THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學
Pao Yue-kong Library
包玉剛圖書館

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

Pao Yue-kong Library, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

http://www.lib.polyu.edu.hk

RADIOMIC SIGNATURES FROM PLAIN RADIOGRAPHS PREDICT KNEE OSTEOARTHRITIS PROGRESSION: FROM SINGLE- TO MULTI-VIEW ANALYSIS

JIANG TIANSHU

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Biomedical Engineering

Radiomic Signatures from Plain Radiographs Predict Knee Osteoarthritis Progression: From Single- to Multi-View Analysis

JIANG Tianshu

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

July 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

JIANG Tianshu (Name of student)

Abstract

Background: Knee osteoarthritis (OA) is a prevalent joint disorder characterised by progressive structural degeneration and heterogeneous disease trajectories. It involves both the tibiofemoral (TF) and patellofemoral (PF) compartments, often with uneven severity, creating challenges for consistent diagnosis and personalised management. Traditional radiographic assessments, such as the Kellgren–Lawrence (KL) grading system, show limited sensitivity, particularly for compartment-specific or early-stage OA. The KL system also suffers from substantial interobserver variability, weak correlation with clinical outcomes, and a non-linear categorical scale that restricts quantitative interpretation. Given the need for scalable, cost-effective, and objective tools, radiomics—the quantitative analysis of imaging features—offers a promising avenue for personalised OA risk prediction. This thesis develops and validates radiomics-based biomarkers for assessing knee OA severity and progression, focusing on compartment-specific modelling and multi-view feature integration.

Methods: Data were derived from multiple large, multicentre knee OA cohorts, including the Multicenter Osteoarthritis Study (MOST), the Osteoarthritis Initiative (OAI), the Hong Kong EHR-derived Knee OA Cohort, the Meniscal Tear and Osteoarthritis Risk (MenTOR) Cohort, and the Knee Injury Cohort at the Kennedy (KICK). All provided standardised radiographs and longitudinal clinical outcomes. (i) Radiomic features were first extracted from the PF joint using semi-automated segmentation and advanced image-processing techniques to build a predictive model for future knee replacement. (ii) To enhance cross-cohort generalisability, a domain adaptation strategy was implemented to harmonise differences in imaging protocols and population characteristics. (iii) A deep-learning-based radiomics framework was then

developed for the TF joint, employing convolutional neural networks to capture complex structural patterns beyond those represented by KL grading. (iv) Finally, a multi-view learning approach combined PF and TF radiomics features to evaluate whether integrated compartmental information improved prediction of OA progression.

Results: Four principal findings were obtained. (i) The PF radiomics score demonstrated additive predictive value to the KL grade, with their combination achieving an area under the receiver operating characteristic curve (AUC) of 0.87 versus 0.84 for KL alone ($p < 0.001$). (ii) The domain-adapted PF model showed strong external generalisability, yielding AUCs of 0.73 (US), 0.70 (Hong Kong), and 0.64 (UK), surpassing alternative models. (iii) The deep-learning TF radiomics framework improved assessment of OA severity and progression, reaching a concordance index (C-index) of 0.85 and an AUC of 0.89, indicating enhanced sensitivity to structural change. (iv) The integrated PF–TF radiomics model further improved predictive accuracy, achieving a C-index of 0.91 and an AUC of 0.93, underscoring the importance of comprehensive, compartment-aware radiographic analysis.

Conclusion: Radiomics analysis of the PF and TF joints offers a powerful, cost-effective, and scalable imaging-based tool for personalised knee OA risk prediction. Through advanced feature extraction, deep learning, and domain adaptation, the proposed models show robust generalisability and clinical relevance. The integration of multi-compartmental radiomic features further enhances predictive precision, enabling comprehensive assessment of disease status. Overall, this thesis refines current radiographic evaluation methods and establishes a methodological framework that may support future precision-medicine approaches, promoting earlier detection and more individualised management of knee OA.

Publications arising from the thesis

Jiang T[#], Lau SH[#], Zhang J, Chan LC, Wang W, Chan PK, Cai J, Wen C[*]. Radiomics signature of osteoarthritis: Current status and perspective. Journal of Orthopaedic Translation. 2024;45:100-6. (Journal article, Q1, IF = 7.8, Chapter 1)

Zhang J[#], **Jiang T**[#], Chan L-C, Lau S-H, Wang W, Teng X, Chan P-K, Cai J, Wen C[*]. Radiomics analysis of patellofemoral joint improves knee replacement risk prediction: Data from the Multicenter Osteoarthritis Study (MOST). Osteoarthritis and Cartilage Open. 2024;6(2):100448. (Journal article, Q1, IF = 2.8, Chapter 3)

Jiang T, Chan LC, Chan PK, Wen C. Predicting Knee Replacement Surgery Risk Using PF Joint RadScore: Cross-Dataset Analysis from the US to Hong Kong. HKOA Annual Congress 2024. (Conference Presentation, Chapter 4)

Jiang T, Zhang Y, Zhang J, Wang W, Chan LC, Chan PK, Hunter DJ, Vincent TL, Cai J, Tan KC, Watt FE[*], Huang Z[*], Wen C[*]. Generalised patellofemoral radiomics score for progressive knee osteoarthritis across multicontinental cohorts (Journal article, npj Digital Medicine, Under Review, Chapter 4)

Jiang T, Chan LC, Chan PK, Wen C. Deep-Learning Radiomics for Survival Analysis on Anticipated Knee Arthroplasty Using Posteroanterior View X-Rays. HKOA Annual Congress 2023. (Conference Presentation, Chapter 5)

Jiang T, Chan LC, Wen C. AI-Powered Radiomics for Predicting Knee Osteoarthritis Progression: Towards Sustainable Patient Management. PolyU Research Student Conference 2025. (Conference Presentation, Best Poster Runner-Up, Chapter 5)

Jiang T, Chan LC, Chan PK, Wen C. Predicting Knee Replacement Surgery Using Deep-learning-Based Radiomics Analysis of Plain Radiographs: Insights from Multicentre Cohort Studies. OARSI 2024 World Congress on Osteoarthritis. (Conference Presentation, Chapter 5)

Jiang T[#], Chan LC[#], Wang W, Lau SH, Chan PK, Chen H, Hunter DJ, Vincent TL, Watt FE[*], Wen C[*]. Moving beyond the Kellgren–Lawrence Grade: A Deep-Learning-Based Radiomics Score Assessing Osteoarthritis Severity and Progression (Journal article, The Lancet Rheumatology, Submitted, Chapter 5)

Note: Corresponding author[*], Co-first author[#]

Other publications during the PhD

Journal articles

Wang W[#], **Jiang T[#]**, Zhang J, Liu J, Chan LC, Lin M, Li J, Ding C, Chiu KY, Fu H, Chan PK, Wen C. Subchondral bone expansion in advanced knee osteoarthritis: Relation with radiographic severity and role in surgical decision-making. Osteoarthritis and Cartilage Open. 2024;6(2):100461. (Q1, IF = 2.8)

