dependent variable and each CopeNVI subscale entered separately as an independent predictor. These analyses extend the bivariate correlations by estimating the proportion of variance in pain intensity attributable to each coping dimension.

Table 5.3: *Regression coefficients for pain intensity (NRS) predicted by CopeNVI subscales*

CopeNVI subscales	Slope	Intercept	R ²	N
Social Support	-1.139	8.702	0.105	142
Avoidance	1.813	2.729	0.362	142
Positive Attitude	-1.441	9.809	0.293	142
Problem Orientation	-0.628	7.696	0.034	142
Transcendent	-0.175	6.174	0.002	142

Prior to model estimation, the distributional properties of pain intensity were visually examined to assess the suitability of parametric analyses. Histograms and Q-Q plots were used to evaluate overall distributional shape and departures from normality, box plots were inspected to identify potential outliers, and residual diagnostic plots were examined to verify model assumptions following regression fitting. Pain intensity scores approximated a normal distribution, with only minimal outliers observed. Residual analyses from the fitted regression models indicated acceptable model fit, with residuals symmetrically distributed around zero and no evident violations of linearity or homoscedasticity assumptions (Figure 5.2).

Avoidance coping emerged as the strongest explanatory factor, accounting for 36.2% of the variance in pain intensity ($R^2 = 0.362$). The positive regression coefficient ($\beta = 1.813$) indicates that higher reliance on avoidant coping strategies was associated with substantially greater reported pain (Table 5.3). Positive Attitude demonstrated the second-largest contribution, explaining 29.3% of the variance ($R^2 = 0.293$), with a negative coefficient ($\beta = -1.441$), consistent with a protective association against pain intensity. Social Support accounted for a more modest proportion of variance ($R^2 = 0.105$; $\beta = -1.139$).

In contrast, Problem Orientation and Transcendent coping showed limited explanatory value, accounting for only 3.4% and 0.2% of the variance in pain intensity, respectively (Table 5.3). These findings suggest that while multiple coping strate-

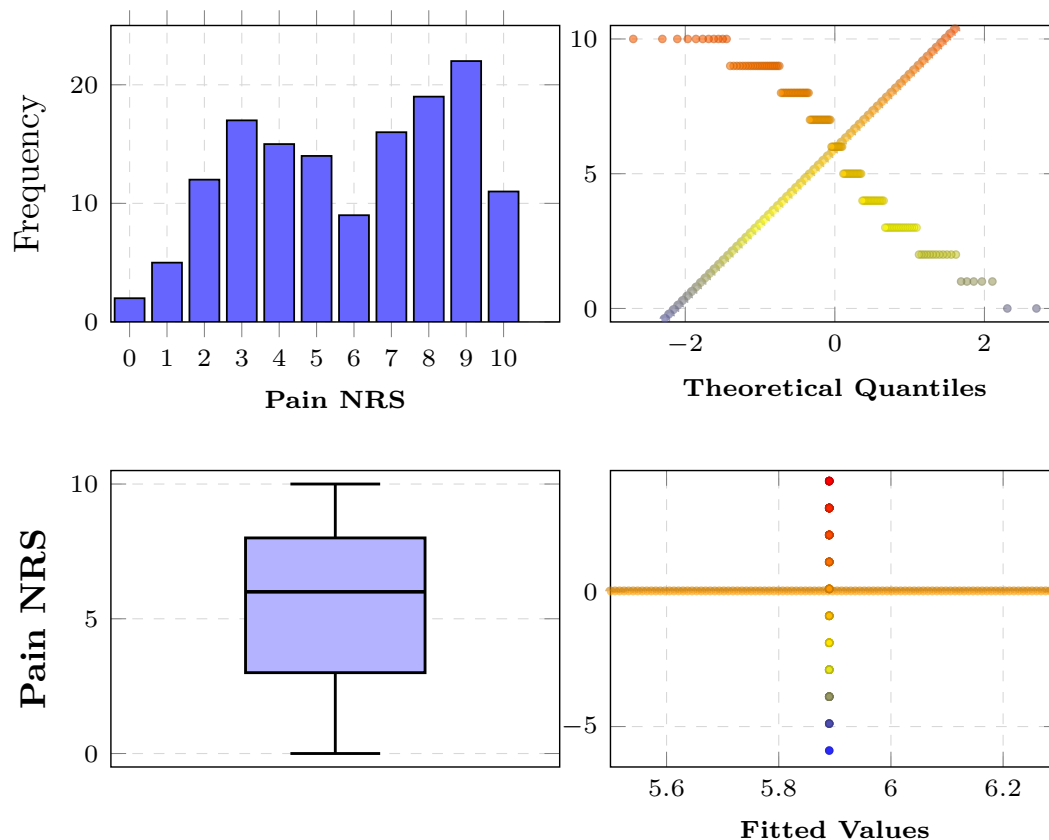


Figure 5.2: Outlier and normality assessment of pain intensity (NRS) scores. Top left: Histogram showing distribution. Top right: Q-Q plot for normality assessment. Bottom left: Box plot showing quartiles and outliers. Bottom right: Residuals plot for regression diagnostics.

gies are theoretically relevant, their empirical univariate associations with pain intensity differ markedly in strength.

Taken together, the univariate regression results identify Avoidance and Positive Attitude as the most salient coping dimensions in relation to pain intensity.

5.3.1.3 Multiple Regression Analysis

Building on the univariable analyses, a multivariable linear regression model was estimated to identify the independent associations between coping strategies and pain intensity while controlling for shared variance among coping dimensions. All five CopeNVI subscales were entered simultaneously as predictors of pain intensity (NRS). The overall model was statistically significant ($F(5, 142) = 19.18$, $p < 0.001$) and explained 41.4% of the variance in pain intensity ($R^2 = 0.414$, adjusted $R^2 = 0.392$), representing a large effect size and confirming coping strategies as a robust set of predictors (Table 5.4).

Table 5.4: Multiple regression analysis: All CopeNVI subscales predicting pain intensity (NRS)

Predictor	B (SE)	β	p-value	95% CI
Social Support	0.244 (0.230)	0.069	0.100	[-0.207, 0.696]
Avoidance	1.428 (0.197)	0.474	0.001**	[1.041, 1.815]
Positive Attitude	-0.781 (0.174)	-0.293	0.001**	[-1.122, -0.439]
Problem Orientation	0.187 (0.224)	0.055	0.100	[-0.251, 0.626]
Transcendent	-0.187 (0.232)	-0.053	0.100	[-0.642, 0.269]

*Model $R^2 = 0.414$, Adjusted $R^2 = 0.392$, $F(5, 142) = 19.182$, ** : statistically significant*

Independent Predictors: Avoidance and Positive Attitude

Among the five coping strategies, only Avoidance and Positive Attitude emerged as significant independent predictors of pain intensity. Avoidance demonstrated the strongest association ($\beta = 0.474$, $B = 1.428$, $SE = 0.197$, $p < 0.001$, 95% CI [1.041, 1.815]). Each one-point increase on the Avoidance scale was associated with an estimated 1.43-point increase on the 0–10 pain NRS. Although attenuated relative to its bivariate correlation ($r = 0.601$), this finding confirms a robust independent association.

Positive Attitude showed a significant protective effect ($\beta = -0.293$, $B = -0.781$, $SE = 0.174$, $p < 0.001$, 95% CI [-1.122, -0.439]). Each one-point increase predicted an approximate 0.78-point reduction in pain intensity, indicating a stable independent effect after accounting for the influence of other coping strategies.

Non-Significant Predictors: Social Support, Problem Orientation, and Transcendent Coping

The remaining coping strategies did not contribute unique explanatory value in the multivariable model. Social Support ($\beta = 0.069$, $p = 0.100$) shifted from a significant negative bivariate correlation to a non-significant positive coefficient, suggesting a possible suppression effect whereby its apparent protective association may be mediated by other coping dimensions, particularly Positive Attitude [196].

Similarly, Problem Orientation ($\beta = 0.055$, $p = 0.100$) did not retain independent significance, indicating that its relationship with pain intensity may be largely accounted for by overlap with other adaptive coping strategies, consistent with models emphasizing indirect pathways via self-efficacy or cognitive reappraisal [179]. Transcendent coping showed minimal association with pain intensity ($\beta = -0.053$, $p = 0.100$), suggesting no unique contribution within this model.

Overall, these findings identify Avoidance and Positive Attitude as the primary

independent coping determinants of pain intensity. This clarification provides a strong empirical foundation for subsequent analyses examining their interactive and mediating roles within the broader biopsychosocial framework.

5.3.1.4 Mediation Analysis

Building on the regression findings, which identified Avoidance and Positive Attitude as the primary independent predictors of pain intensity, while Social Support and Problem Orientation showed only bivariate associations, we conducted a series of mediation analyses. Mediation analysis is used to decompose an observed association into direct and indirect components, allowing assessment of whether the effect of a given predictor on an outcome operates primarily through another intervening variable rather than directly. In this context, the analyses were designed to test whether more distal coping strategies (e.g., Social Support, Problem Orientation) influence pain intensity indirectly by shaping more proximal coping processes, such as Avoidance and Positive Attitude, thereby clarifying the functional pathways within the coping system.

All mediation models were estimated using ordinary least squares path analysis. For each model, we report the total effect (c), and the indirect effect ($a \times b$), with bootstrapped standard errors (5,000 resamples) and 95% confidence intervals. Results are summarized in Table 5.5, and the principal pathways are illustrated in Figure 5.3.

Dominant Indirect Pathways: Problem Orientation as a Foundational Strategy

The largest and most consistent mediation effects identified *Problem Orientation* as a pivotal but predominantly indirect adaptive resource. Its relationship with pain intensity was almost entirely transmitted through the two core determinants identified in the regression analysis, Positive Attitude and Avoidance, indicating minimal direct influence.

- **Problem Orientation → Positive Attitude → Pain.** Higher Problem Orientation was associated with higher Positive Attitude (path $a = 0.495$, $p < 0.001$), which in turn predicted lower pain intensity (path $b = -1.441$, $p < 0.001$). The resulting indirect effect was substantial ($a \times b = -0.713$, $p < 0.001$), accounting for **113.5%** of the total effect. This proportion exceeding 100% indicates a suppression pattern, whereby the indirect protective pathway is stronger than the observed total effect.

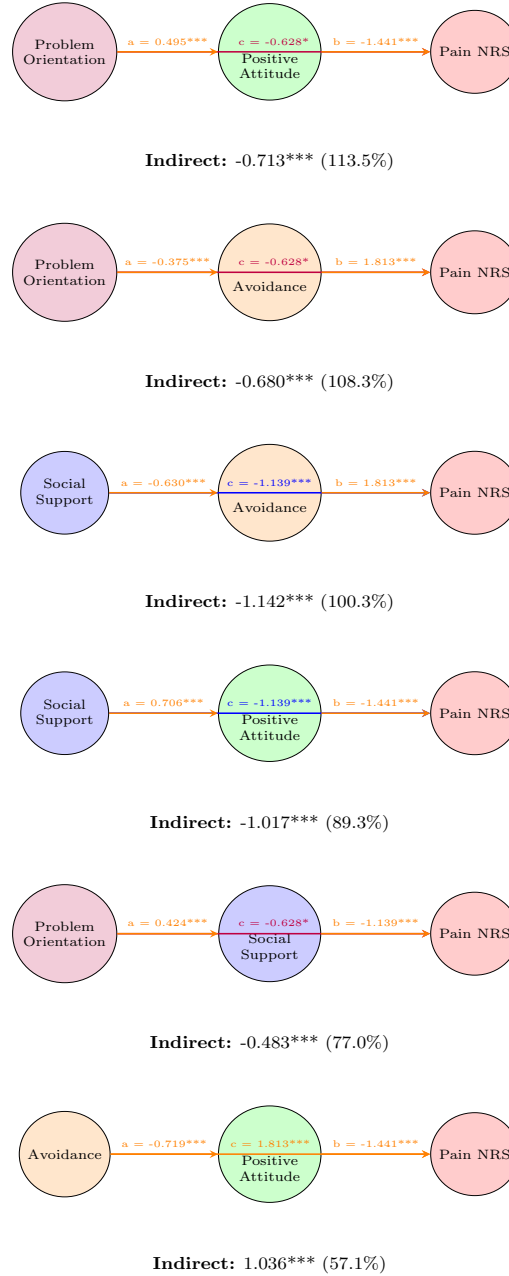


Figure 5.3: Synthesized mediation model depicting primary pathways among coping strategies and pain intensity. Dominant indirect effects are shown, including: Problem Orientation’s protective effect being mediated by Positive Attitude (113.5% of total effect) and Avoidance (108.3%); Social Support’s effect mediated by Avoidance (100.3%) and Positive Attitude (89.3%); and the reciprocal mediation between Avoidance and Positive Attitude (57.1% and 31.9%, respectively). Solid lines represent significant paths ($*** p < 0.001$, $** p < 0.01$, $* p < 0.05$).

- **Problem Orientation → Avoidance → Pain.** A parallel pathway showed that higher Problem Orientation predicted lower Avoidance ($a = -0.375$, $p < 0.001$), which strongly predicted higher pain ($b = 1.813$, $p < 0.001$). This indirect effect ($a \times b = -0.680$, $p < 0.001$) accounted for **108.3%** of the total

effect, again indicating near-complete mediation.

Together, these findings indicate that Problem Orientation does not directly regulate pain, but instead functions as a foundational strategy that exerts its protective influence by promoting adaptive cognitive appraisal and reducing maladaptive avoidance.

Table 5.5: *Mediation analysis: Coping strategies as mediators between other coping strategies and pain intensity*

Predictor → Mediator → Outcome	a (SE)	b (SE)	Indirect (a×b)	% Med.
PO → PA → Pain	0.495 (0.100)	-1.441 (0.189)	-0.713 (0.172)	113.5%***
PO → AV → Pain	-0.375 (0.091)	1.813 (0.204)	-0.680 (0.181)	108.3%***
SS → AV → Pain	-0.630 (0.083)	1.813 (0.204)	-1.142 (0.198)	100.3%***
SS → PA → Pain	0.706 (0.094)	-1.441 (0.189)	-1.017 (0.191)	89.3%***
PO → SS → Pain	0.424 (0.074)	-1.139 (0.282)	-0.483 (0.146)	77.0%***
AV → PA → Pain	-0.719 (0.074)	-1.441 (0.189)	1.036 (0.173)	57.1%***
PA → SS → Pain	0.404 (0.054)	-1.139 (0.282)	-0.460 (0.129)	31.9%***
AV → SS → Pain	-0.462 (0.061)	-1.139 (0.282)	0.526 (0.147)	29.0%***
SS → PO → Pain	0.449 (0.078)	-0.628 (0.284)	-0.282 (0.137)	24.8%*
PA → PO → Pain	0.300 (0.061)	-0.628 (0.284)	-0.188 (0.093)	13.1%*

*Note: Abbreviations: PO = Problem Orientation, PA = Positive Attitude, AV = Avoidance, SS = Social Support, Pain = Pain Intensity (NRS), *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.*

Social Support as a Facilitator of Proximal Coping

A closely analogous pattern was observed for *Social Support*, which likewise showed no meaningful direct association with pain once indirect pathways were accounted for. Instead, its protective effects were fully transmitted through the same proximal determinants.

- **Social Support → Avoidance → Pain.** Greater Social Support predicted lower Avoidance ($a = -0.630$, $p < 0.001$), which in turn predicted higher pain ($b = 1.813$, $p < 0.001$). This pathway constituted the strongest indirect effect in the entire model ($a \times b = -1.142$, $p < 0.001$), accounting for **100.3%** of the total effect.
- **Social Support → Positive Attitude → Pain.** Social Support also strongly predicted higher Positive Attitude ($a = 0.706$, $p < 0.001$), which predicted lower pain ($b = -1.441$, $p < 0.001$). This indirect effect ($a \times b = -1.017$, $p < 0.001$) accounted for **89.3%** of the total effect.

These results indicate that Social Support operates indirectly, facilitative resource whose apparent benefit for pain is almost entirely explained by its capacity to reduce avoidance and enhance positive cognitive coping.

Reciprocal Mediation Among Core Determinants

Significant reciprocal pathways were also observed among the core determinants of pain, indicating a dynamically interdependent coping system.

- **Avoidance → Positive Attitude → Pain.** Higher Avoidance predicted lower Positive Attitude ($a = -0.719, p < 0.001$), which in turn predicted lower pain ($b = -1.441, p < 0.001$). This produced a positive indirect effect ($ab = 1.036, p < 0.001$), indicating that **57.1%** of Avoidance’s detrimental association with pain is mediated through its erosion of positive cognitive appraisal.
- **Positive Attitude → Social Support → Pain.** A secondary reciprocal pathway showed that higher Positive Attitude predicted greater Social Support ($a = 0.404, p < 0.001$), which in turn predicted lower pain ($b = -1.139, p < 0.001$). This indirect effect ($ab = -0.460, p < 0.001$) accounted for **31.9%** of the total effect.

These findings indicate that Positive Attitude and Avoidance are not only direct regulators of pain but also mutually embedded within broader cognitive–social feedback loops.

Secondary and Minor Mediation Pathways

Several additional mediation effects were statistically significant but comparatively smaller in magnitude, further refining the network structure.

- **Problem Orientation → Social Support → Pain** accounted for **77.0%** of the total effect ($ab = -0.483, p < 0.001$), suggesting that Problem Orientation may partially exert its influence by facilitating support-seeking behaviors.
- **Avoidance → Social Support → Pain** accounted for **29.0%** mediation ($ab = 0.526, p < 0.001$), indicating that avoidance coping is associated with social withdrawal, which in turn exacerbates pain.
- Pathways involving *Problem Orientation as a mediator* were consistently weaker. Both **Social Support → Problem Orientation → Pain** (24.8% mediated, $p < 0.05$) and **Positive Attitude → Problem Orientation → Pain** (13.1% mediated, $p < 0.05$) accounted for a relatively small proportion of total effects, reinforcing the interpretation of Problem Orientation as primarily antecedent rather than downstream within the coping hierarchy.

Synthesis: A Hierarchical Coping Architecture

Taken together, the mediation results reveal a structured and hierarchical organization of coping strategies in relation to pain intensity:

1. **Foundational (Distal) Strategies:** *Problem Orientation* and *Social Support* exert strong protective influences almost entirely through indirect pathways, shaping downstream coping responses rather than acting directly on pain.
2. **Proximal Determinants:** *Avoidance* and *Positive Attitude* represent the final common pathways through which the coping system influences pain intensity. These strategies show both strong direct effects and substantial reciprocal mediation.
3. **Directional Asymmetry:** The substantially stronger mediation effects from foundational strategies to proximal determinants, compared with the reverse, clarify the functional direction of influence within the coping system.

This hierarchical architecture explains why Avoidance and Positive Attitude emerged as independent predictors in the regression models: they are the proximal mechanisms through which broader coping resources ultimately affect pain experience. Clinically, these findings suggest that interventions targeting Avoidance and Positive Attitude may yield the most immediate reductions in pain, while strategies enhancing Problem Orientation and Social Support may serve as upstream mechanisms facilitating durable change.

5.3.2 Biopsychosocial Phenotyping via Unsupervised Clustering

The mediation analyses delineated the functional hierarchy within the coping system, illustrating how specific strategies exert influence on pain via both direct and indirect pathways. Yet, chronic pain is inherently multidimensional, with coping representing only one component of a broader biopsychosocial landscape encompassing functional disability, psychological distress, and overall quality of life.

This raises a critical, unresolved question: when considering the full spectrum of clinical and psychosocial variables simultaneously, *do patients naturally cluster into coherent subgroups with distinct, multimodal profiles?*

To address this, we shift from a variable-centered, hypothesis-driven approach focused on coping to a person-centered, data-driven exploration of patient phenotypes. By applying unsupervised clustering across all measured domains, including pain intensity, disability, quality-of-life subdomains, and coping subscales, we examine whether patients clustered into clinically meaningful subgroups and assess whether the central influence of Avoidance and Positive Attitude persists within these emergent, holistic profiles.

5.3.2.1 Identification of Patient Phenotypes

Cluster Tendency Assessment. Before applying clustering algorithms, we evaluated whether the biopsychosocial dataset contained inherent structure using the Hopkins statistic. The resulting value of $H = 0.644 \pm 0.020$ ($n = 142$) indicated moderate-to-strong cluster tendency, consistent with the presence of non-random structure ($H > 0.5$).