Zhao J, **Jiang T**, Lin Y, Chan LC, Chan PK, Wen C, Chen H[*]. Adaptive Fusion of Deep Learning with Statistical Anatomical Knowledge for Robust Patella Segmentation from CT Images. IEEE Journal of Biomedical and Health Informatics. 2024;28(5):2842-53. (Q1, IF = 6.8)

Au MT[#], Ni J[#], Tang K[#], Wang W, Zhang L, Wang H, Zhao F, Li Z, Luo P, Lau LCM, Chan PK, Luo C, Zhou B, Zhu L, Zhang CY, **Jiang T**, Lauwers M, Chan JFW, Yuan S[*], Wen C[*]. Blockade of endothelin receptors mitigates SARS-CoV-2-induced osteoarthritis. Nature Microbiology. 2024;10;2538-52. (Q1, IF = 19.4)

Chong YY, Lau CML[*], **Jiang T**, Wen C, Zhang J, Cheung A, Luk MH, Leung KCT, Cheung MH, Fu H, Chiu KY, Chan PK[*]. Predicting periprosthetic joint infection in primary total knee arthroplasty: a machine learning model integrating preoperative and perioperative risk factors. BMC Musculoskeletal Disorders. 2025;26(241). (Q2, IF = 2.4)

Lau SH, Chan LC, **Jiang T**, Zhang J, Meng X, Wang W, Chan PK, Cai J, Li P, Wen C[*]. Diffusion model-empowered patella shape analysis predicts knee osteoarthritis

outcomes. Osteoarthritis and Cartilage Open. 2025;7(4):100663. (Q1, IF = 2.8)

Conference presentations

Jiang T, Zhao J, Lin Y, Chan LC, Chan PK, Wen C, Chen H. Adaptive Fusion of Deep Learning with Statistical Shape Model for Robust Patella Segmentation from CT Images. HKOA Annual Congress 2022.

Jiang T, Chan LC, Chan PK, Wen C. Topological Data Analysis On 3-Dimensional Patella Shape as New Biomarker for Knee Arthroplasty Type Classification. OARSI 2023 World Congress on Osteoarthritis.

Wang W, **Jiang T**, Zhang J, Chan LC, Chan PK, Wen C. Tibial Plateau Surface Area in Surgical Decision for Osteoarthritic Knees. OARSI 2023 World Congress on Osteoarthritis.

Note: Corresponding author*, Co-first author#

Acknowledgements

First and foremost, I would like to express my deepest gratitude to **Prof. WEN Chunyi**, my PhD supervisor and mentor. I am profoundly thankful that Professor Wen recognised my potential and recruited me into his research group during the challenging period of the pandemic. His decision provided me with the opportunity to embark on this academic journey despite global uncertainties. He has guided me not only academically with his profound expertise and vision in biomedical engineering and artificial intelligence, but has also shaped my approach to research with his meticulous attention to detail and pursuit of scientific excellence. I am truly thankful for his detailed guidance; he has offered valuable insights not only on scientific research but also on life and work, which has greatly benefited me. His patience, encouragement, and unwavering support throughout my doctoral journey have been invaluable. He has been a constant source of inspiration, consistently challenging me to think critically, refine my ideas, and push the boundaries of my work. I am truly fortunate to have had the opportunity to work under his guidance and to learn from his vast experience and wisdom.

I am profoundly thankful to my girlfriend, **Dr. ZHANG Ruotong**, for her love, patience, and understanding. Her unwavering emotional support has sustained me through moments of doubt and fatigue. She has always believed in me, encouraged me to persevere, and brought joy and balance into my life.

My sincere appreciation goes to **Dr. CHAN Lok Chun**, who has not only been a remarkable senior colleague but also the spirit of our AI team. His diligence, intelligence, and dedication to scientific rigour have been truly inspiring. I deeply

admire his ability to navigate complex challenges and his generosity in sharing knowledge and advice. I am also immensely grateful to **Dr. ZHANG Jiang**, a key collaborator and master of radiomics, whose technical insights and collaborative spirit have significantly enriched this research. His expertise and willingness to help have been instrumental to my progress.

I wish to acknowledge **Dr. CHAN Ping-Keung** for providing essential clinical perspectives that strengthened the translational relevance of my work. My heartfelt thanks also go to **Prof. Tonia Vincent** and **Dr. Fiona Watt**, whose support during my exchange at Oxford offered invaluable opportunities for academic growth, collaboration, and broader scientific engagement.

I am also grateful to my main co-authors, **WANG Wei** and **LAU Sing-Hin**, for their substantial contributions, insightful discussions, and teamwork that made the collaborative aspects of this research both productive and enjoyable.

To my colleagues and friends, including **Dr. AU Man-Ting, Dr. NI Junguo, Dr. ZHANG Yuqi, LIU Jun, LI Zhan, YAN Jin, LI Ho-Hin Toby, HSIEH He-Yi, WANG Hantang, ZHANG Charlie Yuli, and ZHANG Alex Yuning**, thank you for the camaraderie, lively discussions, and moments of laughter that have made this PhD journey memorable. Your friendship and support have been invaluable, both inside and outside the lab.

I would also like to extend my thanks to many other colleagues and friends whose names may not be listed here individually but whose encouragement, assistance, and shared experiences have contributed significantly to my growth and the completion of this work.

I would like to extend my sincere gratitude to the providers and contributors of the valuable cohorts used in this research, including the **Multicenter Osteoarthritis Study (MOST) Team**, the **Osteoarthritis Initiative (OAI) Team**, the **Hong Kong Hospital Authority Data Collaboration Laboratory (HADCL) Team**, the **MenTOR Cohort Team**, and the **KICK Cohort Team**. Without the availability of these meticulously collected datasets and the dedication of the researchers and participants involved, this work would not have been possible.

Finally, I extend my deepest gratitude to my family for their unconditional love, patience, and steadfast belief in me. Their constant encouragement and understanding have been my greatest strength throughout this long journey. Without their support, none of this would have been possible.