Determination of Optimal Cluster Number. Cluster validity was assessed through complementary indices. The elbow method (Figure 5.4A) revealed a monotonic decline in within-cluster sum of squares (WSS) without a sharp inflection point. The most largest reduction (31.3%) occurred at $k = 2$, with progressively attenuating gains at $k = 3$ (13.7%) and $k = 4$ (10.4%).

Silhouette analysis as shown in Figure 5.4C, demonstrated systematic improvement across successive analytical refinements. The initial solution (129 patients, 12 variables) yielded a coefficient of 0.259, falling within the weak-structure range. Following feature engineering, specifically, removal of the non-discriminatory Transcendent Coping subscale, and expansion to the full 142-patient sample, the silhouette coefficient improved to 0.313. When the 19 patients occupying the transition zone (individual silhouette < 0.25) were temporarily excluded, the two core poles achieved a coefficient of 0.411, representing a 31.2% improvement and crossing into the moderate-to-strong separation range. To verify that the underlying clustering signal was not artificially sustained by these borderline cases, the Hopkins statistic was recomputed on the reduced sample; the value remained effectively unchanged ($H = 0.645 \pm 0.022$, $n = 123$; Figure 5.4D), confirming that cluster tendency was a genuine property of the data rather than an artefact of patient composition. An exploratory three-cluster solution yielded a silhouette of 0.220, inferior to the two-cluster solution, as expected when a third centroid is positioned within the continuum between two established poles rather than capturing an independent grouping. On the basis of converging quantitative evidence and clinical interpretability, the two-cluster solution ($n = 142$, 11 variables) was retained as the primary analysis.

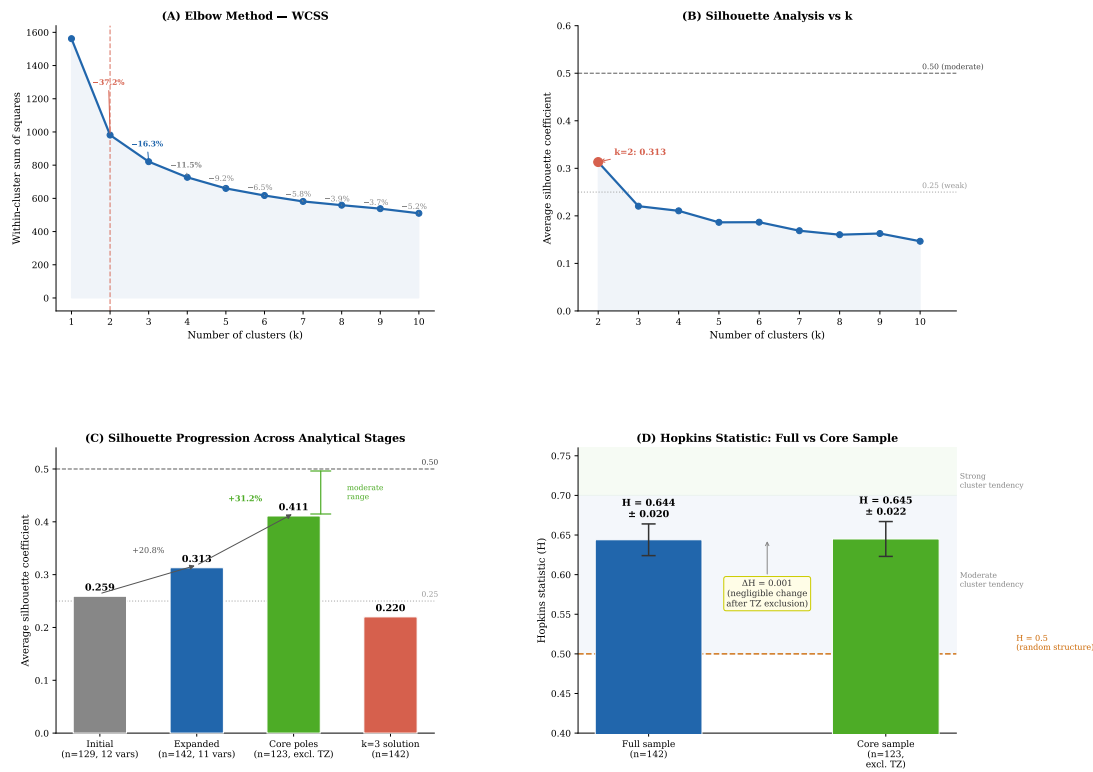


Figure 5.4: Cluster validation overview. (A) *Elbow method.* Within-cluster sum of squares (WCSS) for $k = 1$ to 10. The largest reduction (31.3%) occurs at $k = 2$, followed by progressively smaller decrements (13.7% at $k = 3$, 10.4% at $k = 4$), with no sharp inflection point. (B) *Silhouette analysis.* Average silhouette coefficient for $k = 2$ to 10. Maximum at $k = 2$ (0.313); dashed and dotted lines mark the 0.50 (moderate) and 0.25 (weak) Rousseeuw thresholds. (C) *Silhouette progression across analytical stages.* Initial solution ($n = 129$, 12 variables): 0.259. Expanded ($n = 142$, 11 variables): 0.313. Core poles ($n = 123$, excluding 19 transition-zone patients): 0.411 (+31.2%, moderate-to-strong range). Exploratory $k = 3$ solution: 0.220. (D) *Hopkins statistic stability.* Full sample ($n = 142$): $H = 0.644 \pm 0.020$; core sample ($n = 123$): $H = 0.645 \pm 0.022$ ($\Delta H = 0.001$).

Cluster Formation and Validation. Following the determination that $k = 2$ was the optimal number of clusters, K-means clustering was applied to the full cohort ($N = 142$) using the refined 11-variable biopsychosocial feature set, yielding an *Adaptive* cluster (Cluster 1, $n = 89$, 62.7%) and a *Maladaptive* cluster (Cluster 2, $n = 53$, 37.3%). Internal validation confirmed detectable though overlapping separation (silhouette = 0.313; Davies-Bouldin = 1.292), consistent with the heterogeneous nature of chronic pain populations. The modest aggregate silhouette reflects the presence of a transition zone rather than failed clustering: excluding the 19 transition-zone patients (silhouette < 0.25) raised the core-pole coefficient to 0.411, entering the moderate-to-strong range (Figure 5.4C). As illustrated in Figure 5.5, these 19 borderline patients occupy an intermediate position between the two phenotypic centroids, displaying profiles that are neither

fully Adaptive nor fully Maladaptive but rather a clinically meaningful transition zone.

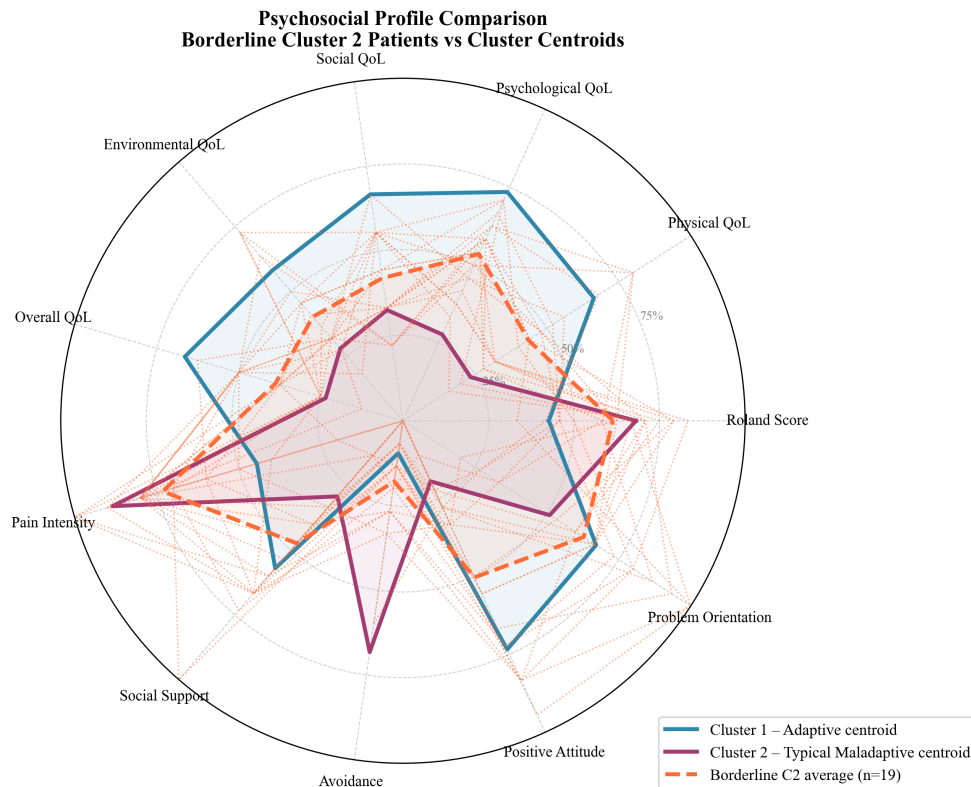


Figure 5.5: Psychosocial profile of borderline Cluster 2 patients relative to cluster centroids. Radar plot comparing normalized domain scores (0–1 scale) for borderline patients initially assigned to Cluster 2 ($n = 19$; dashed orange line, with individual profiles shown as faint dotted traces) against the two reference centroids: Cluster 1 (adaptive phenotype; solid blue) and Cluster 2 (typical maladaptive phenotype; solid purple). Across most quality-of-life and coping domains, the borderline subgroup lies between the two centroids and generally closer to the maladaptive profile, while showing partial proximity to the adaptive centroid in selected domains (e.g., Social Support). The plot highlights the intermediate/transition nature of this subgroup rather than a fully prototypical Cluster 2 pattern.

To assess robustness, the K-means solution was benchmarked against hierarchical clustering (Ward’s linkage) and three density-based algorithms (DBSCAN, OPTICS, HDBSCAN). Hierarchical clustering produced inferior separation (silhouette: 0.202; Davies-Bouldin: 1.566; ARI vs. K-means = 0.370). The density-based methods achieved higher core silhouette values (0.468–0.494) and independently recovered the same two phenotypic poles (Table 5.6), confirming that the two-cluster structure is not an artefact of K-means assumptions. However, all three relegated 28–36% of patients ($n = 40$ –51) to a noise category, effectively declining to classify them. As shown in Figure 5.5, these excluded patients substantially overlap with the 19 borderline individuals who occupy the intermediate transition

Table 5.6: Comparison of clustering solutions across four algorithms. Density-based methods (DBSCAN, OPTICS, HDBSCAN) achieve substantially higher core silhouette scores (0.468–0.494 vs. 0.313 for K-means) by permitting a noise region (28–36% of patients). All three independently recover the same two phenotypic poles. TZ-19 = the 19 transition-zone patients (silhouette < 0.18) from the K-means solution.

Method	Adaptive	Maladaptive	Noise (%)	Silhouette	Davies-Bouldin	ARI vs. K-means	TZ-19 Overlap
K-means	89	53	0 (0%)	0.313	1.292	—	—
DBSCAN	79	17	46 (32.4%)	0.468	0.714	0.471	7/19 (37%)
OPTICS	71	20	51 (35.9%)	0.494	0.713	0.591	10/19 (53%)
HDBSCAN	85	17	40 (28.2%)	0.477	0.701	0.517	8/19 (42%)

zone. Because the clinical objective requires phenotypic assignment for every patient, particularly those with ambiguous profiles who may benefit most from structured assessment, this exclusion behaviour was deemed a disqualifying limitation. K-means was therefore preferred: it recovers the same core structure while retaining full cohort coverage, including the clinically meaningful borderline subgroup visualized in Figure 5.5.

Bootstrap stability analysis (1,000 iterations) further confirmed the solution’s reproducibility (mean ARI = 0.843, $SD = 0.163$; Figure ??). In summary, the K-means two-cluster solution was selected as the final model on the basis of: (1) superior internal validity over hierarchical clustering; (2) cross-algorithm convergence on the same two phenotypes; (3) full cohort coverage, which density-based methods failed to provide; (4) high bootstrap stability; and (5) subsequent validation through large-effect-size differences on clinically meaningful outcomes.

Visualization and Variable Contribution. A PCA biplot (Figure 5.7) illustrated clear separation along the first principal component (PC1, 45.7% variance). Cluster 1 occupied the positive PC1 space, associated with higher quality of life (QoL) and adaptive coping vectors. Cluster 2 occupied the negative PC1 space, associated with higher pain, disability, and avoidance. Psychological QoL (loading = 0.899) and Avoidance (loading = -0.799) were the strongest contributors to cluster separation, confirming their discriminant power.

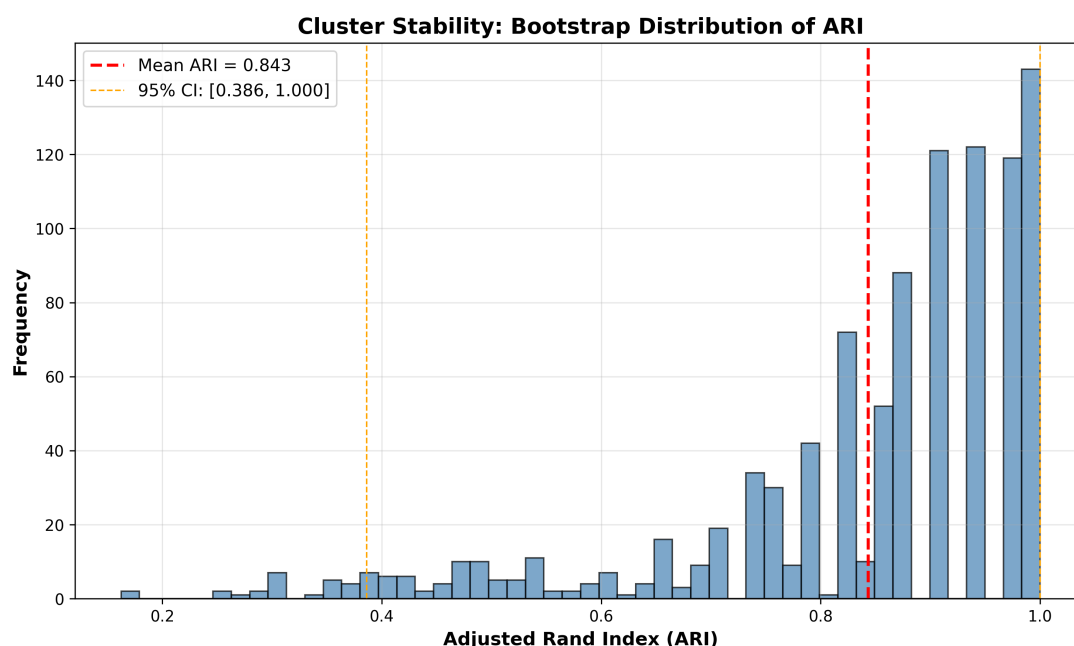


Figure 5.6: *Distribution of Adjusted Rand Index (ARI) values across 1000 bootstrap iterations. The mean ARI of 0.843 ($SD = 0.163$) indicates high cluster stability, demonstrating that the cluster solution is robust to sampling variation.*

5.3.2.2 Psychosocial and Quality-of-Life Differentiators

Comprehensive cluster comparisons revealed profound, large effect-size differences across all psychosocial and clinical domains (Tables 5.7).

Cluster 1: The Adaptive, Lower-Burden Phenotype. This group reported significantly lower pain intensity (NRS $M = 4.32$, $SD = 2.27$) and disability (Roland $M = 10.60$, $SD = 5.84$). They exhibited superior QoL across all domains, with the largest difference in Psychological QoL (Cohen’s $d = 2.12$, $p < 0.001$). Coping was characterized by high use of adaptive strategies, Positive Attitude ($d = 1.39$) and Social Support ($d = 0.82$), and minimal use of Avoidance ($d = -1.42$).

Cluster 2: The Maladaptive, High-Burden Phenotype. This group reported severe pain (NRS $M = 7.98$, $SD = 1.58$) and high disability (Roland $M = 14.96$, $SD = 4.57$). QoL was markedly diminished across all domains (all $d > 1.5$). Coping was dominated by high Avoidance ($d = -1.42$) coupled with low Positive Attitude and Social Support. Problem Orientation showed a small, significant difference ($d = 0.45$), while Transcendent coping did not differ.

To formally identify the key variables driving the separation between the two phenotypes, **Linear Discriminant Analysis (LDA)** was performed [197]. The analysis yielded a single discriminant function that explained 100% of the between-

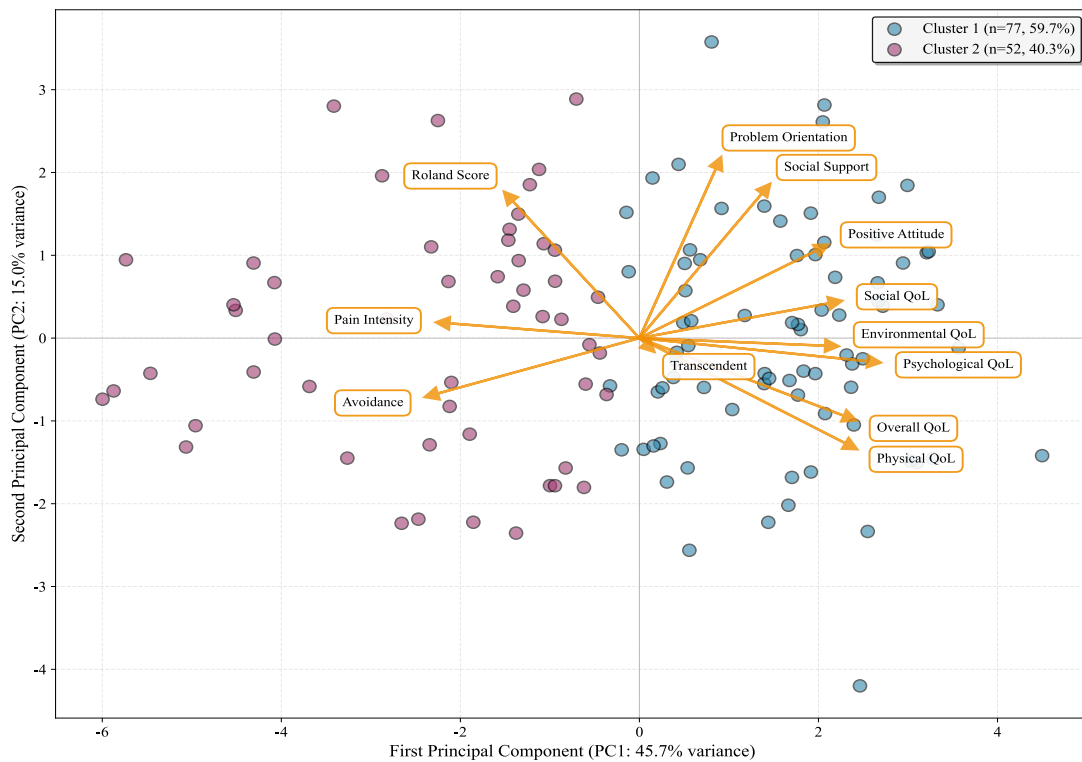


Figure 5.7: PCA biplot showing patient data points (colored by cluster assignment) and variable loading vectors (arrows) in the first two principal components (PC1: 45.7%, PC2: 15.0%). Cluster 1 (blue) is positioned primarily in the positive PC1 direction, while Cluster 2 (purple) is positioned primarily in the negative PC1 direction, showing clear separation. Variables pointing toward Cluster 1 (right) include quality of life domains and adaptive coping strategies, while variables pointing toward Cluster 2 (left) include pain intensity, functional disability, and avoidance coping. Arrow lengths indicate variable contribution strength, with psychological quality of life and avoidance coping showing the strongest contributions to cluster separation.

cluster variance, confirming a clear separation in the discriminant space. The variables with the strongest standardized discriminant coefficients (absolute value > 1.0) were identified as the primary discriminators: Overall QoL/General Health (coefficient = -2.29), Social QoL (coefficient = -1.98), Pain Intensity (coefficient = 1.59), and Positive Attitude (coefficient = -1.03). This result quantifies that the cluster separation is predominantly defined by a combined profile of impaired quality of life, higher pain intensity, and lower use of adaptive coping strategies, thereby providing statistical validation for the clinical interpretation of the derived phenotypes.