Table of contents

List of figures

List of tables

List of abbreviations

OA	Osteoarthritis
KR	Knee Replacement
KL	Kellgren-Lawrence
MRI	Magnetic Resonance Imaging
CT	Computed Tomography
TF	Tibiofemoral
PF	Patellofemoral
IBSI	Image Biomarker Standardisation Initiative
WORMS	Whole-Organ Magnetic Resonance Imaging Score
MOAKS	MRI Osteoarthritis Knee Score
DEXA	Dual-energy X-ray Absorptiometry
PET	Positron Emission Tomography
ROI	Regions of Interest
VOI	Volume of Interest
CNN	Convolutional Neural Network
TMJ	Temporomandibular Joint
SpA	Spondylarthritis
ACLR	Anterior Cruciate Ligament Reconstruction
MOST	Multicenter Osteoarthritis Study
OAI	Osteoarthritis Initiative
EHR	Electronic Health Record
MenTOR	Meniscal Tear and Osteoarthritis Risk
KICK	Knee Injury Cohort at the Kennedy

IRB	Institutional Review Board
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
HK	Hong Kong
HADCL	Hospital Authority Data Collaboration Laboratory
BMI	Body Mass Index
PFOA	Patellofemoral Osteoarthritis
CVAT	Computer Vision Annotation Tool
RadScore	Radiomics Score
mRMR	minimum Redundancy Maximum Relevance
HR	Hazard Ratio
ROC	Receiver Operating Characteristic
AUC	Area Under the Receiver Operating Characteristic Curve
C-index	Concordance index
CI	Confidence Interval
KM	Kaplan–Meier
vKR	virtual Knee Replacement
KOOS	Knee Injury and Osteoarthritis Outcome Score
GPR Score	Generalised Patellofemoral Radiomics Score
PA	Posterior–anterior
Faster R-CNN	Faster Region-based Convolutional Neural Network
MIL	Multiple Instance Learning
DR Score	Deep-learning-based Radiomics Score
PACS	Picture Archiving and Communication Systems

Chapter 1: Introduction and Literature Review

1.1 Background of Knee Osteoarthritis (OA)

Osteoarthritis (OA) is one of the most common chronic musculoskeletal conditions and a leading cause of pain and disability among older adults worldwide (1). It is a progressive degenerative joint disease primarily affecting load-bearing joints, with the knee being the most commonly involved (2, 3), as shown Figure 1-1. Driven by demographic shifts such as population ageing and the global rise in obesity, the prevalence of knee OA has sharply increased in recent decades (4, 5). Epidemiological data indicate that more than 500 million people globally are affected, with substantial implications for individual well-being and public healthcare systems (6, 7).

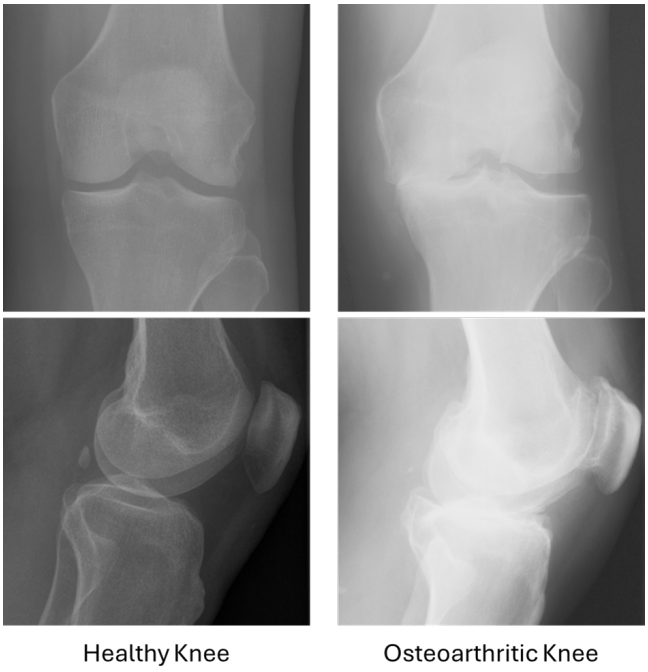


Figure 1-1 Comparison of Normal Knee and Osteoarthritic Knee

Clinically, knee OA manifests as joint pain, stiffness, reduced mobility, and functional

impairment, often leading to a diminished quality of life (8). From a public health perspective, knee OA imposes a significant socioeconomic burden, not only due to the direct costs of medical care—including surgeries such as knee replacement (KR)—but also due to the indirect costs related to lost productivity, long-term disability, and caregiving (9).

Despite its high prevalence, early diagnosis and precise characterisation of knee OA remain challenging (10). Conventional clinical diagnosis of knee OA is heavily reliant on a combination of patient-reported symptoms, physical examination findings, and standard imaging, primarily plain radiography. Radiographs are used to assess structural changes such as joint space narrowing, osteophyte formation, subchondral sclerosis, and cysts—findings that are typically graded using systems like the Kellgren-Lawrence (KL) grading. However, these radiographic changes generally appear in the later stages of the disease, limiting their utility for early detection or risk stratification (10, 11).

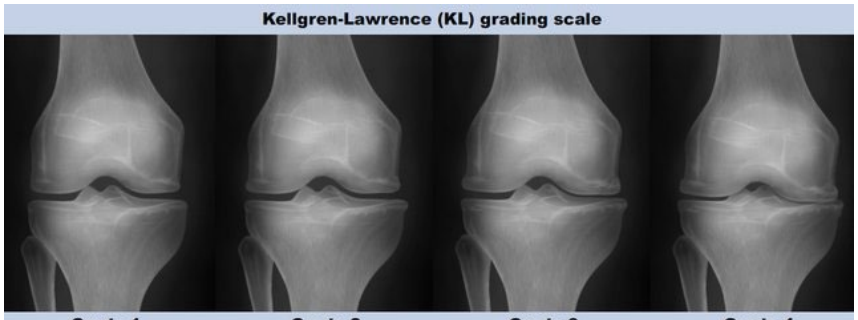
Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) offer more detailed visualisations of joint structures, including cartilage, bone, and soft tissues (12). However, practical challenges such as high cost, limited availability, and exposure to radiation (in the case of CT) restrict their widespread use in routine screening and follow-up (13).

Given the lack of sensitive imaging biomarkers capable of detecting early disease or predicting its trajectory, there is a pressing need for innovative image analysis techniques that can extract more meaningful and predictive information from standard imaging modalities—particularly those that are fast, low-cost, and widely accessible for routine screening and longitudinal follow-up (14).

1.2 Clinical and Radiological Assessment of Knee OA

From a clinical standpoint, the assessment of OA often involves a combination of patient-reported symptoms, physical examination, and imaging. Among these, radiographic imaging has long been the mainstay in both clinical and research settings, owing to its wide availability, low cost, and standardised interpretation protocols (14).

The KL grading system is the most widely adopted radiographic classification for knee OA (Figure 1-2). First introduced in 1957, it categorises disease severity into five grades, ranging from grade 0 (normal) to grade 4 (severe OA), based on visual identification of joint space narrowing, osteophyte formation, subchondral sclerosis, and bony deformity (15).



CLASSIFICATION	Normal	Doubtful	Mild	Moderate	Severe
DESCRIPTION	No features of OA	Minute osteophyte: doubtful significance	Definite osteophyte: normal joint space	Moderate joint space reduction	Joint space greatly reduced: subchondral sclerosis

Figure 1-2 The KL grading system to assess the severity of knee OA (16).