5.3.2.3 Network-Level Characteristics

Having established that clusters differ in their mean levels of clinical and psychosocial variables, we next investigated whether the *relationships* between variables differ between clusters. This network analysis addresses a critical question: *do the*

Table 5.7: Descriptive statistics and between-cluster comparisons.

Variable	C1 μ (SD)	C2 μ (SD)	U	p	d	ES
Pain Intensity	4.32 (2.27)	7.98 (1.58)	18.47	<0.001	-1.81	Large
Roland Disability	10.60 (5.84)	14.96 (4.57)	16.32	<0.001	-0.81	Large
Physical QoL	59.93 (14.70)	30.77 (18.17)	17.03	<0.001	1.80	Large
Psychological QoL	74.19 (12.84)	41.59 (18.55)	18.42	<0.001	2.12	Large
Social QoL	75.97 (10.98)	54.33 (11.73)	17.66	<0.001	1.92	Large
Environmental QoL	69.64 (11.38)	51.02 (12.19)	15.84	<0.001	1.59	Large
Overall QoL	67.05 (16.83)	32.45 (16.65)	18.21	<0.001	2.06	Large
Social Support	2.76 (0.68)	2.18 (0.75)	14.68	<0.001	0.81	Large
Avoidance	1.27 (0.38)	2.26 (0.99)	13.21	<0.001	-1.42	Large
Positive Attitude	3.23 (0.83)	2.07 (0.84)	14.32	<0.001	1.39	Large
Problem Orientation	3.06 (0.77)	2.71 (0.82)	5.91	0.015	0.45	Small
Transcendent	1.63 (0.74)	1.65 (0.88)	0.16	0.691	-0.03	Neg

Note: C1 = Cluster 1, Adaptive ($n = 89$), C2 = Cluster 2, Maladaptive ($n = 53$); μ = mean, SD = standard deviation; U = Mann-Whitney U statistic; d = Cohen's d ; ES = Effect sizes: Large ($|d| \geq 0.8$), Medium ($0.5 \leq |d| < 0.8$), Small ($0.2 \leq |d| < 0.5$), Neg - Negligible ($|d| < 0.2$); QoL = Quality of Life. All p -values two-tailed.

same variables relate to each other differently in each cluster, suggesting different underlying mechanisms?

To model and compare the systems of relationships within each phenotype, we constructed separate undirected network graphs for Clusters 1 and 2. In these graphs, each **node** represents one of the 12 psychosocial variables (e.g., Pain Intensity, Psychological QoL). An **edge** between two nodes was created if their Spearman rank correlation coefficient was robust ($|r| \geq 0.3$), representing a monotonic association that is practically significant. This threshold-based approach focuses on the strongest relationships that likely shape each cluster's internal dynamics. The resulting correlation coefficients are visually summarized in Figure 5.9.

Network Density and Structure: The resulting networks were architecturally opposed. **Cluster 2** (the higher-burden phenotype) exhibited a dense, highly interconnected network with a density of 0.394 and an average clustering coefficient of 0.700, comprising 26 edges. In stark contrast, **Cluster 1**'s network was sparse, with a density of 0.197, a clustering coefficient of 0.117, and only 13 edges, exactly half the density of Cluster 2 (Figure 5.8). This structural dichotomy suggests that in the high-burden phenotype, variables operate within a tightly coupled, potentially reinforcing system, whereas in the adaptive phenotype, these factors are more independent and compartmentalized.

Hub Variables and Mechanistic Differences: Centrality analysis, which identifies the most connected nodes, revealed different operational hubs in each

network (Figure 5.10). In Cluster 1, Psychological QoL emerged as the central hub (degree centrality = 0.545), suggesting psychological well-being is the integrative node in the adaptive pattern. Conversely, in Cluster 2, Avoidance Coping was the dominant hub (degree centrality = 0.727), connecting to 8 other variables and positioning maladaptive coping as the central driver of the high-burden state. This is exemplified within Cluster 2’s dense network by a tightly knit subgroup where strong edges connect Avoidance Coping, Psychological QoL, and Pain Intensity, suggesting a potential “vicious cycle” that may maintain symptom burden e.g., Pain → lower QoL → more Avoidance → less Positive Attitude → worse Pain. The sparse network in Cluster 1 suggests more independent, resilient functioning.

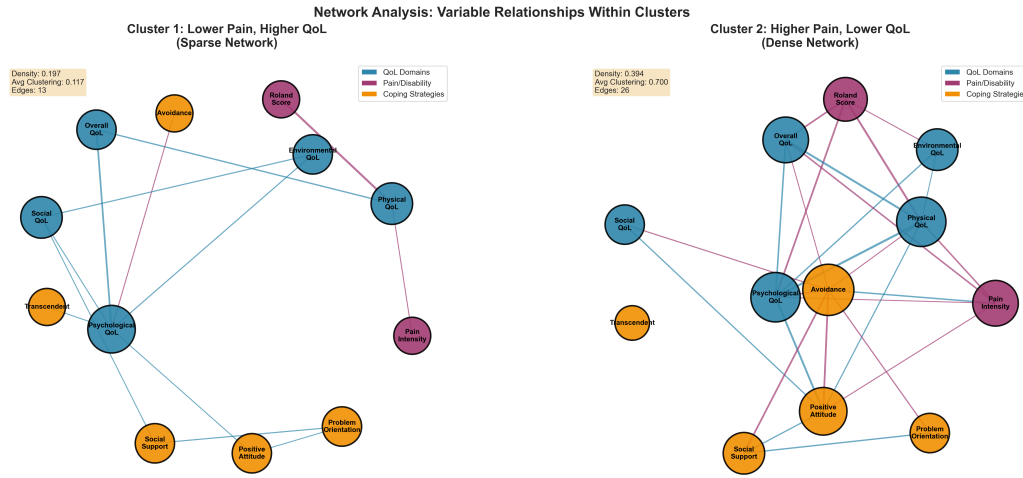


Figure 5.8: Network graphs for each cluster showing variable relationships (edges represent $|r| \geq 0.3$). Cluster 1 (left) exhibits a sparse network (density = 0.197) with Psychological QoL as the central hub. Cluster 2 (right) exhibits a dense network (density = 0.394) with Avoidance Coping as the central hub, indicating a complex maladaptive cycle.

5.3.2.4 Clinical–Radiological Decoupling

Next, to empirically evaluate a core premise of the biopsychosocial model, that clinical burden may be dissociable from structural pathology, we undertook a multilevel analysis of spinal imaging data. The objective was to determine whether the markedly worse clinical presentation of the high-burden psychosocial phenotype was accompanied by correspondingly greater or more localized structural abnormalities on conventional MRI.

First, from 230 disc-level observations, we derived patient-level indices reflecting complementary dimensions of pathology: *Global Burden* (overall extent of degeneration), *Focal Severity* (maximum disc-level abnormality), and *Spatial Pattern* (distribution across spinal regions). Comparisons of these theory-driven composite

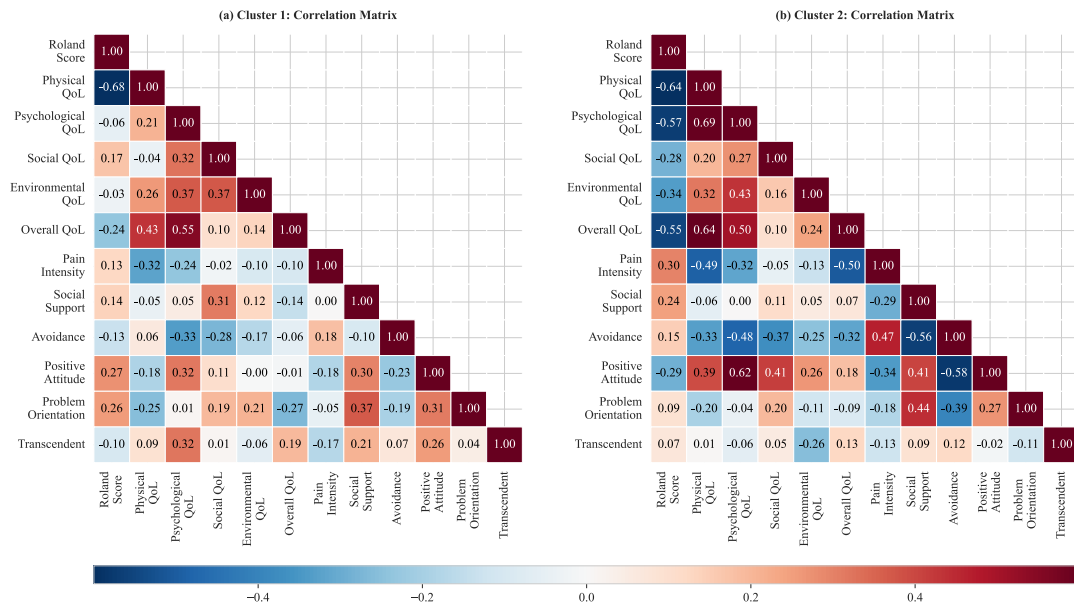


Figure 5.9: Correlation heatmaps for Cluster 1 (left) and Cluster 2 (right) showing Spearman's rank correlations between all clustering variables. The heatmaps illustrate the different correlation patterns between clusters, with Cluster 2 showing stronger and more numerous correlations.

measures revealed **no significant differences** between the Adaptive (Cluster 1) and Maladaptive (Cluster 2) psychosocial phenotypes (all $p > 0.05$).

To address multicollinearity among imaging variables and identify latent, data-driven patterns without a priori assumptions, we performed a Principal Component Analysis (PCA). The first two principal components, which together explained 60.2% of the total variance, represented “Overall Structural Burden” (PC1) and “Disc Degeneration Severity” (PC2). Scores on these data-driven dimensions also failed to separate the two clinical clusters.

We further tested the specific hypothesis that pathology localized to symptom-relevant regions (lumbar, cervical) might still differentiate the phenotypes. As with the global measures, region-specific indices of burden and severity showed no statistically detectable associations with cluster membership (all $p > 0.29$).

The comprehensive results of these nine distinct comparisons are presented in Table 5.8.

Taken together, these analyses provide consistent, multi-method evidence for a statistical dissociation between psychosocial-defined clinical burden and the spectrum of pathology observable on conventional spinal MRI. This finding underscores a central clinical reality: the severity of a patient's pain and disability cannot be reliably inferred from imaging findings alone. It directly challenges a

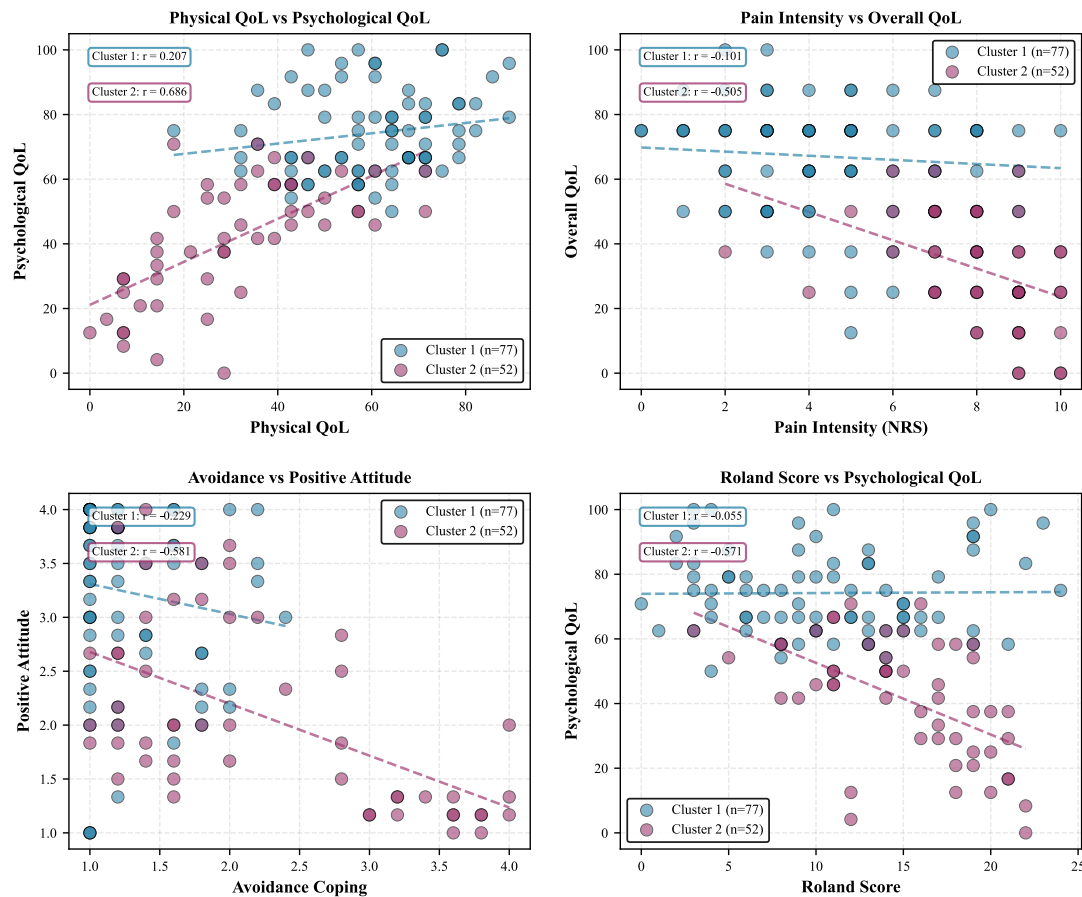


Figure 5.10: Scatter plots of key variable relationships comparing Cluster 1 (blue) and Cluster 2 (purple). The plots illustrate how relationships differ between clusters, with Cluster 2 showing stronger correlations (e.g., Physical QoL $\leftrightarrow$ Psychological QoL, Roland Score $\leftrightarrow$ Psychological QoL).

purely pathoanatomic diagnostic approach and substantiates the necessity of the integrated biopsychosocial assessment framework proposed in this work.

5.3.2.5 Predictive Modelling of Cluster Membership

Given the clear cluster differences and high stability, we next developed predictive models to enable early identification of cluster membership. The clinical utility of the full 12-variable clustering solution was limited by its significant assessment burden, requiring patients to complete a total of 76 items: WHOQoL-BREF (26 items, approximately 10-15 minutes), Cope-NVI (25 items, approximately 5-10 minutes), Roland-Morris Disability Questionnaire (24 items, approximately 5 minutes), and Pain NRS (1 item, few seconds). The total time to complete all four questionnaires would be roughly 20-30 minutes, which may be impractical in busy clinical settings.

To enhance feasibility, we engineered a simplified predictor using a subset of

Table 5.8: Comparison of Structural Pathology by Psychosocial Cluster

Analysis	Metric	Clus. 1 M (SD)	Clus. 2 M (SD)	U	<i>p</i>
Composite Indices					
Overall Burden	Global Burden Index	0.08 (0.92)	-0.12 (0.57)	2125.5	0.553
Focal Severity	Focal Severity Index	-0.01 (0.55)	0.01 (0.47)	1872.0	0.533
Spatial Pattern	Spatial Pattern Index	0.04 (0.87)	-0.06 (0.74)	2068.5	0.750
Principal Comp.					
Data-Driven	PC1: Overall Burden	0.15 (2.57)	-0.22 (1.74)	2028.5	0.901
	PC2: Degeneration Sev.	0.01 (1.94)	-0.01 (1.66)	2009.5	0.973
Regional Analysis					
<i>Lumbar Region</i>					
Burden	Global Burden Index	0.08 (0.81)	-0.11 (0.65)	2218.5	0.299
Severity	Focal Severity Index	0.01 (0.59)	-0.02 (0.48)	2007.0	0.982
<i>Cervical Region</i>					
Burden	Global Burden Index	-0.03 (0.76)	0.05 (0.87)	1930.5	0.588
Severity	Focal Severity Index	-0.02 (0.52)	0.02 (0.59)	1922.5	0.538

Note: Clus. 1: Adaptive, Lower-Burden Phenotype (n=77); Clus. 2: Maladaptive, High-Burden Phenotype (n=52). All tests: Mann-Whitney *U*. Values are standardized scores (*mean* $\approx$ 0, *SD* $\approx$ 1). All *p* $\geq$ 0.05.

four variables selected through a data-driven approach combining statistical importance, clinical relevance, and practical feasibility. Variable selection was based on three criteria: (1) statistical discriminative power, assessed through linear discriminant analysis coefficients and effect sizes (Cohen's *d*) from cluster comparisons; (2) clinical importance, prioritizing core pain and functional measures and key psychosocial differentiators; and (3) clinical feasibility, favoring variables that can be collected quickly in routine practice.

The selected variables were: Pain NRS (1 item), Roland-Morris Disability Questionnaire (24 items), the Psychological domain of WHOQoL (6 items), and the Avoidance subscale of Cope-NVI (5 items). Pain NRS was selected as a top discriminator (LDA coefficient = 1.59, rank #3) with a large effect size (Cohen's *d* = -1.81, rank #4). WHOQoL-Psychological was selected as it showed the largest effect size (Cohen's *d* = 2.12, rank #1) and represents the strongest differentiator between clusters. Avoidance coping was selected for its large effect size (Cohen's *d* = -1.42, rank #7) and clinical importance as a maladaptive coping pattern. Roland Score was included as a core functional disability measure (Cohen's *d* = -0.81, rank #10) that is standard in clinical practice and captures a key outcome domain.

This brings the total to 36 items, a 53% reduction in respondent burden. The simplified assessment condenses the original 20-30 minute comprehensive assessment

into an 8-10 minute assessment (Psychological QoL: 2-3.5 minutes, Avoidance Coping: 1-2 minutes, Roland-Morris: 5 minutes, Pain NRS: few seconds). This more than 50% reduction in administration time makes routine screening feasible in clinical practice.

5.3.2.6 Simplified Predictive Model

Using these four selected variables, we developed predictive models to identify cluster membership. We compared two algorithms: logistic regression and random forest. Both models were trained on 80% of the data ($n = 103$) and tested on the remaining 20% ($n = 26$), with stratification to maintain cluster proportions. Logistic regression achieved a test accuracy of 88.5% with strong discrimination ($\text{AUC-ROC} = 0.950$), while random forest achieved a test accuracy of 92.3% with an AUC-ROC of 0.912. Cross-validation across 5 folds revealed more consistent performance for logistic regression (mean accuracy = 91.5%, $\text{SD} = 4.5\%$) compared to random forest (mean accuracy = 86.9%, $\text{SD} = 8.6\%$). Logistic regression was selected as the final model based on its superior AUC-ROC (0.950 vs. 0.912), which is the preferred metric for binary classification, and its more stable cross-validation performance. Additional performance metrics for the selected logistic regression model included precision = 100.0%, recall = 70.0%, and F1-score = 82.4%.

Clinical Translation (Decision Rules and Risk Scoring): While the logistic regression model provides high accuracy, its clinical utility in fast-paced environments may be limited because it outputs a probability rather than an immediate, actionable decision. To create a tool that could be used at a glance without calculation, we developed simple, interpretable decision rules using a decision tree classifier. The decision tree's complexity was constrained by setting a maximum depth of 3 to prioritize clinical interpretability while maintaining reasonable accuracy.

The pruned decision tree achieved a test accuracy of 88.5%, identical to the logistic regression model, while offering complete transparency. Other performance metrics included a precision of 88.9% and a recall of 80.0%. The tree identified Pain Intensity (NRS) as the root (primary) decision node, followed by Psychological Quality of Life and Avoidance Coping as secondary splits, creating a simple, three-question checklist for risk stratification. The confidence values reported (e.g., 85%, 90%) represent the class probabilities from the leaf nodes, indicating the proportion of training samples in that leaf that belonged to the predicted cluster. For example, a leaf node with 85% confidence for Cluster 2 indicates

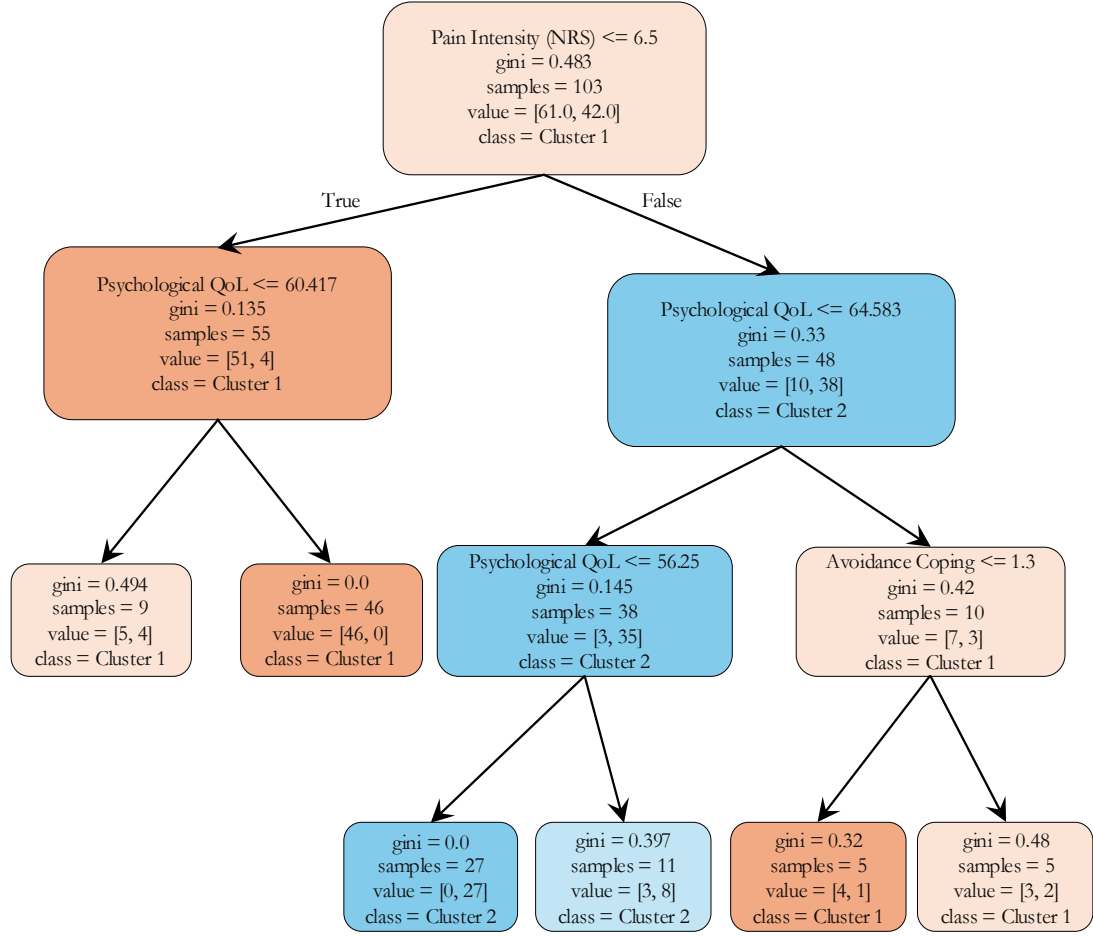


Figure 5.11: Decision tree for cluster classification using the 4 simplified variables. The tree has a maximum depth of 3 levels, with Pain Intensity (NRS) as the root (primary) decision node, followed by Psychological Quality of Life and Avoidance Coping as secondary splits. The visualization depicts a hierarchical classification model for 103 samples, primarily split by Pain Intensity (NRS) at a threshold of 6.5. The left branch (orange) represents the path for lower pain levels, further partitioned by Psychological Quality of Life (QoL), leading predominantly to Cluster 1. The right branch (blue) represents higher pain levels, where classification into Cluster 2 is driven by lower Psychological QoL scores (≤ 56.25). A specific subgroup with high pain but high resilience, characterized by high QoL and low Avoidance Coping, is redirected back into Cluster 1. The tree does not terminate at the high-pain node because it is impure, containing both Cluster 2 patients and this resilient Cluster 1 subgroup that requires additional splits for correct classification. Each node displays the Gini impurity index, sample size, class distribution, and the resulting classification. These simple, interpretable rules achieve 88.5% accuracy, matching the logistic regression model while offering complete transparency for clinical application.

that 85% of the training samples reaching that leaf were assigned to Cluster 2. Figure 5.11 displays the decision tree visualization, showing the simple, clinically interpretable rules that can be applied without computational resources.

While the above decision tree provides binary cluster classification, it lacks a risk

spectrum that could guide treatment intensity. To address this limitation and provide a continuous risk assessment tool, we developed a risk scoring system based on the same four variables used in the simplified predictive model: Pain Intensity (NRS), Roland-Morris Disability score, Psychological Quality of Life (WHOQoL-Psychological), and Avoidance Coping. The scoring system uses the logistic regression coefficients to weight each variable's contribution, with risk scores calculated as the probability of Cluster 2 membership transformed into a 0-200 point scale for clinical interpretability.

The risk scores were categorized into three levels: Low Risk (scores < 50), Moderate Risk (scores 50-150), and High Risk (scores ≥ 150). A total score ≥ 150 was found to predict Cluster 2 membership with 100% specificity and 50% sensitivity. A lower threshold (score ≥ 37) optimized for screening achieved a balance of 90.0% sensitivity and 87.5% specificity, effectively identifying 90% of high-risk patients while maintaining a low false-positive rate. When applied to the test set, the scoring system achieved an accuracy of 80.8% and an AUC of 0.950, demonstrating strong discriminative performance comparable to the logistic regression model (accuracy: 88.5%, AUC: 0.950) while providing a continuous risk spectrum that can guide clinical decision-making regarding treatment intensity and resource allocation.

To facilitate clinical adoption, we developed an interactive web-based interface (Figure 5.12) that implements the logistic regression model with SHAP explanations. The interface requires four clinically feasible inputs: (1) Pain Intensity (NRS, 0–10 single item); (2) Roland-Morris Disability score (24-item checklist, 0–24); (3) Psychological Quality of Life (WHOQoL-BREF, 6 items, transformed to 0–100); and (4) Avoidance Coping (COPE-NVI, 5 items, mean score 1.0–4.0). Together, these four assessments require approximately 5–8 minutes to complete at the point of care.

Once the patient completes the four questionnaires, the scores are automatically transferred to the classifier, which predicts the cluster assignment (Adaptive or Maladaptive), continuous risk score (0-200 scale), and risk category (Low/Moderate/High), alongside SHAP feature importance values quantifying each variable's contribution to the prediction. This explainable AI approach provides transparency into why a particular prediction was made, facilitating clinical trust and adoption. The tool also displays a cluster profile comparison and generates clinical recommendations with priority level and referral pathways, as illustrated in Figure 5.12.

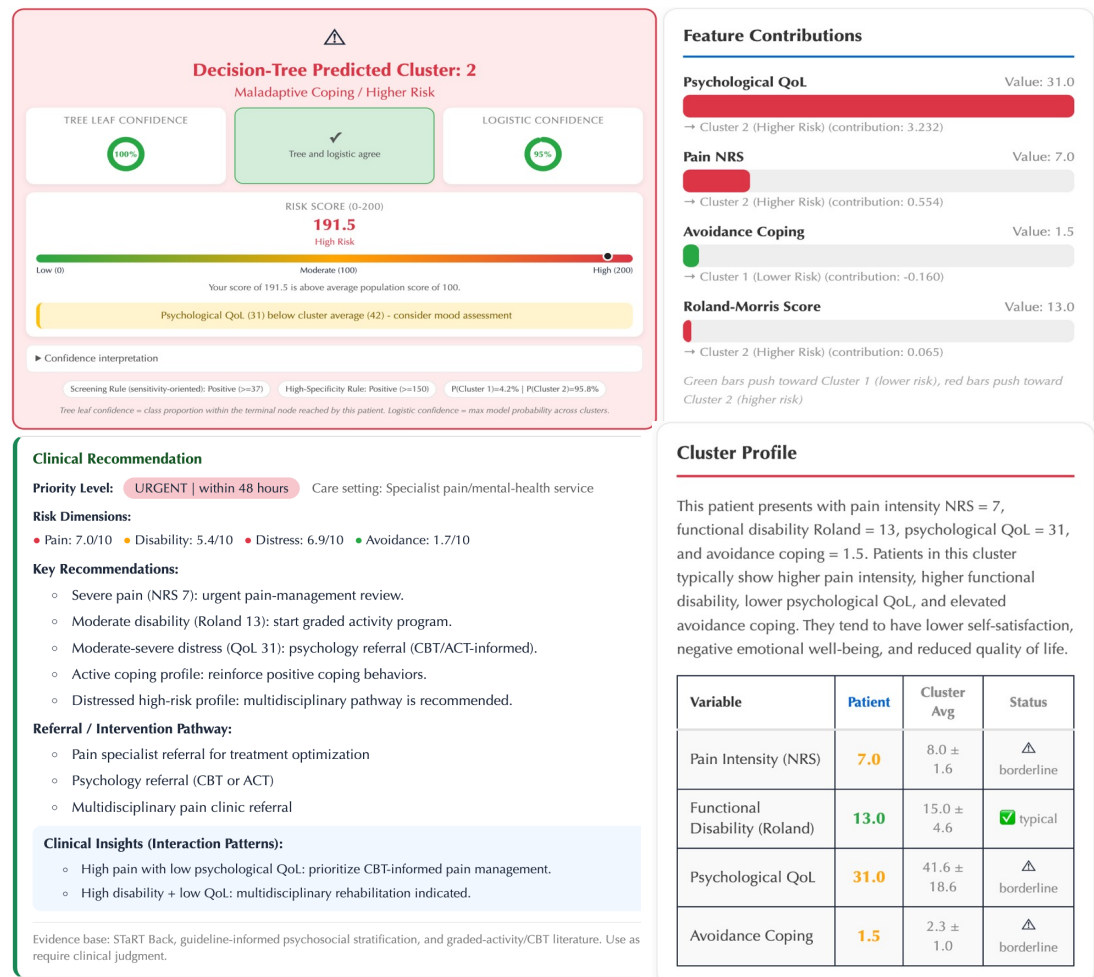


Figure 5.12: Interactive clinical screening report for a Maladaptive phenotype patient. The interface integrates prediction summary, SHAP feature contributions, cluster profile comparison, and adaptive clinical recommendations. Interactive clinical screening report generated for a patient classified as **Cluster 2 (Maladaptive phenotype)**. The interface integrates four modules: **(Top-left)** Prediction summary showing decision tree and logistic regression confidence, risk score (0-200 scale) with population average reference, and screening threshold indicators. **(Top-right)** SHAP feature contribution plot displaying the direction and magnitude of each variable's influence on the prediction (red = toward Cluster 2/higher risk; green = toward Cluster 1/lower risk). **(Bottom-right)** Cluster profile comparison showing patient values against cluster averages with borderline/typical status indicators. **(Bottom-left)** Adaptive clinical recommendation module providing priority level, care setting, risk dimensions, key recommendations, referral pathways, and clinical insights based on interaction patterns.

Together, these tools enable rapid (< 10 minutes), accurate ($AUC \approx 0.95$) identification of high-risk biopsychosocial phenotypes in clinical settings. The $> 50\%$ reduction in assessment time facilitates the practical implementation of tailored, phenotype-specific treatment pathways, moving from a one-size-fits-all approach to personalized spine care.

5.4 Discussion

5.4.1 Integrated Interpretation and Clinical Implications

This study employed a dual analytical strategy, first variable-centered, then person-centered, to elucidate the biopsychosocial architecture of spinal pain. The findings consistently converge on a model where Avoidance and Positive Attitude are the primary, independent determinants of pain intensity. This aligns with contemporary frameworks that distinguish these as core maladaptive and adaptive pathways, respectively. Avoidance, a key component of the fear-avoidance model, has been robustly confirmed as a driver of both pain intensity and disability in diverse populations as well [198]. Conversely, our identification of Positive Attitude as a primary protective factor resonates with the growing field of positive psychology, which conceptualizes adaptive coping, positive reframing, and finding meaning as cornerstones for thriving with chronic pain [199].

The non-significant direct effects of strategies like Social Support and Problem Orientation, coupled with their strong indirect effects (mediating 77-113% of relationships), reveal a functional hierarchy. These foundational strategies exert influence predominantly by shaping the proximal determinants, enhancing Positive Attitude and reducing Avoidance, rather than through direct effects on pain. This pattern clarifies why bivariate associations can vanish in multivariate models and underscores the importance of examining indirect causal networks [198, 200].

When this coping hierarchy is embedded within the full biopsychosocial context, pain, disability, and quality of life (QoL) core domains, patients naturally segregate into two distinct, stable clusters. Cluster 1 (Adaptive, Lower-Burden Phenotype) exhibits lower pain, higher QoL, and a coping profile anchored by Positive Attitude. Its sparse network structure suggests that these adaptive resources operate in a relatively decoupled, resilient manner, a pattern aligned with the concept of “thriving” in chronic pain [199]. In contrast, Cluster 2 (Maladaptive, High-Burden Phenotype) is defined by severe pain, catastrophically low QoL, and a coping profile dominated by Avoidance. Critically, its network is dense and interconnected, with Avoidance as the central hub. This topology suggests a vicious

cycle consistent with established fear-avoidance and mutual maintenance models, where pain, disability, and maladaptive cognition reinforce one another [198, 201]. The successful identification of such distinct subgroups using data-driven methods aligns with the growing use of machine learning for pain phenotyping, which aims to predict treatment response and tailor management [202].

These data-driven phenotypes reveal that the most critical determinants of patient distress are psychosocial, a conclusion powerfully reinforced by the complete clinical–radiological decoupling observed in this cohort. This finding provides strong empirical support for a core principle of the biopsychosocial model: that the illness experience may be substantially dissociated from structural disease [30]. In our cohort, patients in the maladaptive, high-burden cluster exhibited severe pain and disability that appeared independent of the level or pattern of pathology visible on conventional MRI. This dissociation suggests that for a distinct clinical phenotype, functional burden may not be reliably inferred from imaging findings alone.

This decoupling carries immediate diagnostic and therapeutic implications. It challenges the clinical heuristic that seeks the cause of disability primarily in imaging findings. Instead, it underscores that the clinical state of the maladaptive cluster is a product of functional and perceptual processes, specifically, the dense network of fear-avoidance and maladaptive cognition that characterizes this phenotype, rather than structural severity. Consequently, effective management must pivot decisively away from a pathoanatomic diagnostic model. Treatment should target the specific psychosocial mechanisms defining this subgroup, such as using cognitive-behavioral approaches to disrupt avoidance cycles and fostering psychological flexibility to build resilience [203]. Consequently, for the maladaptive, high-burden phenotype identified in this study, effective management requires moving beyond a pathoanatomic diagnosis. This aligns with modern frameworks for chronic pain, which advocate a fundamental shift in treatment goals from seeking a structural cause to improving functional capacity and self-management [204, 203, 205].

Clinical Translation: From Stratification to Action

The translation of these phenotypic insights into a practical screening tool addresses a major barrier to biopsychosocial implementation: assessment burden. By reducing the full 12-variable battery to four core predictors (Pain Intensity, Disability, Psychological QoL, Avoidance Coping), the tool enables rapid, point-of-care stratification without requiring specialised scoring software or lengthy patient

questionnaires. This approach is consistent with contemporary calls for integrated, co-managed care that addresses both pathoanatomic and psychosocial factors simultaneously [203].

The derived decision rules and risk score directly inform stratified, mechanism-targeted care pathways:

- **For Cluster 1 (Adaptive Copers):** Care should focus on *maintenance and prevention*. Reinforce adaptive coping via positive psychology interventions that foster engagement, relationships, and meaning [199], and support self-management. Standard rehabilitation is likely adequate.
- **For Cluster 2 (Maladaptive Copers):** This group requires *intensive, multimodal intervention* to disrupt its dense maladaptive network. Treatment must concurrently:
 1. **Target the Central Hub (Avoidance):** Employ graded exposure and Cognitive-Behavioral Therapy (CBT) to break the fear-avoidance cycle [198].
 2. **Build Adaptive Resources:** Use Acceptance and Commitment Therapy (ACT) to foster psychological flexibility, pain acceptance, and values-based action [206, 207].
 3. **Leverage Foundational Strategies:** Actively facilitate social support and problem-solving skills to indirectly bolster primary adaptive targets.

The network density observed in the Maladaptive phenotype (Figure 5.8) implies that isolated interventions may be insufficient; an integrated, co-managed approach targeting multiple nodes simultaneously is warranted [203].

5.4.2 Limitations

Several methodological boundaries shaped what can be concluded from this study and, perhaps more importantly, what remains to be explored.

First, the cross-sectional design precludes causal inference and tracking of phenotypic stability. Future longitudinal studies would benefit from testing whether these phenotypes predict trajectories of chronicity or differential treatment response, a logical next step supported by phenotyping research [202]. Second, the single-center cohort may limit generalizability; external validation in diverse cultural and healthcare settings is essential. Third, while self-report is appropriate for subjective constructs, future work could integrate objective behavioral or physiological measures. Finally, although radiological decoupling was clear, other unmeasured neurophysiological or inflammatory factors may yet play a role.

Separately, the predictive models, while promising, require prospective validation in independent cohorts and integration into clinical decision support systems to evaluate their real-world impact on treatment selection and patient outcomes.

5.4.3 Methodological Considerations and Strengths

The methodological framework of this study was designed to provide a comprehensive, multi-layered understanding of patient heterogeneity in spinal pain by employing sequential, complementary analytical strategies, each addressing a distinct research question. First, mediation analysis was used to move beyond simple association and elucidate the functional hierarchy and potential causal pathways within the coping system [196]. This approach allowed us to distinguish foundational strategies (e.g., Problem Orientation) with indirect effects from proximal determinants (Avoidance, Positive Attitude) with direct effects on pain, thereby clarifying the mechanisms through which coping influences outcomes. Second, to transition from these variable-centered mechanisms to person-centered profiles, we employed unsupervised machine learning (k-means clustering). This data-driven technique avoided imposing a priori categorical assumptions, allowing distinct biopsychosocial phenotypes to emerge directly from the patterns in the data [197]. The robustness of the two-cluster solution was supported by the convergence of multiple validation metrics (elbow method, silhouette analysis, bootstrap stability) and demonstrated high reproducibility.

Third, building on the identification of these phenotypes, network analysis was applied to move beyond comparing mean levels of variables. This analysis revealed fundamental differences in the systemic interconnectivity of clinical and psychosocial factors between clusters [208]. The discovery of a dense, maladaptive network in the high-burden phenotype, with Avoidance as a central hub, provides a novel, mechanistic lens that suggests therapeutic interventions may need to target interconnected system dynamics rather than isolated symptoms.

Finally, to bridge the gap between statistical identification and clinical application, we developed a simplified predictive tool using a principled, tri-criteria approach that balanced statistical discriminative power, clinical relevance, and practical feasibility. This process ensures the tool has genuine translational relevance for point-of-care risk stratification.

5.5 Conclusion

This study demonstrates the critical importance of psychosocial data in understanding and stratifying patients with spinal pain. Through a multi-method analysis of patient-reported outcome measures (PROMs), we identified a robust, two-cluster phenotypic solution. The "Adaptive, Lower-Burden" phenotype, characterized by positive cognitive coping and resilient network dynamics, contrasts sharply with the "Maladaptive, High-Burden" phenotype, defined by avoidance coping, dense psychological interconnectivity, and severe impairment. Critically, these distinct clinical profiles were completely dissociated from the severity or pattern of structural pathology on MRI, providing empirical validation for a biopsychosocial model of care.

These findings offer more than a clinical typology; they provide the validated data architecture for the PROM module within the proposed AI ecosystem for spine care. The specific measures that successfully defined these phenotypes, pain intensity (NRS), disability (Roland-Morris), quality of life (WHOQoL), and coping strategies (COPE-NVI), are precisely the multimodal inputs required for the PROM module of the framework. Furthermore, the predictive model derived from this analysis, which prioritizes Psychological QoL, Pain Intensity, Avoidance Coping, and Functional Disability, exemplifies the translation of complex PROM data into a clinically actionable tool for early risk stratification.

Together, this work operationalizes the key psychosocial dimensions that, when integrated with automated imaging analysis and large language model synthesis, can enable a truly personalized, mechanism-informed approach to spine care. This shifts the focus from isolated pathology to comprehensive patient function.

6

Proof-of-Concept: Feasibility and Interpretability of the Integrated Workflow

6.1 Introduction

This chapter presents an empirical proof-of-concept pilot study to evaluate the clinical feasibility and interpretability of a workflow that integrates patient-reported outcome measures (PROMs) with automated spine MRI analysis. While Chapters 4 and 5 validated the core analytical components, LumbarAI for imaging and clustering for psychosocial phenotyping, this study assesses the *clinical sense-making potential* of their combined outputs in a simulated clinical setting.