Despite its widespread use, the KL system presents several notable limitations. It lacks sensitivity in detecting early-stage disease, as the radiographic changes it relies upon often manifest only in more advanced stages. Consequently, subtle degenerative changes may go unnoticed, delaying the opportunity for early intervention (17, 18). Another significant shortcoming lies in its generalised scoring across the entire joint,

with no ability to differentiate between structural changes in the tibiofemoral and patellofemoral compartments. Given the compartment-specific nature of OA progression in many patients, this limitation reduces the clinical granularity required for personalised treatment planning.

Furthermore, KL grading is inherently subjective. Its reliance on visual interpretation introduces a high degree of inter- and intra-observer variability, particularly in borderline or intermediate grades (19). Compounding this issue is the known discordance between radiographic severity and clinical symptoms. Patients with similar KL grades can report widely differing levels of pain and functional impairment, while some individuals with severe radiographic changes may remain asymptomatic. These inconsistencies reduce the predictive value of KL grading for disease progression and treatment needs, highlighting the need for more robust and sensitive imaging biomarkers.

To address these limitations, advancements in quantitative imaging techniques have been introduced (Table 1-1). MRI, for example, allows for detailed visualisation of cartilage, bone marrow, menisci, and synovial tissue. Semi-quantitative MRI scoring systems, such as the Whole-Organ Magnetic Resonance Imaging Score (WORMS) (20) and the MRI Osteoarthritis Knee Score (MOAKS) (21), have been developed to assess the status of multiple joint tissues comprehensively. These systems provide more detailed information than conventional radiographs and can identify early pathological changes. However, their application in routine clinical practice remains limited due to time-consuming interpretation, high cost, and the need for expert radiological input (13).

CT offers excellent spatial resolution for bone but provides limited soft tissue contrast and involves radiation exposure, which restricts its use in early disease detection or

longitudinal follow-up (22). Ultrasound is another imaging modality capable of assessing superficial soft tissues and synovial inflammation, though it is highly operator-dependent and less informative for deep joint structures (23, 24). Other advanced imaging techniques, such as dual-energy X-ray absorptiometry (DEXA), positron emission tomography (PET), and hybrid imaging modalities, have shown promise in research settings but have not yet achieved widespread clinical adoption for OA.

Table 1-1 Comparing medical imaging technologies

Type	X-ray	CT	MRI	Ultrasound
Advantages	• Fast • Cheap • Low radiation	• Fast • Medium price • detailed images in 3D	• More detailed images • No radiation	• Real-time • Relatively Cheap • No radiation
Disadvantages	• less detail in 2D	• High radiation	• Very expensive • Slow	• Low quality • Dependent on technician's skill

Abbreviations: OA = Osteoarthritis, CT = Computed Tomography, MRI = Magnetic Resonance Imaging, 3D = Three-Dimensional, 2D = Two-Dimensional

In parallel with these hardware-based advancements, methodological innovations in image analysis have emerged. Techniques such as statistical shape modelling (25), bone texture analysis (26-28), and automated segmentation (29, 30) have enabled more detailed assessments of joint morphology. These methods mark a shift from qualitative to quantitative interpretation of images.

Among them, radiomics has gained particular attention for its ability to extract a large number of quantitative features that describe the intensity, shape, and texture of joint

structures from medical images. Radiomics transforms medical images into structured, high-dimensional data that can be mined using machine learning algorithms to uncover patterns associated with disease severity, progression, and treatment response.

The emergence of radiomics reflects a growing recognition of the limitations of traditional radiographic grading and a broader push toward precision medicine. By enabling data-driven, compartment-specific, and scalable image analysis, radiomics offers the potential to bridge the gap between imaging and individualised clinical care in knee OA.

1.3 Radiomics and Deep Learning in Medical Imaging

Radiomics is an emerging field at the intersection of medical imaging, computational analysis, and machine learning (Figure 1-3). It aims to transform standard medical images into a high-dimensional, mineable dataset by extracting a vast number of quantitative features that describe the shape, intensity distribution, and textural patterns of tissues and anatomical structures (31, 32). Unlike traditional image interpretation, which relies on qualitative and often subjective assessments by clinicians, radiomics enables objective and reproducible analysis by identifying imaging biomarkers that may not be visible to the naked eye (33).

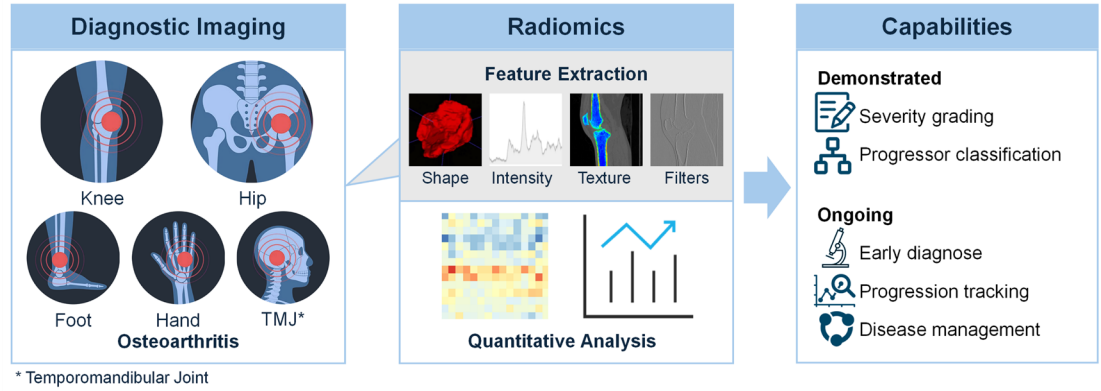


Figure 1-3 Overview of Radiomics Workflow and Capabilities in Osteoarthritis Research (11)

The standard radiomics workflow involves several critical steps: image acquisition, preprocessing, segmentation, feature extraction, feature selection or dimensionality reduction, and predictive modelling. Medical images may be obtained using various modalities, including X-ray, CT, and MRI. Image preprocessing techniques such as resolution standardisation and intensity normalisation are typically applied to mitigate variability arising from scanner settings and acquisition protocols. Segmentation defines the regions or volumes of interest (ROIs or VOIs) from which radiomic features are extracted. This can be performed manually, semi-automatically, or using fully automated algorithms.

The extracted features fall into several categories. First-order features describe the distribution of voxel intensities within a region and are typically derived from histogram-based statistics. Second-order and higher-order features, which include gray-level co-occurrence matrices, run-length matrices, and size zone matrices, capture spatial relationships and textural heterogeneity within tissues. Shape descriptors characterise geometric properties of anatomical structures. Due to the high dimensionality of the resulting feature space, feature selection and dimensionality reduction techniques such as principal component analysis, mutual information filtering,

or recursive feature elimination are used to identify the most informative and non-redundant features. These selected features are then input into machine learning models—such as logistic regression, random forests, support vector machines, or ensemble methods—for classification or prediction tasks.