The central aim is to demonstrate whether a unified view of structural and psychosocial data, when presented coherently, can: (a) be integrated into a radiological workflow without disruption, (b) influence specialist reasoning toward more personalized treatment prioritization, and (c) be perceived as clinically plausible and useful. This pilot serves as a critical, real-world **feasibility and sense-making step**, probing the core assumptions of the conceptual framework presented in Chapter 3 prior to any investment in full automation.

Clarification of Scope and Claims Importantly, this proof-of-concept is not intended to constitute a full system integration or a clinical validation study. Rather, it represents a manual, researcher-mediated integration designed to test feasibility, interpretability, and clinical sense-making prior to automation. The synthesis of imaging outputs and PROMs was performed using predefined rule-

based mappings (severity thresholds, lookup tables, and conditional logic) rather than end-to-end AI orchestration. Full system-level integration would require substantial multidisciplinary development, including clinical workflow engineering, regulatory alignment, and prospective deployment, which are beyond the scope of this dissertation.

6.2 Methods for the Integrated Pilot Study

6.2.1 Study Design and Participants

A qualitative pilot case series was conducted to obtain preliminary evidence on the usability and impact of a multimodal workflow. The study involved the following elements:

- **Patients:** Three patients were purposively sampled from the dataset presented in Chapter 5, to represent the spectrum of different psychosocial phenotypes: one *Adaptive Coper* (Patient A), one *Maladaptive Coper* (Patient C), and one patient with an intermediate profile (Patient B).
- **Clinicians:** Two experienced musculoskeletal radiologists participated. Both were familiar with routine spine MRI reporting but were naive to the specific AI tools and PROMs profiles used in this study.

6.2.2 Study Procedures

The evaluation followed a structured, three-phase protocol for each patient case:

1. **Baseline MRI-Only Reporting:** Radiologists were shown spine MRI images without additional context and asked to generate a standard report using voice transcription. Only active dictation time was recorded.
2. **AI-Assisted Reporting:** Radiologists were provided with a draft report generated by the LumbarAI model, complete with explainable AI (XAI) overlays (heatmaps and annotations) as illustrated in Appendix A. They were asked to review, edit, and refine this draft. The time for this task was recorded.
3. **PROMs-Informed Clinical Planning:** Radiologists were given a summary of the patient's key PROMs (NRS pain, Roland-Moris disability score, WHOQoL subdomains scores, Cope-NVI profile, and phenotype cluster). They were explicitly asked if this information altered their treatment recommendations from Phase 1. Any changes were documented.

Finally, unstructured debriefing interviews were conducted. Feedback was analyzed using thematic analysis to identify key insights about workflow integration and

multimodal data value.

6.2.3 Nature of the Manual Integration

The integration demonstrated in this pilot was deliberately manual and modular, a standard methodological approach in translational AI research for isolating conceptual value prior to automation. The process involved two distinct components. First, LumbarAI generated outputs in the form of pathology-specific severity scores and structured labels, as visualized in Appendix A. Second, these discrete outputs were manually mapped to standardized descriptive statements using a predefined, transparent rule-based logic (numeric grade-to-severity translation via conditional statements). No large language model or autonomous narrative generator was used. This controlled, non-generative approach was chosen to isolate the conceptual value of multimodal synthesis from confounding effects related to generative model behavior. Consequently, the pilot produced a hybrid clinical draft report that combined AI-generated scores with manually synthesized prose (see Appendix A). This design allowed radiologists to evaluate the quality and utility of the integrated clinical content itself, rather than the fluency of automated text generation.

6.3 Summary of Validated Core Components

This proof-of-concept builds upon two independently validated modules whose performance is a prerequisite for any meaningful integration:

- **Module 1: LumbarAI.** The imaging model (Chapter 4) provides accurate, multi-pathology detection. Its key output for integration is a structured radiology report with explainable visualizations.
- **Module 2: Psychosocial Phenotyping.** The clustering model (Chapter 5) robustly identifies two patient phenotypes. Its key output is a patient’s cluster assignment and profile of core determinants.

6.4 Results

6.4.1 Integrated Patient Profiles

The three patients represented a clear gradient of biopsychosocial burden (Table 6.1). The manual synthesis for the Maladaptive Coper (Patient C), for example, combined the LumbarAI finding of “severe L4-L5 central stenosis” with

the PROMs-based label “Maladaptive Coper with high avoidance and catastrophic disability” to generate the clinical insight: “*Structural severity is compounded by a maladaptive psychosocial profile, suggesting interventions targeting coping behaviors may be prerequisite for successful engagement with physical rehabilitation.*”

Table 6.1: Summary of dominant PROMs, clinical, and imaging findings across pilot patients.

	Patient A (Adaptive)	Patient B (Intermed.)	Patient C (Maladaptive)
Age, Gender	60, M	65, M	35, F
Pain Location	Lumbar	Lumbar	Lumbar
BMI	29.0	25.0	27.5
Pain Duration	>12 weeks	>12 weeks	3 years
Pain (NRS 0–10)	7.0	9.0	10.0
Disability (0–24)	6	8	19
Physical QOL (0–100)	78.6	57.1	32.1
Psychological QOL (0–100)	66.7	75.0	25.0
Avoidance Coping (1–4)	1.0	1.0	2.0
Positive Attitude (1–4)	4.0	1.0	1.7
Dominant imaging findings	L3-L4 canal stenosis, L4-L5 protrusion & foraminal stenosis	L2-L3 herniation & foraminal stenosis, L5-S1 spondylolisthesis	L3-L4 foraminal stenosis & degeneration (Pfirrmann 5), L4-L5 severe canal stenosis

6.4.2 Reporting Efficiency

AI-assisted reporting resulted in substantial reductions in active reporting time for both radiologists (Table 6.2). Radiologist 1 showed a 75–85% reduction, while Radiologist 2 showed a 40–57% reduction. Both noted the AI draft was highly concordant with their initial assessment, requiring minimal edits.

Table 6.2: Reporting time efficiency: MRI-only vs. AI-assisted editing.

Radiologist	Patient	MRI-Only (min:sec)	AI-Assisted (min:sec)
Radiologist 1	A	1:20	0:25
	B	1:45	0:20
	C	1:51	0:20
Radiologist 2	A	2:20	1:00
	B	2:50	1:08
	C	2:26	1:20

6.4.3 Impact on Treatment Planning

The integration of PROMs data directly altered clinical reasoning. For the Adaptive Coper (Patient A), radiologists maintained their initial conservative management plan. For the Maladaptive Coper (Patient C), both fundamentally

reprioritized their recommendations, explicitly moving “psychological therapy” and “coping skills training” to a first-line intervention, while demoting structural interventions to a later stage.

6.4.4 Radiologist Feedback and Qualitative Insights

Both participating radiologists provided reflective debriefing comments following the proof-of-concept evaluation. Given the exploratory nature of this pilot, their full verbatim statements are reported below to maximize analytic transparency, followed by a thematic interpretation.

Radiologist 1 *“The comparison between specialist reporting and reporting obtained with artificial intelligence tends to be very similar to the specialist’s reporting. So this is definitely the future, because it will already provide you with a perfect report. In fact, artificial intelligence could potentially even correct some of the radiologist’s oversights or some details the radiologist might have missed, so the future certainly lies more with artificial intelligence than with direct specialist reporting.”*

This radiologist also noted adapting his reporting style after reviewing the AI outputs, beginning to incorporate more standardized terminology (e.g., Pfirrmann grading) that he had not previously used consistently.

Radiologist 2 *“Probably in a patient with multiple and pathological findings as in this case, assistance with artificial intelligence can reduce the reporting time by highlighting the existing problems.”*

This observation on efficiency gains was particularly emphasized for cases with multiregional or complex pathology, aligning with the measured time savings for Patient C.

Thematic Interpretation and Workflow Impact Analytic interpretation of these statements and subsequent discussions identified three cross-cutting themes with direct workflow implications:

1. **Efficiency and Augmented Accuracy:** Both experts perceived high concordance between AI-assisted and conventional reporting. They emphasized the tool’s potential to reduce oversight and expedite the reporting process,

especially for complex cases, thereby augmenting rather than replacing their expertise.

2. **Enhanced Contextualization:** The integration of patient-reported outcome measures (PROMs) was noted as valuable for interpreting symptom–imaging discordance. Radiologists reported that this biopsychosocial context supported more coherent clinical reasoning and a clearer understanding of why a patient’s experience may not correlate with structural findings.
3. **Tailored Clinical Communication:** Both participants highlighted the utility of incorporating psychosocial context into the report narrative. They suggested this could improve interdisciplinary communication, for instance, by using contextual phrases such as “*findings consistent with chronic degenerative changes in a patient reporting preserved function*” to guide treatment planning.

Qualitatively, this pilot suggested that the integrated framework could enhance interpretive confidence and guide more personalized reporting. The complete draft reports generated by LumbarAI for the three evaluated patient cases, including the explainable AI (XAI) visualizations (heatmaps and annotations) reviewed by the radiologists, are provided in Appendix A.

6.5 Discussion

This pilot study provides empirical evidence that integrating the framework’s core modules adds tangible clinical value, empirically probing and de-risking the key requirements articulated in Chapter 3.

6.5.1 Synthesis of Evidence

The study demonstrates *workflow feasibility* and *clinical sense-making* through significant time savings and positive usability feedback. Most importantly, the changed treatment plans for the Maladaptive Coper provide concrete evidence of *personalized clinical utility*, showing how a synthesized view can pivot specialist reasoning from a purely structural to a biopsychosocial approach.

6.5.2 Limitations and the Path to Automation

The primary limitation of this study is its deliberately small scale and illustrative design. With only three cases, no claims of generalizability, population-level effectiveness, or diagnostic performance are made. Instead, the study addresses a different and prerequisite question: whether multimodal integration of imaging

and psychosocial data is clinically intelligible, workflow-compatible, and capable of altering expert reasoning when presented in a coherent form. Within this scope, the proof-of-concept achieves its objective.

The manual synthesis step, while methodologically intentional for this phase, is the key barrier to scalability. This illustrates the necessity and defines the precise functional requirement for the envisioned clinically governed RAG-based LLM Synthesis Module (Module 3): to automate this rule-based mapping and narrative assembly.

6.5.3 Future Directions

The clear next step, directly informed by this proof-of-concept, is the technical development of Module 3. The radiologists’ validated need for “narrative tailoring” provides a concrete specification for an LLM: to generate draft reports that perform the manual synthesis demonstrated here, but at scale. A subsequent controlled study should quantitatively measure the impact of such automated, integrated reports on diagnostic confidence and clinical decision-making.

6.6 Conclusion

This proof-of-concept chapter bridges the validation of standalone AI components (Chapters 4 and 5) and the vision of an integrated clinical AI ecosystem (Chapter 3). By demonstrating that a manually synthesized multimodal profile can improve reporting efficiency and alter treatment planning in clinically meaningful ways, it provides **robust preliminary evidence of feasibility, interpretability, and clinical plausibility**.

The study confirms that the major translational barrier is no longer the accuracy of individual AI models, but the *automation of synthesis*. Therefore, this work concludes that the proposed framework is conceptually sound and practically plausible, delivering a clear, user-informed mandate for the next phase of development: engineering the AI synthesizer required to realize personalized, biopsychosocial spine care.

7

Summary, General Discussion, and Conclusions

7.1 Thesis Summary

This dissertation was driven by the core problem that the management of chronic spinal pain remains isolated, with structural imaging and psychosocial assessment operating in parallel, often discordant, clinical pathways. To bridge this gap, it proposed, developed, and evaluated a modular AI framework designed to integrate these dimensions for personalized care. The following summary recapitulates the journey and contributions of each empirical chapter.

Chapter 3 established the foundational blueprint. Moving beyond a critique of isolated AI tools, it synthesized stakeholder needs into a principled three-component architecture for a clinically integrated AI ecosystem: (1) automated spine MRI grading, (2) multimodal psychosocial phenotyping, and (3) AI-synthesized reporting. This chapter transformed the identified clinical problem into a concrete engineering and research agenda.

Chapter 4 delivered the first pillar: a clinically-aligned imaging module. By benchmarking and then outperforming a state-of-the-art model, LumbarAI demonstrated that AI could achieve specialist-level agreement in quantifying structural pathology. More importantly, its hybrid, curriculum-based design proved that AI could be engineered to mirror, and accelerate, the established radiological workflow, providing not just predictions but explainable, workflow-ready outputs.

Chapter 5 delivered the second pillar and a pivotal scientific insight: a robust,

data-driven typology of patient experience. It identified Avoidance and Positive Attitude as the primary determinants of pain intensity and revealed two stable biopsychosocial phenotypes (Adaptive and Maladaptive Copers) with starkly different disability and quality-of-life profiles. Critically, it provided powerful empirical evidence for a core tenet of the biopsychosocial model: these disabling phenotypes were completely dissociated from the severity or extent of findings on spinal MRI.

Chapter 6 integrated the first two pillars into a proof-of-concept, addressing the framework’s core integrative premise. It demonstrated that a manually synthesized view of imaging and phenotype data was clinically intelligible, significantly reduced specialist reporting time, and, most importantly, altered treatment prioritization in a clinically plausible manner, specifically by highlighting psychosocial interventions for high-risk patients. This chapter de-risked the fundamental assumption that such synthesis provides actionable value beyond isolated assessment.

Collectively, these chapters fulfill the dissertation’s original aim: to move from identifying the integration gap to providing and validating a coherent, component-by-component blueprint for closing it.

7.2 General Discussion

The empirical work presented leads to several higher-order conceptual discussions that challenge prevailing paradigms and clarify the path forward for AI in personalized medicine.

7.2.1 Imaging Severity and the Limits of Structural Explanation

A cornerstone finding of this thesis is the robust statistical dissociation between imaging-derived pathology and patient-reported suffering (Chapter 5). This ‘clinical–radiological decoupling’ is not an artifact but a fundamental characteristic of chronic spinal pain that our data powerfully quantifies. It carries a profound implication: **imaging AI, no matter how accurate, is inherently limited as a standalone explanatory tool for symptom burden.** The current dominant paradigm in medical AI, focusing narrowly on beating human performance on detection or grading tasks, is therefore insufficient for conditions where illness is not reducible to anatomy. Our work argues that the evaluation of imaging AI must evolve to assess its performance within integrative models that include

patient context, moving beyond diagnostic accuracy to predictive and explanatory validity for clinical outcomes.

7.2.2 Coping Phenotypes as Organizing Principles

If spinal morphology is a poor organizer of clinical reality, what serves better? Our analysis identifies **coping strategy profiles** as a more potent organizing principle. The mediation hierarchy (where foundational strategies like Problem Orientation operate through proximal ones like Avoidance) and the distinct, stable clusters (Adaptive vs. Maladaptive) provide a coherent, data-driven schema for the ‘psychosocial’ component of pain. This contributes to biopsychosocial models by moving them from a heuristic checklist to a quantifiable, mechanistic framework. The phenotype is not a secondary modifier of a structural diagnosis; for a substantial patient subgroup, it is the primary determinant of disability. This recenters the clinical question from “What do we see?” to “How does this person cope with what we see?”

7.2.3 AI as Clinical Decision Support

The proof-of-concept (Chapter 6), though small in scale, is pivotal for its demonstration of **clinical sense-making**. Its value lies not in automation, the synthesis was deliberately manual, but in proving that the output of integration is intelligible and action-guiding for specialists. This distinction is crucial. It shifts the objective of clinical AI from pure task automation (replacing the radiologist) to **augmentation and sense-making** (equipping the radiologist with synthesized context). The observed change in treatment prioritization, where psychosocial interventions were moved forward for maladaptive copers, is a direct example of AI enabling a more holistic, biopsychosocial reasoning process. This validates the framework’s core design principle: AI should synthesize information to support decisions, not make them autonomously.

7.3 Future Work

This thesis provides a validated foundation; the following directions outline the necessary work to translate it into practice.

- **Automation of Synthesis (Module 3):** The immediate next step is the development of the LLM-based synthesis module. This involves engineering a constrained, deterministic language system that codifies the rule-based integra-

tion logic validated in Chapter 6, enabling scalable generation of contextualized draft reports.

- **Prospective Validation:** A controlled, prospective study is required to quantitatively measure the impact of integrated AI reports on clinical outcomes (e.g., time to appropriate treatment, patient satisfaction, functional improvement) compared to standard care.
- **Workflow Integration:** Research must address the health systems challenges of embedding the framework into live clinical workflows, including interoperability with hospital information systems, defining referral triggers, and managing the handoff of synthesized reports to treating teams.
- **Generalization:** The framework’s architecture is generalizable. Future work should test its application to other musculoskeletal domains (e.g., osteoarthritis, shoulder pain) where similar discordance between imaging and symptoms exists, and to other multimodal data sources (e.g., genomics, wearable sensors).

7.4 General Conclusion

This dissertation set out to address a fundamental disconnect in spine care. It has demonstrated that the siloed assessment of a multidimensional condition is not only a clinical inefficiency but a scientific and operational error. Through the consecutive development and validation of a modular AI framework, this work provides a coherent alternative.

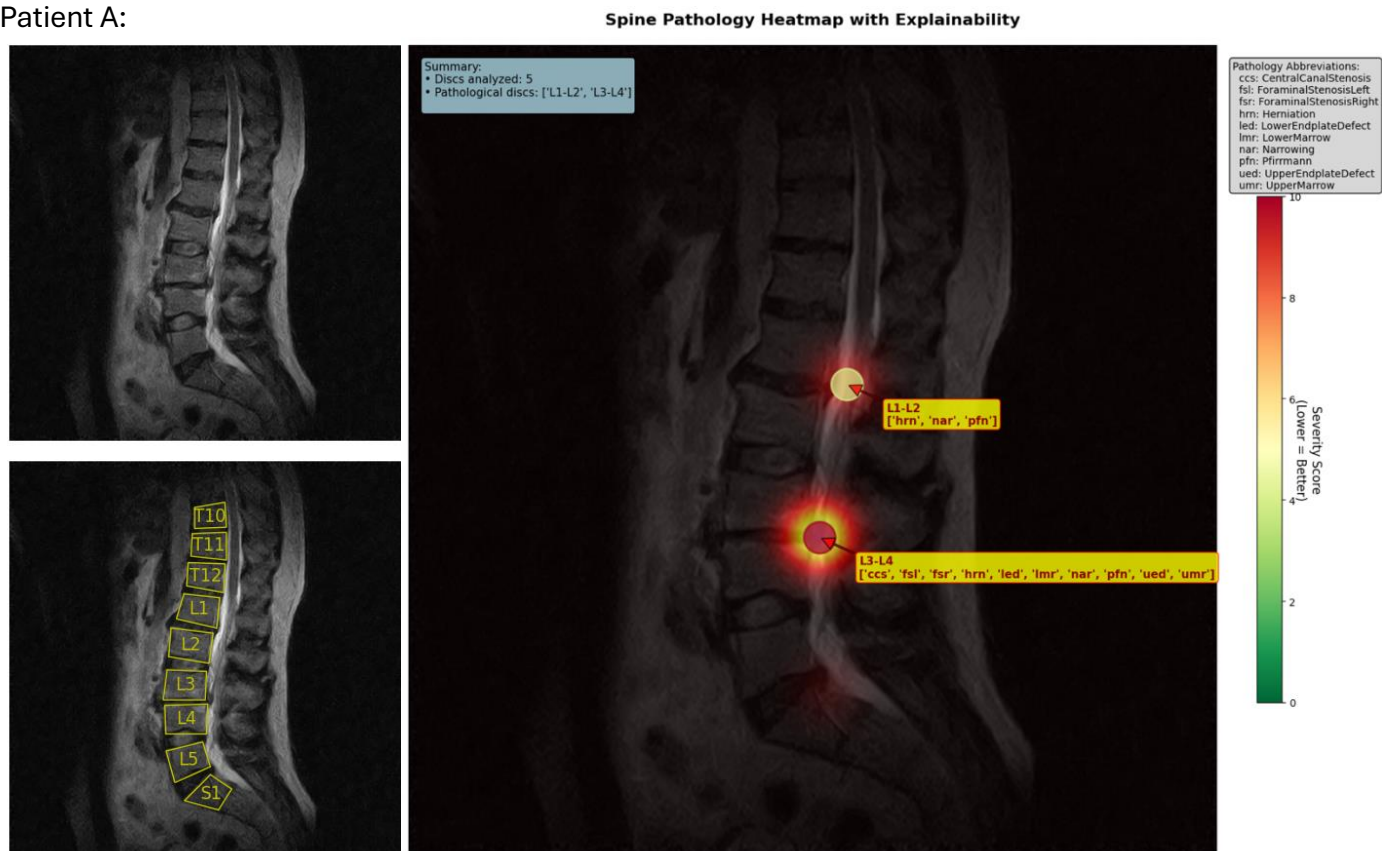
The thesis demonstrated that AI can reliably quantify both the structural and psychosocial pillars of chronic pain. It challenged the primacy of the structural narrative by providing robust evidence that disabling phenotypes are independent of imaging severity. Finally, it enabled a new paradigm by demonstrating that synthesizing these pillars alters specialist reasoning in a clinically meaningful way, turning the biopsychosocial model from an abstract ideal into an engineerable workflow.

The contribution therefore extends beyond spine care. It offers a template for AI in medicine at large: a move from isolated, task-specific models toward **integrative, sense-making systems** that respect clinical complexity. By proving that technology can help reconcile the long-observed paradox between pathology and illness, this work points toward a future where AI does not simply see what we see, but helps us understand what it means.

A

Appendix: LumbarAI Output and Draft Report

Patient A:



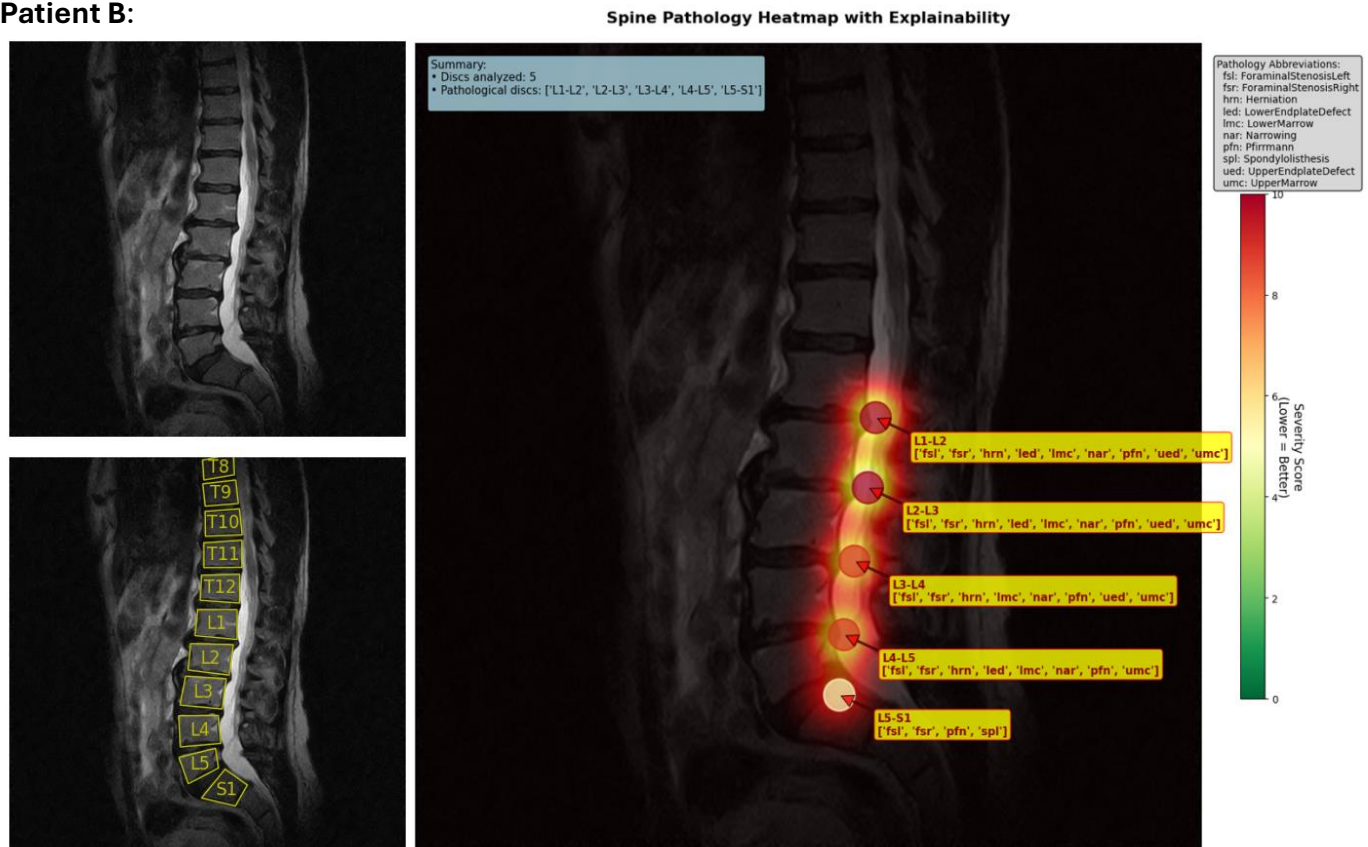
AI-Generated MRI interpretation output:

At L1-L2, there is Pfirrmann grade 4 disc degeneration, narrowing grade 3. At L3-L4, severe disc degeneration with Pfirrmann grade 5, narrowing grade 4, central canal stenosis grade 3 which represents severe stenosis, upper and lower endplate defects present, upper and lower marrow changes present, bilateral foraminal stenosis, and herniation present. At L5-S1, findings show mild disc degeneration Pfirrmann grade 3.

Summary for Patient A:

Severe degeneration at L3-L4 (grade 5). Most severe pathology concentrated at L3-L4 with extensive multi-faceted changes including central canal stenosis (grade 3), endplate defects, marrow changes, bilateral foraminal stenosis, and herniation. Herniation also present at L1-L2. Overall pattern shows a heterogeneous distribution with localized severe pathology at L3-L4 and well-preserved adjacent levels.

Patient B:



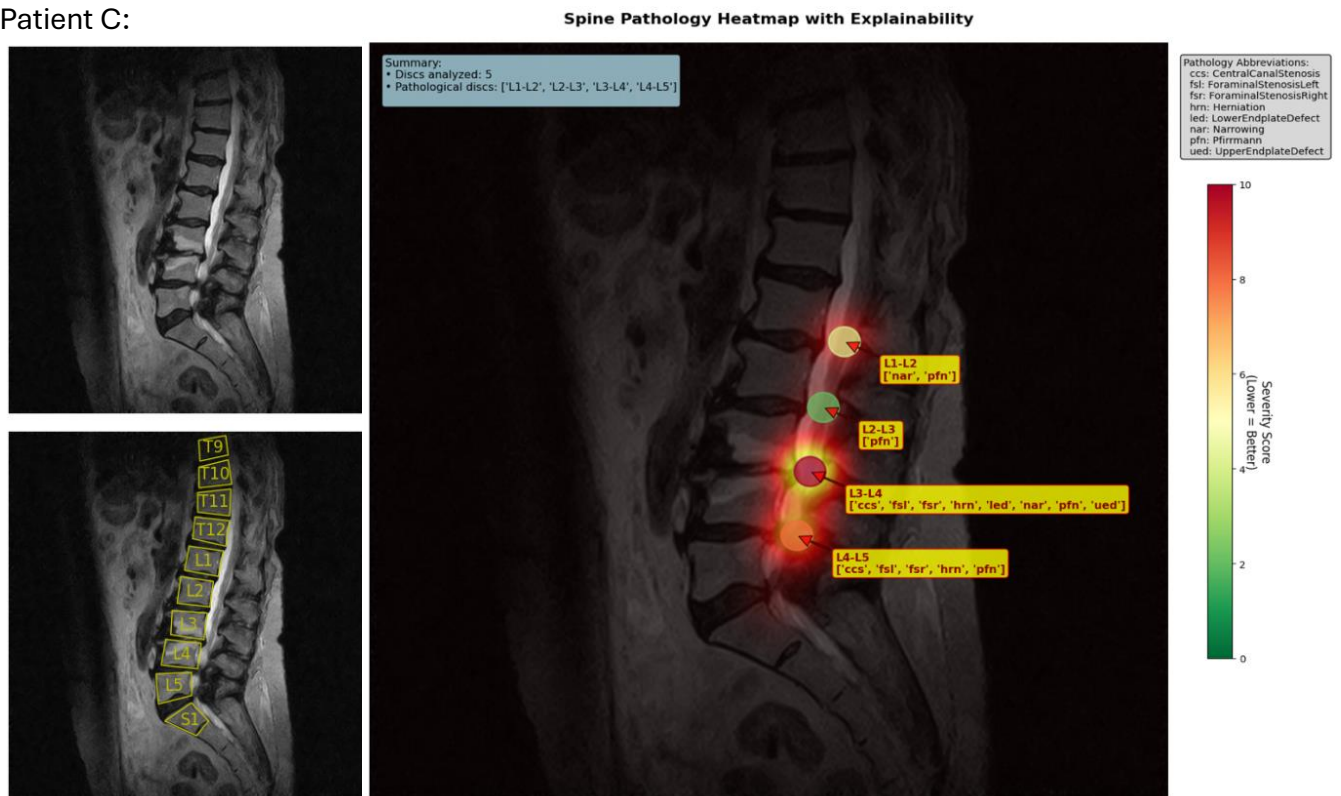
AI-Generated MRI interpretation output:

At L1-L2, there is Pfirsman grade 4 disc degeneration, narrowing grade 3, upper and lower endplate defects present, upper and lower marrow changes present, bilateral foraminal stenosis, and herniation present. At L2-L3, the most severe pathology is observed with Pfirsman grade 5 disc degeneration, narrowing grade 4, upper and lower endplate defects present, upper and lower marrow changes present, bilateral foraminal stenosis, and herniation present. At L3-L4, findings include Pfirsman grade 4 disc degeneration, narrowing grade 3, upper endplate defect present, lower endplate defect absent, upper and lower marrow changes present, bilateral foraminal stenosis, and herniation present. At L4-L5, there is Pfirsman grade 4 disc degeneration, narrowing grade 3, central canal stenosis grade 1, upper endplate defect absent, lower endplate defect present, upper and lower marrow changes present, bilateral foraminal stenosis, and herniation present. At L5-S1, findings show Pfirsman grade 3 disc degeneration, narrowing grade 2, spondylolisthesis present.

Summary for Patient B:

The MRI finding show Pfirsman grades range from 3 to 5, with most severe degeneration at L2-L3 (grade 5). Widespread endplate defects present at L1-L2, L2-L3, L3-L4, and L4-L5. Bilateral foraminal stenosis present at all levels except L1-L2. Herniation present at L1-L2, L2-L3, L3-L4, and L4-L5. Unique finding of spondylolisthesis at L5-S1, which is not present in the other two patients. Central canal stenosis uniformly grade 1 across all levels. Overall pattern demonstrates extensive multi-level pathology with widespread endplate defects and foraminal stenosis.

Patient C:



AI-Generated MRI interpretation output:

At L1-L2, there is Pfirrmann grade 3 disc degeneration, narrowing grade 3. At L2-L3, findings include Pfirrmann grade 3 disc degeneration, narrowing grade 2. At L3-L4, the most severe pathology is observed with Pfirrmann grade 5 disc degeneration, narrowing grade 4, central canal stenosis grade 2, upper and lower endplate defects present, upper and lower marrow changes present, bilateral foraminal stenosis, and herniation present. At L4-L5, there is Pfirrmann grade 4 disc degeneration, narrowing grade 2, severe central canal stenosis grade 3, bilateral foraminal stenosis, and herniation present. At L5-S1, findings show mild/moderate disc degeneration with Pfirrmann grade 3.

Summary for Patient C:

Moderate to severe degenerative changes with Pfirrmann grades ranging from 3 to 5. Most severe disc degeneration at L3-L4 (grade 5) with extensive multi-faceted pathology including endplate defects, marrow changes, bilateral foraminal stenosis, and herniation. Significant central canal stenosis at L4-L5 (grade 3) and L3-L4 (grade 2). Herniation present at L3-L4 and L4-L5. Overall pattern shows moderate to severe degeneration with notable central canal stenosis.

Bibliography

- [1] WHO low back pain factsheet. Low back pain, jun 2023. URL <https://www.who.int/news-room/fact-sheets/detail/low-back-pain>.
- [2] Manuela L Ferreira, Katie De Luca, Lydia M Haile, Jaimie D Steinmetz, Garland T Culbreth, Marita Cross, Jacek A Kopec, Paulo H Ferreira, Fiona M Blyth, Rachelle Buchbinder, et al. Global, regional, and national burden of low back pain, 1990–2020, its attributable risk factors, and projections to 2050: a systematic analysis of the global burden of disease study 2021. *The Lancet Rheumatology*, 5(6):e316–e329, 2023.
- [3] Mengyi Zhu, Jiarui Zhang, Diefei Liang, Junxiong Qiu, Yuan Fu, Zhaopei Zeng, Jingjun Han, Junmeng Zheng, and Liling Lin. Global and regional trends and projections of chronic pain from 1990 to 2035: Analyses based on global burden of diseases study 2019. *British Journal of Pain*, 19(2):125–137, 2025.
- [4] Nan Yang, Jingkai Di, Weihao Wang, and Haoyu Feng. Global burden of low back pain from 1990 to 2021: a comprehensive analysis of risk factors and trends using the global burden of disease study 2021. *BMC Public Health*, 25(1):1886, 2025.
- [5] Mohsen Naghavi, Giulia Zamagni, Cristiana Abbafati, Benedetta Armocida, Antonella Agodi, Gianfranco Alicandro, Franca Barbic, Martina Barchitta, Matteo Bauckneht, Massimiliano Beghi, et al. State of health and inequalities among italian regions from 2000 to 2021: a systematic analysis based on the global burden of disease study 2021. *The Lancet Public Health*, 10(4):e309–e320, 2025.
- [6] Francis Fatoye, Tadesse Gebrye, Cormac G Ryan, Ushotanefe Useh, and Chidozie Mbada. Global and regional estimates of clinical and economic burden of low back pain in high-income countries: a systematic review and meta-analysis. *Frontiers in public health*, 11: 1098100, 2023.
- [7] Damian Hoy, Lyn March, Peter Brooks, Fiona Blyth, Anthony Woolf, Christopher Bain, Gail Williams, Emma Smith, Theo Vos, Jan Barendregt, et al. The global burden of low back pain: estimates from the global burden of disease 2010 study. *Annals of the rheumatic diseases*, 73(6):968–974, 2014.
- [8] World Health Organization et al. Who releases guidelines on chronic low back pain. *Geneva: WHO*, 2023.
- [9] Amir Qaseem, Devan Kansagara, Jennifer S Lin, Reem A Mustafa, Timothy J Wilt, and Clinical Guidelines Committee of the American College of Physicians*. The development of clinical guidelines and guidance statements by the clinical guidelines committee of the american college of physicians: update of methods. *Annals of internal medicine*, 170(12): 863–870, 2019.
- [10] YiFan Wu, Sarah Wulf Hanson, Garland Culbreth, Caroline Purcell, Peter Brooks, Jacek Kopec, Lyn March, Anthony D Woolf, Maja Pasovic, Erin Hamilton, et al. Assessing the impact of health-care access on the severity of low back pain by country: a case study within the gbd framework. *The Lancet Rheumatology*, 6(9):e598–e606, 2024.
- [11] Jan Hartvigsen, Mark J Hancock, Alice Kongsted, Quinette Louw, Manuela L Ferreira, Stéphane Genevay, Damian Hoy, Jaro Karppinen, Glenn Pransky, Joachim Sieper, et al. What low back pain is and why we need to pay attention. *The Lancet*, 391(10137): 2356–2367, 2018.
- [12] Paul F Beattie and Steven P Meyers. Magnetic resonance imaging in low back pain:

BIBLIOGRAPHY

- general principles and clinical issues. *Physical therapy*, 78(7):738–753, 1998.
- [13] Christian WA Pfirrmann, Alexander Metzdorf, Marco Zanetti, Juerg Hodler, and Norbert Boos. Magnetic resonance classification of lumbar intervertebral disc degeneration. *spine*, 26(17):1873–1878, 2001.
- [14] Yue-Hui Zhang, Chang-Qing Zhao, Lei-Sheng Jiang, Xiao-Dong Chen, and Li-Yang Dai. Modic changes: a systematic review of the literature. *European Spine Journal*, 17(10):1289–1299, 2008.
- [15] Seunghun Lee, Joon Woo Lee, Jin Sup Yeom, Ki-Jeong Kim, Hyun-Jib Kim, Soo Kyo Chung, and Heung Sik Kang. A practical mri grading system for lumbar foraminal stenosis. *American Journal of Roentgenology*, 194(4):1095–1098, 2010.
- [16] Estanislao Arana, FM Kovacs, Ana Royuela, A Estremera, H Sarasibar, Guillermo Amengual, Isabel Galarraga, C Martínez, A Muriel, V Abaira, et al. Influence of nomenclature in the interpretation of lumbar disk contour on mr imaging: a comparison of the agreement using the combined task force and the nordic nomenclatures. *American journal of neuroradiology*, 32(6):1143–1148, 2011.
- [17] Roger Chou, Amir Qaseem, Douglas K Owens, Paul Shekelle, and Clinical Guidelines Committee of the American College of Physicians*. Diagnostic imaging for low back pain: advice for high-value health care from the american college of physicians. *Annals of internal medicine*, 154(3):181–189, 2011.
- [18] Waleed Brinjikji, Patrick H Luetmer, Bryan Comstock, Brian W Bresnahan, LE Chen, RA Deyo, Safwan Halabi, JA Turner, AL Avins, K James, et al. Systematic literature review of imaging features of spinal degeneration in asymptomatic populations. *American journal of neuroradiology*, 36(4):811–816, 2015.
- [19] Waleed Brinjikji and Phillip M Young. Update on state of the art magnetic resonance angiography techniques. *Journal of Vascular Diagnostics*, pages 9–16, 2015.
- [20] Jon D Lurie, Anna N Tosteson, Tor D Tosteson, Eugene Carragee, John Carrino, Jay Kaiser, Roberto T Blanco Sequeiros, Amy Rosen Lecomte, Margaret R Grove, Emily A Blood, et al. Reliability of readings of magnetic resonance imaging features of lumbar spinal stenosis. *Spine*, 33(14):1605–1610, 2008.
- [21] Ivan SK Thong, Mark P Jensen, Jordi Miró, and Gabriel Tan. The validity of pain intensity measures: what do the nrs, vas, vrs, and fps-r measure? *Scandinavian journal of pain*, 18(1):99–107, 2018.
- [22] R Padua, L Padua, E Ceccarelli, E Romanini, G Zanoli, R Bondi, and A Campi. Italian version of the roland disability questionnaire, specific for low back pain: cross-cultural adaptation and validation. *European Spine Journal*, 11(2):126–129, 2002.
- [23] Jeremy Fairbank. The oswestry low back pain disability questionnaire. *Physiotherapy*, 1980.
- [24] Suzanne M Skevington, Mahmoud Lotfy, and Kathryn A O’Connell. 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. *Quality of life Research*, 13(2):299–310, 2004.
- [25] John E Ware, Kristin K Snow, Mark Kosinski, and Barbara Gandek. Sf-36 health survey. *Manual and interpretation guide*, 2, 1993.
- [26] Claudio Sica, Cristina Magni, Marta Ghisi, Gianmardo Altoè, Cecilia Sighinolfi, Luigi Rocco Chiri, Sandro Franceschini, et al. Coping orientation to problems experienced-nuova versione italiana (cope-nvi): uno strumento per la misura degli stili di coping. *Psicoterapia cognitiva e comportamentale*, 14(1):27, 2008.
- [27] Michael JL Sullivan, Scott R Bishop, and Jayne Pivik. The pain catastrophizing scale: development and validation. *Psychological assessment*, 7(4):524, 1995.
- [28] Jay R Ebert, Peter K Edwards, Sven Klinken, Michael Fallon, David J Wood, and Gregory C Janes. 10-year clinical and mri-based outcomes of a randomized controlled trial evaluating a 6-week return to full weightbearing after matrix-induced autologous chondrocyte implantation. *Orthopaedic Journal of Sports Medicine*, 13(10):23259671251383136, 2025.
- [29] Ashwin Bhadresha, Owen John Lawrence, and Michael JH McCarthy. A comparison of