In parallel with handcrafted radiomics, deep learning has gained significant attention for its potential to automate the feature learning process. Convolutional neural networks (CNNs), in particular, are well-suited for image analysis tasks due to their ability to learn hierarchical representations of image content. Unlike traditional radiomics, where features are predefined and extracted based on expert-designed algorithms, CNNs learn relevant patterns directly from raw image data through optimisation during model training. This data-driven approach allows CNNs to capture complex and abstract visual characteristics, potentially revealing subtle disease signatures that might be overlooked by handcrafted methods. However, deep learning models often require large, labelled datasets and are typically perceived as less interpretable than traditional radiomics, which may limit their acceptance in clinical practice.

In the field of musculoskeletal imaging, and specifically in OA research, the application of radiomics is still in its early stages but is expanding rapidly. Traditional OA diagnosis and progression studies have relied on clinical assessments and visual grading systems such as the Kellgren-Lawrence scale. In contrast, radiomics offers a quantitative and scalable approach that holds promise for detecting early disease, predicting disease trajectory, and evaluating treatment outcomes.

Recent studies have demonstrated the potential of radiomics to enhance the understanding of OA pathology (34-38), as shown in Table 1-2. For instance, radiomics features extracted from radiographs have been used to differentiate between knees with

and without patellofemoral OA, outperforming clinical variables in classification tasks. Radiomics has also been applied to CT and MRI data to study the microstructural integrity of subchondral bone, cartilage, and other joint tissues. Some studies have combined radiomics with clinical and biomechanical data to develop integrative models for OA prediction and management. Furthermore, emerging work has utilised CNN-based models to learn from entire joint images, demonstrating superior performance to traditional radiographic grading in detecting structural severity and predicting long-term outcomes.

Table 1-2 Radiomics applications of various OA joints based on different imaging modalities.

Application	X-ray	CT	MRI
Classification	Knee OA – Healthy vs OA (26-28) Hand OA – Healthy vs OA (39) Knee & Hand OA – Healthy vs OA (40)	TMJ OA – Healthy vs OA (41, 42) TMJ OA – Morphology Variability (43) Hand OA – Young vs Older (44)	SpA – Healthy vs OA (35) Knee OA – Healthy vs ACLR (36) Knee OA – Healthy vs OA (45) (46-48) (37) Ankle OA – Dancer vs Normal Person (49)
Detection	Knee OA – Disease Grade (50) Hand OA – Osteophyte (51) Ankle OA – History of Injury (52)	Ankle OA – Deformity Characteristics (25) (53)	Knee OA – Disease Grade (54)
Prediction	Knee OA –Progression vs Non-progression (26, 55) Knee & Hand OA – Progression vs Non-progression (40) Hip OA – Risk of Incident (56, 57)	TMJ OA – Risk of Incident (34)	Knee OA – Treatment response of Vitamin D (38) Knee OA – High Risk vs Low Risk (54) Knee OA – 1-year Onset (58) Knee OA – Risk of Incident (59)

Abbreviations: OA = Osteoarthritis, TMJ = Temporomandibular Joint, SpA = Spondylarthritis, ACLR = Anterior Cruciate Ligament Reconstruction.

Despite its promise, radiomics in OA research still faces several challenges (11, 33).

Many existing studies have small sample sizes, lack external validation, or are limited to specific imaging protocols. Variability in segmentation methods and image acquisition settings also affects feature reproducibility. Standardisation efforts, such as those led by the IBSI, aim to address these issues and enhance the clinical translatability of radiomic biomarkers.

Overall, radiomics and deep learning are reshaping how medical images are analysed in OA research. By enabling high-throughput, quantitative analysis of joint structures, these technologies offer new opportunities for early diagnosis, personalised risk assessment, and precision medicine approaches in the management of knee OA. As data availability and methodological maturity improve, radiomics is poised to become a key tool in the next generation of OA imaging research.

1.4 Gap Analysis and Conceptual Framework

Despite the growing interest in radiomics and deep learning for musculoskeletal imaging, the translation of these approaches into OA research—particularly knee OA—remains relatively limited. While several studies have demonstrated proof-of-concept applications, the field still faces notable methodological, technical, and clinical gaps that hinder broader adoption and impact.

One of the most pressing limitations is the lack of compartment-specific analysis in conventional OA imaging research. The knee joint comprises anatomically and biomechanically distinct compartments, including the tibiofemoral (TF) and patellofemoral (PF) joints, which may exhibit independent disease patterns. However, most existing studies rely on global grading systems such as the KL grade that do not differentiate between compartments. As a result, disease involvement in the PF joint—

which is prevalent in early OA and significantly contributes to pain and disability—is often under-recognised or overlooked. This oversight reduces the sensitivity of radiographic assessments and limits the opportunity to identify patients who may benefit from compartment-targeted management strategies. While some recent studies have explored PF radiomics separately, they remain few, and further validation is needed.

Another gap relates to generalisability. Radiomics models are often developed and tested within a single cohort, under tightly controlled imaging protocols. This raises concerns about their robustness and clinical applicability across diverse populations and clinical settings. Differences in scanner hardware, acquisition parameters, patient demographics, and disease presentation can introduce variability in radiomic feature distributions. Without external validation or domain adaptation strategies, models may perform well in one dataset but fail to generalise beyond it. Few OA studies to date have addressed this challenge systematically, and this remains a major barrier to clinical translation.

There is also a methodological imbalance in the current literature. Most published radiomics studies in OA have focused on traditional, handcrafted feature extraction, with fewer incorporating deep learning-based models. While handcrafted features offer transparency and interpretability, they are limited in their ability to capture higher-order interactions or complex anatomical patterns. Deep learning approaches, especially CNNs, are capable of learning more abstract and powerful representations from imaging data. However, they require large, annotated datasets and are more difficult to interpret. Comparative studies evaluating both approaches in OA are scarce, and little is known about how they may complement or outperform each other in practical

applications.

Moreover, there is limited research on integrating information across joint compartments. Given the compartmental heterogeneity of OA, analysing the TF and PF joints in isolation may lead to suboptimal models. A multi-view approach that combines radiomic features from both compartments could potentially capture a more complete picture of joint health and disease trajectory. Yet few studies have investigated such integrative strategies, and the potential synergistic value of multi-compartment analysis remains largely unexplored.

Finally, the field lacks standardised pipelines and clinical endpoints. Variations in preprocessing methods, feature extraction tools, and model evaluation criteria make it difficult to compare results across studies or establish best practices. Additionally, many studies use proxy endpoints such as structural grading or binary OA classification, which may not reflect patient-centred outcomes like functional decline or the need for knee replacement surgery.

Taken together, these gaps highlight the need for systematic, multi-cohort, and multi-compartment investigations using both traditional and deep learning radiomics approaches. There is also a clear imperative to develop models that go beyond disease classification to support risk stratification, individualised prognosis, and clinical decision-making. The current thesis addresses these unmet needs by (i) building predictive radiomics models focused on the PF and TF joints, (ii) validating model generalizability across cohorts using domain adaptation, and (iii) exploring multi-view integration to enhance predictive performance. These contributions aim to advance the field toward clinically meaningful, image-based tools for personalised OA management.

This thesis is grounded in the evolving paradigm of precision medicine in knee OA, where the goal is to move beyond global structural assessments toward personalised risk profiling and targeted intervention strategies. Within this paradigm, radiomics provides a novel computational approach to extract quantifiable imaging biomarkers from routine radiographs, enabling a shift from qualitative to data-driven evaluation of joint health. The conceptual framework of this research is designed to operationalise this shift by systematically addressing the methodological gaps identified in prior literature and establishing a logical structure for the four studies included in this thesis.

The framework is organised around three core principles: compartment-specific modelling, generalizability across populations, and multi-view integration. These principles guide the progression from hypothesis generation to model development, validation, and clinical interpretation.

The first principle—compartment-specific modelling—recognises that the knee is not a uniform joint but rather a complex structure composed of the TF and PF compartments, each with distinct patterns of degeneration and clinical relevance. Existing radiographic grading systems, such as the KL grade, provide a global assessment that overlooks compartmental differences. To address this limitation, the first study in this thesis focuses on the PF joint, extracting handcrafted radiomic features to develop a predictive model for knee replacement risk. This study highlights the underexplored prognostic value of the PF compartment and establishes a precedent for targeted radiomics analysis in OA.

Building upon this, the second principle—generalisability—is addressed in the second study, which evaluates the robustness of the PF radiomics model across multiple large-scale OA cohorts with varying demographics and imaging protocols. To overcome

domain shifts and reduce the risk of performance drop-off in external datasets, domain adaptation techniques are employed. This component of the framework emphasises the importance of developing clinically applicable models that maintain stability across real-world imaging settings.

The third principle—multi-view integration—serves as the foundation for the latter part of the thesis. The third study introduces a deep learning-based radiomics framework centred on the TF joint. Unlike handcrafted features, this model leverages convolutional neural networks to automatically learn structural patterns associated with OA severity and progression from raw image data. This approach complements the PF analysis and brings in a second axis of modelling sophistication. The final study synthesises both lines of work by integrating PF and TF radiomics features into a unified multi-view model. This integrative approach is designed to capture the heterogeneity of OA across compartments and provide a more comprehensive prediction of disease trajectory.

Collectively, the conceptual framework illustrates how each study builds upon the previous one to advance the field of OA imaging. Rather than treating radiomics as a static tool, the framework treats it as a flexible, layered methodology that can be adapted to specific anatomical targets, validated across diverse populations, and expanded through integrative modelling. It also underscores the complementary strengths of traditional radiomics and deep learning techniques, showing how both can contribute to a broader vision of image-based phenotyping and risk stratification.

By aligning each study with a specific gap and a guiding principle, the framework ensures methodological coherence and theoretical continuity across the thesis. This structured approach not only facilitates the development of clinically relevant imaging biomarkers but also positions the research to make meaningful contributions to the

personalised management of knee OA.

1.5 Rationale and Significance of the Study

In the context of knee OA, radiomics offers several key advantages (11). First, it enables objective, reproducible, and high-dimensional quantification of joint structures, such as subchondral bone and cartilage, from routine imaging like X-rays and MRIs. Second, radiomics facilitates data-driven modelling, allowing the identification of imaging biomarkers that are predictive of clinical endpoints, including pain severity, function loss, and risk of knee replacement. Third, by incorporating machine learning and deep learning algorithms, radiomics can support automated risk stratification and personalised disease management strategies, moving toward the goals of precision medicine.

Moreover, radiomics can be particularly valuable in early-stage OA detection and progression prediction, areas where traditional imaging methods fall short. By capturing microstructural changes before gross morphological alterations become evident, radiomics may offer the potential for timely intervention, thereby delaying or even preventing irreversible joint damage. It also supports cross-compartmental analysis of the tibiofemoral (TF) and patellofemoral (PF) joints, accounting for OA's heterogeneity.

Recent advancements in computational infrastructure, open-access imaging databases, and radiomics standardisation initiatives (e.g., Image Biomarker Standardisation Initiative [IBSI] (60)) further enhance the feasibility of developing generalizable and clinically translatable radiomic models.

In sum, radiomics represents a cutting-edge approach that could redefine how knee OA is detected, monitored, and managed. This thesis builds on this evolving field, aiming to develop and validate radiomics-based tools that not only outperform traditional imaging assessments but also support individualised patient care by integrating multi-compartmental and multi-view radiographic information.

This thesis contributes to the emerging field of radiomics in musculoskeletal imaging by systematically investigating the utility of quantitative features extracted from plain radiographs—a widely available, low-cost imaging modality. Unlike most prior work that focused on the TF compartment or MRI-based features, this study uniquely emphasises the underexplored PF joint, validates model performance across international cohorts using domain adaptation, and introduces a multi-view strategy that integrates PF and TF compartments for a more holistic assessment of disease risk.

This thesis advances the field of knee OA imaging through four key contributions:

Compartment-Specific Risk Prediction: By focusing on the PF joint, which is often underrepresented in traditional assessments, the first study demonstrates that PF radiomic features extracted from standard radiographs can effectively predict future knee replacement risk. This highlights the potential of targeted, compartment-aware radiomic models to capture clinically meaningful patterns beyond global KL grading.

Cross-Cohort Generalizability: A major barrier to the clinical translation of imaging biomarkers is the lack of validation across diverse populations and imaging protocols. The second study addresses this by applying domain adaptation techniques to improve model robustness across international OA cohorts. This work provides an important step toward developing radiomics models that are deployable in real-world clinical

settings.

Deep Learning for TF Radiomics: The third study introduces a deep learning-based radiomics framework focused on the TF joint. Unlike hand-crafted radiomic features, this approach allows automated learning of complex structural signatures from raw image data. It demonstrates that convolutional neural networks can match or exceed the performance of KL grading, offering a scalable and potentially more sensitive tool for OA severity assessment.

Integrative Multi-View Modelling: Recognising the multifactorial and compartmental nature of OA, the fourth study integrates PF and TF radiomics features into a unified multi-view model. This integrative strategy significantly enhances prediction accuracy and reflects a more holistic approach to disease modelling, consistent with the goals of precision medicine.