- magnetic resonance imaging muscle fat content in the lumbar paraspinal muscles with patient-reported outcome measures in patients with lumbar degenerative disk disease and focal disk prolapse. *Global spine journal*, 6(4):401–410, 2016.
- [30] George L Engel. The need for a new medical model: a challenge for biomedicine. *Science*, 196(4286):129–136, 1977.
 - [31] Panzhang Wang, Tao Li, Lei Yu, Liang Zhou, and Tao Yan. Towards an effective framework for integrating patient-reported outcomes in electronic health records. *Digital Health*, 8: 20552076221112152, 2022.
 - [32] Jakub Fusiak, Kousha Sarpari, Inger Ma, Ulrich Mansmann, and Verena S Hoffmann. Practical applications of methods to incorporate patient preferences into medical decision models: a scoping review. *BMC Medical Informatics and Decision Making*, 25(1):109, 2025.
 - [33] Jeppe Eriksen, Ann Bygholm, and Pernille Bertelsen. The purpose of patient-reported outcome (pro) post its digitalization and integration into clinical practice: an interdisciplinary redefinition resembling pros theoretical and practical evolvement. *Applied Sciences*, 10(21):7507, 2020.
 - [34] Pat Croskerry. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic medicine*, 78(8):775–780, 2003.
 - [35] Zunaid Ismail Vally, Razia AG Khammissa, Gal Feller, Raoul Ballyram, Michaela Beetge, and Liviu Feller. Errors in clinical diagnosis: a narrative review. *Journal of International Medical Research*, 51(8):03000605231162798, 2023.
 - [36] Irena Kovačević, Jadranka Pavić, Biljana Filipović, Štefanija Ozimec Vulinec, Boris Ilić, and Davorina Petek. Integrated approach to chronic pain—the role of psychosocial factors and multidisciplinary treatment: a narrative review. *International journal of environmental research and public health*, 21(9):1135, 2024.
 - [37] Geert Litjens, Thijs Kooi, Babak Ehteshami Babenko, Nico Karssemeijer, Cornelis L Hendriks, and Jeroen A van der Laak. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
 - [38] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
 - [39] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Automated detection, labelling and radiological grading of clinical spinal mris. *Scientific Reports*, 14(1):14993, 2024.
 - [40] Dalwinder Singh, Jimmy Singla, Mohammad Khalid Imam Rahmani, Sultan Ahmad, Masood Ur Rehman, Sudan Jha, Deepak Prashar, Jabeen Nazeer, et al. Lumbar spine disease detection: enhanced cnn model with improved classification accuracy. *IEEE Access*, 11:141889–141901, 2023.
 - [41] Yasser Mahmoud, Kaleb Horvath, and Yi Zhou. Deep learning for predicting rehabilitation success: Advancing clinical and patient-reported outcome modeling. *Electronics*, 14(6): 1082, 2025.
 - [42] Yunsoo Lee, Tariq Ziad Issa, and Alexander R Vaccaro. State-of-the-art applications of patient-reported outcome measures in spinal care. *JAAOS-Journal of the American Academy of Orthopaedic Surgeons*, 31(20):e890–e897, 2023.
 - [43] Ali Algarni. Careassist gpt improves patient user experience with a patient centered approach to computer aided diagnosis. *Scientific Reports*, 15(1):22727, 2025.
 - [44] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.
 - [45] Mattia Savardi, Alberto Signoroni, Sergio Benini, Filippo Vaccher, Matteo Alberti, Pietro Ciolli, Nunzia Di Meo, Teresa Falcone, Marco Ramanzin, Barbara Romano, et al. Upskilling or deskilling? measurable role of an ai-supported training for radiology residents: a lesson from the pandemic. *Insights into Imaging*, 16(1):23, 2025.
 - [46] Vasantha Gurusamy, Anandhavalli Muniasamy, Shruthi A Gauthaman, and Hameed Abdulla. Human-ai collaboration in healthcare: The evolving role of healthcare professionals.

BIBLIOGRAPHY

- In *Evaluation and Assessment of AI-Driven Systems in Hospitals*, pages 341–364. IGI Global Scientific Publishing, 2026.
- [47] Alemu Sisay Nigru, Sergio Benini, Matteo Bonetti, Graziella Bragaglio, Michele Frigerio, Federico Maffezzoni, and Riccardo Leonardi. External validation of spinenetv2 on a comprehensive set of radiological features for grading lumbosacral disc pathologies. *North American Spine Society Journal (NASSJ)*, 20:100564, 2024.
 - [48] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Spinenetv2: automated detection, labelling and radiological grading of clinical mr scans. *arXiv preprint arXiv:2205.01683*, 2022.
 - [49] Damian Hoy, Christopher Bain, Gail Williams, Lyn March, Peter Brooks, Fiona Blyth, Anthony Woolf, Theo Vos, and Rachelle Buchbinder. A systematic review of the global prevalence of low back pain. *Arthritis & rheumatism*, 64(6):2028–2037, 2012.
 - [50] Aimin Wu, Lyn March, Xuanqi Zheng, Jinfeng Huang, Xiangyang Wang, Jie Zhao, Fiona M Blyth, Emma Smith, Rachelle Buchbinder, and Damian Hoy. Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the global burden of disease study 2017. *Annals of translational medicine*, 8(6):299, 2020.
 - [51] Nicholas LS Roberts, W Cliff Mountjoy-Venning, Mina Anjomshoa, Joseph Adel Mattar Banoub, and Yasin Jemal Yasin. Gbd 2017 disease and injury incidence and prevalence collaborators. global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the global burden of disease study (vol 392, pg 1789, 2018). *Lancet*, 393(10190):E44–E44, 2019.
 - [52] Theo Vos, Stephen S Lim, Cristiana Abbafati, Kaja M Abbas, Mohammad Abbasi, Mitra Abbasifard, Mohsen Abbasi-Kangevari, Hedayat Abbastabar, Foad Abd-Allah, Ahmed Abdelalim, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The lancet*, 396(10258):1204–1222, 2020.
 - [53] Massimo Allegri, Silvana Montella, Fabiana Salici, Adriana Valente, Maurizio Marchesini, Christian Compagnone, Marco Baciarello, Maria Elena Manferdini, and Guido Fanelli. Mechanisms of low back pain: a guide for diagnosis and therapy. *F1000Research*, 5: F1000–Faculty, 2016.
 - [54] Chris Maher, Martin Underwood, and Rachelle Buchbinder. Non-specific low back pain. *The Lancet*, 389(10070):736–747, 2017.
 - [55] Hiroaki Nakashima, Yasutsugu Yukawa, Kota Suda, Masatsune Yamagata, Takayoshi Ueta, and Fumihiko Kato. Abnormal findings on magnetic resonance images of the cervical spines in 1211 asymptomatic subjects. *Spine*, 40(6):392–398, 2015.
 - [56] Crystian B Oliveira, Chris G Maher, Rafael Z Pinto, Adrian C Traeger, Chung-Wei Christine Lin, Jean-Francois Chenot, Maurits Van Tulder, and Bart W Koes. Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. *European Spine Journal*, 27(11):2791–2803, 2018.
 - [57] Coen J Itz, JW Geurts, M Van Kleef, and P Nelemans. Clinical course of non-specific low back pain: A systematic review of prospective cohort studies set in primary care. *European journal of pain*, 17(1):5–15, 2013.
 - [58] Rolf-Detlef Treede, Winfried Rief, Antonia Barke, Qasim Aziz, Michael I Bennett, Rafael Benoliel, Milton Cohen, Stefan Evers, Nanna B Finnerup, Michael B First, et al. Chronic pain as a symptom or a disease: the iasp classification of chronic pain for the international classification of diseases (icd-11). *pain*, 160(1):19–27, 2019.
 - [59] Jo Nijs, Adri Apeldoorn, Hank Hallegraef, Jacqui Clark, Rob Smeets, Anneleen Malfliet, E Lluch Girbes, Margot De Kooning, and Kelly Ickmans. Low back pain: guidelines for the clinical classification of predominant neuropathic, nociceptive, or central sensitization pain. *Pain physician*, 18(3):E333–46, 2015.
 - [60] Rahman Shiri, Jaro Karppinen, Päivi Leino-Arjas, Svetlana Solovieva, and Eira Viikari-Juntura. The association between obesity and low back pain: a meta-analysis. *American journal of epidemiology*, 171(2):135–154, 2010.
 - [61] Aline Ramond, Celine Bouton, Isabelle Richard, Yves Roquelaure, Christophe Baufreton, Erick Legrand, and Jean-François Huez. Psychosocial risk factors for chronic low back

- pain in primary care—a systematic review. *Family practice*, 28(1):12–21, 2011.
- [62] Marina B Pinheiro, Manuela L Ferreira, Kathryn Refshauge, Juan R Ordoñana, Gustavo C Machado, Lucas R Prado, Christopher G Maher, and Paulo H Ferreira. Symptoms of depression and risk of new episodes of low back pain: a systematic review and meta-analysis. *Arthritis care & research*, 67(11):1591–1603, 2015.
 - [63] Amir Qaseem, Timothy J Wilt, Robert M McLean, Mary Ann Forciea, and Clinical Guidelines Committee of the American College of Physicians*. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the american college of physicians. *Annals of internal medicine*, 166(7):514–530, 2017.
 - [64] Tarcisio Folly de Campos. Low back pain and sciatica in over 16s: assessment and management nice guideline [ng59]. *Journal of Physiotherapy*, 63(2):120, 2017.
 - [65] Nicholas Henschke, Christopher G Maher, Kathryn M Refshauge, Robert D Herbert, Robert G Cumming, Jane Bleasel, John York, Anurina Das, and James H McAuley. Prevalence of and screening for serious spinal pathology in patients presenting to primary care settings with acute low back pain. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 60(10):3072–3080, 2009.
 - [66] Karin Verkerk, Pim AJ Luijsterburg, Harard S Miedema, Annelies Pool-Goudzwaard, and Bart W Koes. Prognostic factors for recovery in chronic nonspecific low back pain: a systematic review. *Physical therapy*, 92(9):1093–1108, 2012.
 - [67] Manohar M Panjabi. Clinical spinal instability and low back pain. *Journal of electromyography and kinesiology*, 13(4):371–379, 2003.
 - [68] Michael T Modic and Jeffrey S Ross. Lumbar degenerative disk disease. *Radiology*, 245(1):43–61, 2007.
 - [69] Jeffrey G Jarvik and Richard A Deyo. Diagnostic evaluation of low back pain with emphasis on imaging. *Annals of internal medicine*, 137(7):586–597, 2002.
 - [70] Michael T Modic, PM Steinberg, JS Ross, TJ Masaryk, and JR Carter. Degenerative disk disease: assessment of changes in vertebral body marrow with mr imaging. *Radiology*, 166(1):193–199, 1988.
 - [71] Jeroen C van Rijn, Nina Klemetso, Johannes B Reitsma, Charles BLM Majoie, Frans J Hulsmans, Wilco C Peul, Jan Stam, Patrick M Bossuyt, and Gerard J den Heeten. Observer variation in mri evaluation of patients suspected of lumbar disk herniation. *American Journal of Roentgenology*, 184(1):299–303, 2005.
 - [72] Jeffrey G Jarvik, Laura S Gold, Bryan A Comstock, Patrick J Heagerty, Sean D Rundell, Judith A Turner, Andrew L Avins, Zoya Bauer, Brian W Bresnahan, Janna L Friedly, et al. Association of early imaging for back pain with clinical outcomes in older adults. *Jama*, 313(11):1143–1153, 2015.
 - [73] Eugene J Carragee, Todd F Alamin, Jonothan L Miller, and John M Carragee. Discographic, mri and psychosocial determinants of low back pain disability and remission: a prospective study in subjects with benign persistent back pain. *The spine journal*, 5(1):24–35, 2005.
 - [74] Geert Crombez, Christopher Eccleston, Stefaan Van Damme, Johan WS Vlaeyen, and Paul Karoly. Fear-avoidance model of chronic pain: the next generation. *The Clinical journal of pain*, 28(6):475–483, 2012.
 - [75] Dennis C Turk and Leah M Adams. Using a biopsychosocial perspective in the treatment of fibromyalgia patients. *Pain management*, 6(4):357–369, 2016.
 - [76] Nick Black. Patient reported outcome measures could help transform healthcare. *Bmj*, 346, 2013.
 - [77] Gillian A Hawker, Samra Mian, Tetyana Kendzerska, and Melissa French. Measures of adult pain: Visual analog scale for pain (vas pain), numeric rating scale for pain (nrs pain), mcgill pain questionnaire (mpq), short-form mcgill pain questionnaire (sf-mpq), chronic pain grade scale (cpgs), short form-36 bodily pain scale (sf-36 bps), and measure of intermittent and constant osteoarthritis pain (icoap). *Arthritis care & research*, 63(S11):S240–S252, 2011.
 - [78] Jeremy CT Fairbank and Paul B Pynsent. The oswestry disability index. *Spine*, 25(22):2940–2953, 2000.

BIBLIOGRAPHY

- [79] Martin Roland and Jeremy Fairbank. The roland–morris disability questionnaire and the oswestry disability questionnaire. *Spine*, 25(24):3115–3124, 2000.
- [80] JM1 Valderas, A Kotzeva, M Espallargues, G Guyatt, CE Ferrans, MY Halyard, DA Re-vicki, T Symonds, A Parada, and J Alonso. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Quality of life research*, 17(2):179–193, 2008.
- [81] Ala S Al Kafri, Sud Sudirman, Abir J Hussain, Paul Fergus, Dhiya Al-Jumeily, Mohammed Al-Jumaily, and Haya Al-Askar. A framework on a computer assisted and systematic methodology for detection of chronic lower back pain using artificial intelligence and computer graphics technologies. In *International Conference on Intelligent Computing*, pages 843–854. Springer, 2016.
- [82] Federico D’Antoni, Fabrizio Russo, Luca Ambrosio, Luca Bacco, Luca Vollero, Gianluca Vadalà, Mario Merone, Rocco Papalia, and Vincenzo Denaro. Artificial intelligence and computer aided diagnosis in chronic low back pain: A systematic review. *International journal of environmental research and public health*, 19(10):5971, 2022.
- [83] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [84] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.
- [85] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.
- [86] Yun Jiang, Yuan Zhang, Xin Lin, Jinkun Dong, Tongtong Cheng, and Jing Liang. Swinbts: A method for 3d multimodal brain tumor segmentation using swin transformer. *Brain sciences*, 12(6):797, 2022.
- [87] Md Sabbir Hossain, Mostafijur Rahman, Mumtahina Ahmed, Ashifur Rahman, Md Mohsin Kabir, MF Mridha, and Jungpil Shin. An in-depth review and analysis of deep learning methods and applications in spinal cord imaging. *Healthcare Analytics*, page 100429, 2025.
- [88] Yu-Cherng Chang, Cinthia Del Toro, Joseph P Gjola, Thiago A Braga, and Ty K Subhawong. Can artificial intelligence in spine imaging affect current practice?: Practical developments and their clinical status. *North American Spine Society Journal (NASSJ)*, page 100621, 2025.
- [89] Soolmaz Abbasi, Haoyu Lan, Jeiran Choupan, Nasim Sheikh-Bahaei, Gaurav Pandey, and Bino Varghese. Deep learning for the harmonization of structural mri scans: a survey. *BioMedical Engineering OnLine*, 23(1):90, 2024.
- [90] Terence P McSweeney, Aleksei Tiulpin, Simo Saarakkala, Jaakko Niinimäki, Rhydian Windsor, Amir Jamaludin, Timor Kadir, Jaro Karppinen, and Juhani Määtä. External validation of spinenet, an open-source deep learning model for grading lumbar disk degeneration mri features, using the northern finland birth cohort 1966. *Spine*, 48(7):484, 2023.
- [91] Alexandra Grob, Markus Loibl, Amir Jamaludin, Sebastian Winklhofer, Jeremy CT Fairbank, Tamás Fekete, François Porchet, and Anne F Mannion. External validation of the deep learning system ”spinenet” for grading radiological features of degeneration on mris of the lumbar spine. *European Spine Journal*, 31(8):2137–2148, 2022.
- [92] Xingkai Wu, Qianbo Song, Jiaxiang Zhou, Zhiyu Zhou, Guangru Cao, Kebing Jin, and Qian Du. External validation of spinenetv2 deep learning system for automated lumbar spine mri analysis: A multi-pathology diagnostic agreement study. *European Spine Journal*, pages 1–11, 2025.
- [93] Jiwoon Seo and Joon Woo Lee. Magnetic resonance imaging grading systems for central canal and neural foraminal stenoses of the lumbar and cervical spines with a focus on the lee grading system. *Korean Journal of Radiology*, 24(3):224, 2023.
- [94] James Meacock, Moritz Schramm, Senthil Selvanathan, Stuart Currie, Deborah Stocken, David Jayne, and Simon Thomson. Systematic review of radiological cervical foraminal

- grading systems. *Neuroradiology*, 63(3):305–316, 2021.
- [95] Dong Hwa Heo, Seok Woo Kim, Hsien-Hao Chang, and Choon-Keun Park. Usefulness of oblique lumbar magnetic resonance imaging for nerve root anomalies and extraforaminal entrapment lesions. *Asian Spine Journal*, 12(3):423, 2018.
 - [96] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.
 - [97] Chaoyi Li, Meng Li, Can Peng, and Brian C Lovell. Dynamic curriculum learning via in-domain uncertainty for medical image classification. In *International conference on medical image computing and computer-assisted intervention*, pages 747–757. Springer, 2023.
 - [98] Jun Luo, Gene Kitamura, Dooman Arefan, Emine Doganay, Ashok Panigrahy, and Shandong Wu. Knowledge-guided multiview deep curriculum learning for elbow fracture classification. In *International Workshop on Machine Learning in Medical Imaging*, pages 555–564. Springer, 2021.
 - [99] Scott D Tagliaferri, Maia Angelova, Xiaohui Zhao, Patrick J Owen, Clint T Miller, Tim Wilkin, and Daniel L Belavy. Artificial intelligence to improve back pain outcomes and lessons learnt from clinical classification approaches: three systematic reviews. *NPJ digital medicine*, 3(1):93, 2020.
 - [100] Paolo Giaccone, Elisabetta de Rinaldis, Federico D’Antoni, Fabrizio Russo, Luca Ambrosio, Giorgia Petrucci, Mario Merone, Leandro Pecchia, Sergio Iavicoli, Gianluca Vadalà, et al. Exploring the biopsychosocial impact of chronic low back pain in workers through artificial intelligence-driven phenotyping. *JOR spine*, 8(3):e70110, 2025.
 - [101] Héctor Hernández-Lázaro, María Teresa Mingo-Gómez, Sandra Jiménez-del Barrio, Silvia Lahuerta-Martín, Ignacio Hernando-Garijo, Ricardo Medrano-de-la Fuente, and Luis Ceballos-Laita. Researcher’s perspective on musculoskeletal conditions in primary care physiotherapy units through the international classification of functioning, disability, and health (icf): a scoping review. *Biomedicine*, 11(2):290, 2023.
 - [102] Earvin S Tio, Melissa C Misztal, and Daniel Felsky. Evidence for the biopsychosocial model of suicide: a review of whole person modeling studies using machine learning. *Frontiers in psychiatry*, 14:1294666, 2024.
 - [103] Michal Pruski, Simone Willis, and Kathleen Withers. A narrative review of the use of prompts and machine learning to impact value-based clinical decision-making. *BMC Medical Informatics and Decision Making*, 25:250, 2025.
 - [104] Benedict J Alter, Nathan P Anderson, Andrea G Gillman, Qing Yin, Jong-Hyeon Jeong, and Ajay D Wasan. Hierarchical clustering by patient-reported pain distribution alone identifies distinct chronic pain subgroups differing by pain intensity, quality, and clinical outcomes. *PLoS One*, 16(8):e0254862, 2021.
 - [105] Christophe Tanguay-Sabourin, Matt Fillingim, Marc Parisien, Gianluca V Guglietti, Azin Zare, Jax Norman, Ronrick Da-ano, Jordi Perez, Scott J Thompson, Marc O Martel, et al. A data-driven biopsychosocial framework determining the spreading of chronic pain. *medRxiv*, pages 2022–07, 2022.
 - [106] Hayato Shigetoh, Teppei Abiko, Michie Ohyama, Eiji Sakata, and Shin Murata. Subgroup characteristics in care workers with low back pain: cluster analysis-based severity of central sensitivity syndromes and low back pain. *European Spine Journal*, 33(4):1447–1454, 2024.
 - [107] Albert FG Leentjens and Stephen L Smith. Will the biopsychosocial model of medicine survive in the age of artificial intelligence and machine learning?, 2023.
 - [108] Karen Ribbons, Sarah Johnson, Elizabeth Ditton, Adrian Wills, Gillian Mason, Traci Flynn, Jodie Cochrane, Michael Pollack, Frederick Rohan Walker, and Michael Nilsson. Using presurgical biopsychosocial features to develop an advanced clinical decision-making support tool for predicting recovery trajectories in patients undergoing total knee arthroplasty: protocol for a prospective observational study. *JMIR research protocols*, 12(1):e48801, 2023.
 - [109] Michelle W Tong, Katharina Ziegeler, Virginie Kreutzinger, and Sharmila Majumdar. Explainable ai reveals tissue pathology and psychosocial drivers of opioid prescription for non-specific chronic low back pain. *Scientific Reports*, 15(1):30690, 2025.