Together, these studies demonstrate that radiomics—both handcrafted and deep learning-based—can complement and potentially surpass traditional assessment tools for knee OA. The thesis provides a structured framework for evaluating radiomics models at multiple levels: localised (PF/TF), technical (handcrafted vs. deep learning), and integrated (multi-view), thus offering a roadmap for future research and clinical implementation.

Ultimately, this work contributes toward the long-term goal of developing clinically applicable, data-driven tools for early detection, progression monitoring, and individualised treatment planning in knee OA.

1.6 Aim and Objectives of the Thesis

This thesis aims to explore the value of radiomics features extracted from plain radiographs for predicting knee OA severity and progression. By analysing PF and TF joints separately and in combination, the study investigates how radiomics may complement or enhance traditional radiographic grading systems in different clinical contexts.

Objective 1: To develop and evaluate a radiomics-based model using PF join features extracted from plain radiographs for identifying individuals at higher risk of undergoing KR. This includes assessing whether PF radiomic features offer added predictive value compared to conventional KL grading.

Objective 2: To assess the generalisability of the PF-based radiomics model by applying domain adaptation techniques across multiple international cohorts. This objective focuses on evaluating the robustness and transferability of radiomics models when applied to populations with differing demographics and imaging protocols.

Objective 3: To construct a deep-learning-based radiomics model focused on the TF joint to estimate OA severity and structural progression. The objective is to compare this data-driven model with KL grading and explore its potential to detect structural disease features in a more automated and scalable way.

Objective 4: To investigate the integration of PF and TF radiomic features into a unified, multi-view predictive model and examine whether this combination improves prediction performance for OA progression. The objective is to evaluate whether a compartment-aware, integrative approach provides complementary insights not

captured by single-joint analyses.

1.7 Structure of the thesis

This thesis adopts a thesis-by-publication format, consisting of four peer-reviewed or submitted articles that collectively address the development, validation, and integration of radiomics-based approaches for predicting knee OA progression. Each article represents a distinct but connected component of the overall research framework, contributing to the central aim of advancing quantitative imaging biomarkers for compartment-specific and comprehensive OA assessment.

The structure of the thesis is as follows:

Chapter 1: Introduction and Literature Review

Introduces the clinical background of knee osteoarthritis, reviews current imaging-based assessment methods, and highlights the evolution of radiomics in OA research. It identifies key research gaps, outlines the rationale and significance of the study, states the aims and objectives, and presents the thesis structure.

Chapter 2: Study Populations and Data Sources

This chapter provides a detailed description of the five multicentre cohorts used throughout the thesis. It outlines the origin, design, and demographic characteristics of each dataset, including the Multicenter Osteoarthritis Study (MOST), Osteoarthritis Initiative (OAI), Hong Kong EHR-derived Knee OA Cohort, MenTOR Cohort, and KICK Cohort. By consolidating this information in one place, the chapter avoids redundancy in later chapters and offers a clear reference for understanding the clinical

and imaging data sources that underpin the thesis.

Chapter 3 to Chapter 6: Research Articles

Each chapter presents one full-length article, framed by a short preamble to contextualise its role within the broader thesis.

Chapter 3: Radiomics Analysis of the Patellofemoral Joint

Develops a PF-specific radiomics model to predict knee replacement risk, evaluating its performance relative to KL grading.

Chapter 4: Generalisability of the PF Radiomics Model Across Multiple Cohorts

Investigates the robustness and clinical translatability of the PF model through domain adaptation across diverse datasets.

Chapter 5: Deep-Learning Radiomics Analysis of the Tibiofemoral Joint

Explores a CNN-based approach for TF joint analysis, offering an alternative to handcrafted features and expanding on automated feature learning.

Chapter 6: Multi-View Radiomics Integration

Combines PF and TF radiomic features to assess whether joint-compartment integration enhances predictive performance.

Chapter 7: Discussion, Conclusion, and Future Directions

Synthesises the findings from the four studies, discusses cross-cutting themes, evaluates

methodological and clinical implications, and situates the research within the broader field of OA imaging and prediction. Summarises the key contributions of the thesis, discusses its limitations, and outlines potential avenues for further research in radiomics-based OA assessment and beyond.

Appendix and References

Includes consolidated bibliographic references and supplementary materials as necessary for transparency and reproducibility.

Chapter 2: Study Populations and Data Sources

2.1 Overview of Multicentre Cohorts

To ensure the robustness, generalisability, and translational relevance of the proposed radiomics-based models for knee OA, this thesis draws upon five independent cohorts from diverse geographical regions and healthcare systems. These multicentre cohorts encompass a wide range of OA severity, demographic variability, and imaging practices, providing a solid foundation for both model development and external validation.

The rationale for utilising multiple cohorts lies in the heterogeneous nature of knee OA, which is influenced by anatomical, functional, and socio-clinical factors that may vary substantially across populations. A model trained and tested within a single cohort may be vulnerable to cohort-specific biases and lack external validity. By contrast, a multicohort approach facilitates the development of more generalisable models and enables the assessment of domain shift and clinical applicability across different settings.

The datasets used in this thesis include the Multicenter Osteoarthritis Study (MOST) (61) and the Osteoarthritis Initiative (OAI) (62) from the United States; the Hong Kong Electronic Health Record (EHR)-derived Knee OA Cohort from Hong Kong; and two United Kingdom-based cohorts: the Meniscal Tear and Osteoarthritis Risk (MenTOR) cohort and the Knee Injury Cohort at the Kennedy (KICK). These cohorts collectively support a comprehensive investigation of radiomic signatures from different knee compartments, facilitate domain adaptation across imaging and population differences, and allow multi-level validation of prediction models.

The following sections introduce each cohort in detail, including their study design, participant characteristics, imaging protocols, and their specific role within the thesis. A summary at the end of the chapter will illustrate how each dataset contributes to different parts of the research presented.

2.2 The Multicenter Osteoarthritis Study (MOST)

The MOST is a large, prospective, longitudinal cohort study designed to investigate risk factors for the incidence and progression of knee osteoarthritis in older adults. It focuses specifically on individuals aged 50 to 79 years who were at elevated risk for developing knee OA, due to factors such as obesity, prior knee injury, or family history.

Initiated in 2003, the study recruited participants from multiple clinical centres across the United States and conducted regular follow-up assessments over a period extending up to 84 months. Standardised imaging protocols and clinical evaluations were employed to collect comprehensive data on radiographic, symptomatic, and functional outcomes related to knee OA.

The MOST cohort is particularly valuable for radiomics research due to its high-quality, standardised imaging datasets, extensive clinical annotation, and long-term follow-up, enabling robust training and internal validation of predictive models for OA severity and progression. All procedures were conducted with approval from the relevant institutional review boards (IRBs), and informed consent was obtained from all participants. The study is registered at ClinicalTrials.gov (ID: NCT03033238).