BIBLIOGRAPHY

- [110] Bargava Subramanian, Naveen Kumarasami, Praveen Shastry, Raghotham Sripadraj, Kalyan Sivasailam, Abinaya Ramachandran, Sudhir MP, Kishore Prasath Venkatesh, et al. Ai-driven mri spine pathology detection: a comprehensive deep learning approach for automated diagnosis in diverse clinical settings. *arXiv preprint arXiv:2503.20316*, 2025.
- [111] Caroline Constant, Carl-Eric Aubin, Hilal Maradit Kremers, Diana V Vera Garcia, Cody C Wyles, Pouria Rouzrokh, and Annalise Noelle Larson. The use of deep learning in medical imaging to improve spine care: a scoping review of current literature and clinical applications. *North American Spine Society Journal (NASSJ)*, 15:100236, 2023.
- [112] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20(1):310, 2020.
- [113] J Raymond Geis, Adrian P Brady, Carol C Wu, Jack Spencer, Erik Ranschaert, Jacob L Jaremko, Steve G Langer, Andrea Borondy Kitts, Judy Birch, William F Shields, et al. Ethics of artificial intelligence in radiology: summary of the joint european and north american multisociety statement. *Radiology*, 293(2):436–440, 2019.
- [114] Mohamad Bydon and Asimina Dominari. Patient-reported outcomes in spine surgery: the wealth of data. *Neurosurgery*, 96(3S):S148–S153, 2025.
- [115] Margaret Bevans, Alyson Ross, and David Cella. Patient-reported outcomes measurement information system (promis): efficient, standardized tools to measure self-reported health and quality of life. *Nursing outlook*, 62(5):339–345, 2014.
- [116] Christian Quinones, Deepak Kumbhare, Bharat Guthikonda, and Stanley Hoang. Scoping review of machine learning and patient-reported outcomes in spine surgery. *Bioengineering*, 12(2):125, 2025.
- [117] Sadhana Kalidindi and R Prasanna Vadana. Overview of the potentially transforming role of large language models (llms) in creating patient-friendly summaries for radiology reports. *J Radiol Clin Imaging*, 7(10.26502), 2024.
- [118] Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384, 2024.
- [119] Rong Chen, Siyun Zhang, Yiyi Zheng, Qiuhua Yu, and Chuhuai Wang. Enhancing treatment decision-making for low back pain: a novel framework integrating large language models with retrieval-augmented generation technology. *Frontiers in Medicine*, 12:1599241, 2025.
- [120] Katharina Ziegeler, Virginie Kreutzinger, Michelle W Tong, Cynthia T Chin, Emma Bahroos, Po-Hung Wu, Noah Bonnheim, Aaron J Fields, Jeffrey C Lotz, Thomas M Link, et al. Information extraction from lumbar spine mri radiology reports using gpt4: Accuracy and benchmarking against research-grade comprehensive scoring. *Diagnostics*, 15(7):930, 2025.
- [121] Jen-Tang Lu, Stefano Pedemonte, Bernardo Bizzo, Sean Doyle, Katherine P Andriole, Mark H Michalski, R Gilberto Gonzalez, and Stuart R Pomerantz. Deep spine: automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. In *Machine Learning for Healthcare Conference*, pages 403–419. PMLR, 2018.
- [122] Aric Lee, Wilson Ong, Andrew Makmur, Yong Han Ting, Wei Chuan Tan, Shi Wei Desmond Lim, Xi Zhen Low, Jonathan Jiong Hao Tan, Naresh Kumar, and James TPD Hallinan. Applications of artificial intelligence and machine learning in spine mri. *Bio-engineering*, 11(9):894, 2024.
- [123] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.
- [124] Laura Steffny, Nanna Dahlem, Lea Reichl, Kevin Gisa, Tobias Greff, and Dirk Werth. Design of a human-in-the-loop centered ai-based clinical decision support system for professional care planning. In *HHA1 2023: Augmenting Human Intellect*, pages 263–273. IOS Press, 2023.
- [125] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *International conference on machine learning*, pages 1321–1330, 2017.

- [126] Yarin Gal and Zoubin Ghahramani. Uncertainty in deep learning. *University of Cambridge*, 2016.
- [127] Xuanzhe Yang, Yuming Zhang, Yi Li, and Zixiang Wu. Performance of artificial intelligence in diagnosing lumbar spinal stenosis: a systematic review and meta-analysis. *Spine*, pages 10–1097, 2024.
- [128] Ningjing Chen, Daniel Yee Tak Fong, and Janet Yuen Ha Wong. 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. *Scandinavian Journal of Work, Environment & Health*, 49(7):487, 2023.
- [129] XrayService. X-ray service s.r.l. <https://www.xrayservice.it/>, 2025. Accessed: January 15, 2024.
- [130] Keshav Shree Mudgal and Neelanjan Das. The ethical adoption of artificial intelligence in radiology. *BJR — Open*, 2(1):20190020, 2020.
- [131] Mark D Kovacs, Joseph Mesterhazy, David Avrin, Thomas Urbania, and John Mongan. Correlate: a pacs-and ehr-integrated tool leveraging natural language processing to provide automated clinical follow-up. *Radiographics*, 37(5):1451–1460, 2017.
- [132] Olya Rezaeian, Alparslan Emrah Bayrak, and Onur Asan. Explainability and ai confidence in clinical decision support systems: Effects on trust, diagnostic performance, and cognitive load in breast cancer care. *arXiv preprint arXiv:2501.16693*, 2025.
- [133] Warren B Geftter, Mathias Prokop, Joon Beom Seo, Suhail Raoof, Curtis P Langlotz, and Hiroto Hatabu. Human-ai symbiosis: a path forward to improve chest radiography and the role of radiologists in patient care, 2024.
- [134] Jonathan Herington, Melissa D McCradden, Kathleen Creel, Ronald Boellaard, Elizabeth C Jones, Abhinav K Jha, Arman Rahmim, Peter JH Scott, John J Sunderland, Richard L Wahl, et al. Ethical considerations for artificial intelligence in medical imaging: deployment and governance. *Journal of Nuclear Medicine*, 64(10):1509–1515, 2023.
- [135] Francesca Coppola, Lorenzo Faggioni, Michela Gabelloni, Fabrizio De Vietro, Vincenzo Mendola, Arrigo Cattabriga, Maria Adriana Cocozza, Giulio Vara, Alberto Piccinino, Silvia Lo Monaco, et al. Human, all too human? an all-around appraisal of the “artificial intelligence revolution” in medical imaging. *Frontiers in psychology*, 12:710982, 2021.
- [136] HIPAA Compliance Assistance. Summary of the hipaa privacy rule. *Office for Civil Rights*, 2003.
- [137] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A practical guide, 1st ed.*, Cham: Springer International Publishing, 10(3152676):10–5555, 2017.
- [138] EU Artificial Intelligence Act. The eu artificial intelligence act. *European Union*, 2024.
- [139] Kevin Donnelly et al. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279, 2006.
- [140] Clement J McDonald, Stanley M Huff, Jeffrey G Suico, Gilbert Hill, Dennis Leavelle, Raymond Aller, Arden Forrey, Kathy Mercer, Georges DeMoor, John Hook, et al. Loinc, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry*, 49(4):624–633, 2003.
- [141] Curtis P Langlotz. Radlex: a new method for indexing online educational materials, 2006.
- [142] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- [143] Yubo Shi, Weichun Guo, and Ling Yu. Non-specific low back pain: interpretation of north american spine society (nass) guidelines for evidence-based medicine. *Zhongguo xiu fu Chong Jian wai ke za zhi= Zhongguo Xiu fu Chongjian Waike Zazhi= Chinese Journal of Reparative and Reconstructive Surgery*, 35(10):1336–1340, 2021.
- [144] Kieran O’Sullivan, Mary O’Keeffe, and Peter O’Sullivan. Nice low back pain guidelines: opportunities and obstacles to change practice, 2017.
- [145] Kilem L Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

BIBLIOGRAPHY

- [146] Justin Q Ly. Systematic approach to interpretation of the lumbar spine mr imaging examination. *Magnetic Resonance Imaging Clinics of North America*, 15(2):155–166, 2007.
- [147] Nathan Hannemann and Liem T Bui-Mansfield. The “abcde” approach to the systematic assessment of lumbar spine mr examination. *Contemporary Diagnostic Radiology*, 41(24):1–7, 2018.
- [148] W Brinjikji, FE Diehn, JG Jarvik, CM Carr, DF Kallmes, MH Murad, and PH Luetmer. Mri findings of disc degeneration are more prevalent in adults with low back pain than in asymptomatic controls: a systematic review and meta-analysis. *American Journal of Neuroradiology*, 36(12):2394–2399, 2015.
- [149] Linda Smith and Chen Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008.
- [150] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.
- [151] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.
- [152] Tyler J Richards, Adam E Flanders, Errol Colak, Luciano M Prevedello, Robyn L Ball, Felipe Kitamura, John Mongan, Maryam Vazirabad, Hui-Ming Lin, Anne Kendell, et al. The rsna lumbar degenerative imaging spine classification (lumbardisc) dataset. *arXiv preprint arXiv:2506.09162*, 2025.
- [153] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer methods and programs in biomedicine*, 208:106236, 2021.
- [154] Bernard Rosner, Robert J Glynn, and Mei-Ling T Lee. The wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*, 62(1):185–192, 2006.
- [155] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.
- [156] David F Fardon, Alan L Williams, Edward J Dohring, F Reed Murtagh, Stephen L Gabriel Rothman, and Gordon K Sze. Lumbar disc nomenclature: version 2.0: Recommendations of the combined task forces of the north american spine society, the american society of spine radiology and the american society of neuroradiology. *The spine journal*, 14(11):2525–2545, 2014.
- [157] Alessandro Trento, Salvatore Rapisarda, Nicola Bresolin, Andrea Valenti, and Enrico Giordan. Artificial intelligence and its impact on the management of lumbar degenerative pathology: a narrative review. *Medicina*, 61(8):1400, 2025.
- [158] Nicholas Shepard, Mohammad Samim, Yong Kim, and Afshin Razi. A practical approach to spine magnetic resonance imaging. *JBJS reviews*, 8(3):e0099, 2020.
- [159] Farid Ghareh Mohammadi and Ronnie Sebro. Artificial intelligence impact on burnout in radiologists—alleviation or exacerbation? *JAMA Network Open*, 7(11):e2448720–e2448720, 2024.
- [160] Suresh Pavuluri, Rohit Sangal, John Sather, and R Andrew Taylor. Balancing act: the complex role of artificial intelligence in addressing burnout and healthcare workforce dynamics. *BMJ Health & Care Informatics*, 31(1):e101120, 2024.
- [161] Mohammad Ennab and Hamid Mcheick. Enhancing interpretability and accuracy of ai models in healthcare: a comprehensive review on challenges and future directions. *Frontiers in Robotics and AI*, 11:1444763, 2024.
- [162] Veronika Cheplygina, Marleen De Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.
- [163] Aleksandr Minin, Olga Leonova, Aleksandr Krutko, Elizaveta Elgaeva, Denis Antonets, Dmitriy Shtokalo, and Yakov Tsepilov. Spinescan: a deep learning model for lumbar spine mri annotation and pfirrmann grading assessment. *European Spine Journal*, pages 1–9, 2025.

- [164] Yonghong Liu, Javed Rashid, Salah Mahmoud Boulaaras, Muhammad Shoaib Saleem, and Muhammad Faheem. Automated lumbar spine degenerative classification using deep learning: A comprehensive evaluation based on rsna 2024. *International Journal of Computational Intelligence Systems*, 18(1):328, 2025.
- [165] Hua-Dong Zheng, Yue-Li Sun, De-Wei Kong, Meng-Chen Yin, Jiang Chen, Yong-Peng Lin, Xue-Feng Ma, Hong-Shen Wang, Guang-Jie Yuan, Min Yao, et al. Deep learning-based high-accuracy quantitation for lumbar intervertebral disc degeneration from mri. *Nature communications*, 13(1):841, 2022.
- [166] Weicong Zhang, Ziyang Chen, Zhihai Su, Zhengyan Wang, Jinjin Hai, Chengjie Huang, Yuhan Wang, Bin Yan, and Hai Lu. Deep learning-based detection and classification of lumbar disc herniation on magnetic resonance images. *JOR spine*, 6(3):e1276, 2023.
- [167] Robin Y Park, Rhydian Windsor, Amir Jamaludin, and Andrew Zisserman. Multi-view and multimodal radiological grading using spinal mris. 2025.
- [168] Arnesh Batra, Arush Gumber, and Anushk Kumar. M-scan: A multistage framework for lumbar spinal canal stenosis grading using multi-view cross attention. *arXiv preprint arXiv:2503.01634*, 2025.
- [169] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- [170] Craig T Hartrick, Juliann P Kovan, and Sharon Shapiro. The numeric rating scale for clinical pain measurement: a ratio measure? *Pain Practice*, 3(4):310–316, 2003.
- [171] Mark P Jensen, Paul Karoly, and Sandra Braver. Measurement of pain. *Pain assessment and management: An applied approach*, pages 45–88, 2001.
- [172] Martin Roland and Richard Morris. A study of the natural history of back pain: part i: development of a reliable and sensitive measure of disability in low-back pain. *spine*, 8(2): 141–144, 1983.
- [173] Giovanni De Girolamo, Paola Rucci, Paolo Scocco, Angela Becchi, Francesco Coppa, Angela D’Addario, Elisabetta Daru, Diego De Leo, Lorella Galassi, Lara Mangelli, et al. Quality of life assessment: validation of the italian version of the whoqol-brief. *Epidemiology and Psychiatric Sciences*, 9(1):45–55, 2000.
- [174] Whoqol Group et al. Development of the world health organization whoqol-bref quality of life assessment. *Psychological medicine*, 28(3):551–558, 1998.
- [175] Kenneth Lyne and Derek Roger. A psychometric re-assessment of the cope questionnaire. *Personality and Individual Differences*, 29(2):321–335, 2000.
- [176] Chiara Foà, Annalisa Tonarelli, Luca Caricati, and Laura Fruggeri. Cope-nvi-25: validazione italiana della versione ridotta della coping orientation to the problems experienced (cope-nvi). *Psicologia della salute: quadrimestrale di psicologia e scienze della salute*: 2, 2015, pages 123–140, 2015.
- [177] Anna Cabak, Anna D’browska-Zimakowska, Aleksandra Truszczyńska, Patryk Rogala, Katarzyna Laprus, and Wiesław Tomaszewski. Strategies for coping with chronic lower back pain in patients with long physiotherapy wait time. *Medical science monitor: international medical journal of experimental and clinical research*, 21:3913, 2015.
- [178] Jennifer A Haythornthwaite, Lynette A Menefee, Leslie J Heinberg, and Michael R Clark. Pain coping strategies predict perceived control over pain. *Pain*, 77(1):33–39, 1998.
- [179] Javier Martinez-Calderon, Carmen Zamora-Campos, Santiago Navarro-Ledesma, and Alejandro Luque-Suarez. The role of self-efficacy on the prognosis of chronic musculoskeletal pain: a systematic review. *The Journal of Pain*, 19(1):10–34, 2018.
- [180] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.
- [181] Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.
- [182] James MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1

BIBLIOGRAPHY

- (14):281–297, 1967.
- [183] Joe H Ward. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.
 - [184] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 2:224–227, 1979.
 - [185] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.
 - [186] Samuel S Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.
 - [187] Howard Levene. Robust tests for equality of variances. *Contributions to probability and statistics*, pages 278–292, 1960.
 - [188] Bernard L Welch. The generalization of ‘student’s’ problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.
 - [189] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 18(1):50–60, 1947.
 - [190] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.
 - [191] Ronald A Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.
 - [192] Carlo E Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.
 - [193] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.
 - [194] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 1988.
 - [195] Harald Cramér. *Mathematical methods of statistics*. Princeton University Press, 1946.
 - [196] Andrew F Hayes. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications, 2017.
 - [197] Alaa Tharwat, Tarek Gaber, Abdelhameed Ibrahim, and Aboul Ella Hassanien. Linear discriminant analysis: A detailed tutorial. *AI communications*, 30(2):169–190, 2017.
 - [198] Mazyad A Alotaibi, Ahmed S Alhowimel, Faris A Alodaibi, Dalia Alimam, Mshari Alghadier, and Ragab K Elnaggar. Fear-avoidance beliefs increase the perception of pain and disability in saudi arabian patients with chronic low back pain. *Medicine*, 104(30):e43510, 2025.
 - [199] Helena Widdrington, Charlotte Krahé, Katie Herron, Kimberley Smith, and Mary Gemma Cherry. Thriving when living with chronic pain: A qualitative evidence synthesis of individuals’ experiences. *British journal of health psychology*, 30(3):e70000, 2025.
 - [200] Roni Chaaya, Michel Sfeir, Samar El Khoury, Sandrella Bou Malhab, and Myriam El Khoury-Malhame. Adaptive versus maladaptive coping strategies: insight from lebanese young adults navigating multiple crises. *BMC Public Health*, 25(1):1464, 2025.
 - [201] Sotaro Shimada, Ardith Z Doorenbos, Ellen Goldstein, and Dahee Wi. A systematic review of pain catastrophizing and chronic musculoskeletal pain. *Pain Management Nursing*, 2025.
 - [202] Akshit Soota, Angela C Incollingo Rodriguez, Benjamin C Nephew, Paula Gardiner, Jean A King, Natalia Morone, and Carolina Ruiz. Machine learning based phenotyping of the response to mindfulness for chronic low back pain. *medRxiv*, pages 2025–08, 2025.
 - [203] Paul Dougherty and Gregory Beehler. Biopsychosocial management of chronic lower back pain: The role of manual therapy and behavioral health (bh) co-management. *Frontiers in Musculoskeletal Disorders*, 3:1646096, 2025.
 - [204] Randy S Roth, Michael E Geisser, and David A Williams. Interventional pain medicine: retreat from the biopsychosocial model of pain. *Translational behavioral medicine*, 2(1):106–116, 2012.

BIBLIOGRAPHY

- [205] Pavan Tankha. Shifting perspectives: How to manage chronic pain with the biopsychosocial model. URL <https://www.capc.org/blog/shifting-perspectives-how-to-manage-chronic-pain-with-the-biopsychosocial-model/>. Blog post. Edited by Melissa Baron. Clinical review by Andrew Esch, MD, MBA.
- [206] Lena Danielsson, Svein Bergvik, Are Hugo Pripp, and Gunnvald Kvarstein. The efficacy of a clustered group-based acceptance and commitment therapy for patients with chronic pain—a randomized controlled semi-crossover trial. *Canadian Journal of Pain*, 9(1): 2515106, 2025.
- [207] Jae QJ Liu, Yim Wah Mak, Aled LY Tang, Crystal Kwan, Fadi Al Zoubi, Timmy KT Wong, Gordon SH Tsang, Heidi CW Kwong, Sabrina WT Lai, Sam PS Sze, et al. Effects of acceptance and commitment therapy plus exercise for older adults with chronic low back pain: A preliminary cluster randomized controlled trial with qualitative interviews. *The Journal of Pain*, 30:105350, 2025.
- [208] Denny Borsboom. A network theory of mental disorders. *World psychiatry*, 16(1):5–13, 2017.