2.3 The Osteoarthritis Initiative (OAI)

The OAI is a large-scale, prospective, multicentre observational cohort study initiated

in 2004 to investigate the onset and progression of knee OA in adults aged 45 to 79 years. It enrolled participants with symptomatic knee OA or those at increased risk based on risk factors such as obesity, prior knee injury or surgery, family history, or the presence of hand OA.

Participants were recruited from four clinical sites in the United States: Baltimore (Maryland), Columbus (Ohio), Pittsburgh (Pennsylvania), and Pawtucket (Rhode Island). The cohort includes both radiographically confirmed OA cases and high-risk individuals without radiographic disease at baseline (KL grade 0 in both knees). Annual assessments included questionnaires, clinical examinations, radiographs, and MRI over an 8-year period, with biological samples collected for biomarker analysis.

The OAI dataset provides a robust longitudinal resource for investigating structural, functional, and symptomatic progression in knee OA. It has been widely used for imaging biomarker development, machine learning applications, and evaluation of physical and biochemical predictors of disease trajectory. The study was approved by the institutional review boards at all participating sites, and all participants provided written informed consent. The study adheres to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines and is registered at ClinicalTrials.gov (ID: NCT00080171).

2.4 Hong Kong EHR-derived Knee OA Cohort (HK Cohort)

The Hong Kong (HK) cohort was derived from the territory-wide EHR system managed by the Hospital Authority Data Collaboration Laboratory (HADCL). This database integrates longitudinal medical records from public hospitals and clinics across the region, encompassing approximately 80% of the population (7.49 million people).

Participants included individuals with either a clinical diagnosis of knee osteoarthritis or presenting with knee pain or established OA risk factors. The EHR data provided comprehensive information on patient demographics, clinical diagnoses, procedures, medication prescriptions, and laboratory results, enabling population-level analyses of disease characteristics and progression patterns.

The study was conducted in accordance with the Declaration of Helsinki and received ethical approval from the Institutional Review Board of The Hong Kong Polytechnic University (Ref. No. HSEARS20221121003).

2.5 The Knee Injury Cohort at the Kennedy (KICK)

The KICK is a prospective longitudinal study designed to investigate the long-term outcomes of acute knee injuries. The cohort comprises individuals aged 16 to 50 years who were enrolled within eight weeks of sustaining an acute knee injury. Recruitment occurred between November 1, 2010, and November 28, 2014, across six hospitals and clinics in London, United Kingdom.

Participants were followed for up to five years, enabling the examination of structural, symptomatic, and functional changes over time. The study provides valuable insight into post-traumatic osteoarthritis development in a relatively young population. Ethical approval for the KICK study was granted by the South East London Research Ethics Committee (REC 10/H0706/44).

2.6 The Meniscal Tear and Osteoarthritis Risk (MenTOR) Cohort

The MenTOR cohort is a longitudinal observational study aimed at evaluating the progression of knee osteoarthritis in patients with chronic meniscal pathology. The

study enrolled adults who underwent arthroscopic surgical intervention or knee drainage (arthrocentesis) for symptomatic meniscal tears.

Participants were followed over a five-year period to assess structural and symptomatic outcomes, providing insights into the interplay between meniscal injury, surgical treatment, and OA progression. The cohort is particularly relevant for understanding early-stage joint degeneration in a surgically managed population.

Ethical approval for the MenTOR study was granted by the South Central – Oxford B Research Ethics Committee (REC 15/SC/0551).

2.7 Summary of Cohort Usage Across Chapters

This thesis incorporates five multicentre cohorts to support the development, validation, and generalisation of radiomics-based approaches for knee OA assessment. These cohorts differ in geographic origin, imaging protocols, population demographics, and disease stage, providing a comprehensive foundation for the studies presented. Importantly, the MOST cohort serves as the primary dataset across several chapters because it uniquely provides both PA and lateral radiographs, together with consistent imaging protocols and long-term follow-up—features not available in the other cohorts. In contrast, OAI includes only PA radiographs, and the HK cohort, despite its larger scale, is retrospective with heterogeneous clinical settings and substantial missing data, limiting its suitability for model development. The specific cohort usage in each chapter is summarised below:

Chapter 3 presents the development of a PF joint radiomics model for predicting knee replacement risk. This study exclusively uses the MOST cohort, which provides high-

quality lateral radiographs and long-term follow-up data in an older adult population.

Chapter 4 extends the PF joint radiomics model by applying domain adaptation techniques to improve generalisability across distinct populations and imaging environments. The model is developed on the MOST dataset, refined using the HK cohort, and externally validated using the MenTOR and KICK cohorts from the United Kingdom.

Chapter 5 shifts focus to the TF joint and presents a deep-learning-based radiomics model for evaluating disease severity and progression risk. This chapter integrates data from MOST, OAI, MenTOR, and KICK cohorts to support model training and validation.

Chapter 6 explores a multi-view radiomics approach, integrating features from both the PF and TF joint to enhance the prediction of knee OA progression. This study utilises imaging data from MOST, MenTOR, and KICK cohorts, allowing for comparative analysis of clinical and radiomics-based predictors across multiple compartments and populations.

This cohort structure was designed to maximise both the depth and breadth of radiomics evaluation, ensuring robustness, clinical relevance, and international generalisability of the proposed models.

Chapter 3: Radiomics Analysis of the Patellofemoral Joint: A Predictive Model for Knee Replacement Surgery Risk

3.1 Chapter overview

Knee OA is a complex, heterogeneous disease that does not affect the joint uniformly. Among the various anatomical compartments of the knee, the PF joint plays a crucial role in early structural deterioration and symptom development, particularly in anterior knee pain and reduced mobility. Yet, despite the increasing recognition of its clinical relevance, the PF joint remains underrepresented in imaging-based OA research. Traditional radiographic assessment tools, including the widely used KL grading system, offer a global score for the knee and do not differentiate between the TF and PF compartments. This lack of compartmental granularity can lead to misclassification of disease severity and may obscure important predictive markers specific to the PF region.

The clinical implications of this gap are substantial. It is well established that PF joint involvement contributes significantly to pain, functional impairment, and disease burden, especially in the early stages of OA (63, 64). Yet patients with isolated PF joint damage often go undetected until the disease progresses to a more advanced, multicompartmental state (65). This delay in detection limits opportunities for early intervention and tailored treatment. Moreover, surgical decision-making, including the consideration of knee arthroplasty, often depends on an accurate understanding of compartment-specific structural damage (66). Therefore, improved assessment tools that focus explicitly on the PF joint are essential for enhancing diagnostic precision and optimising treatment strategies.

Radiomics offers a novel avenue to address this need. By extracting high-dimensional quantitative features from standard radiographic images, radiomics enables detailed analysis of bone and joint structure that is both reproducible and sensitive to subtle variations. While most radiomics research in OA has centred on the TF joint, the PF compartment presents an untapped opportunity for predictive modelling. The anterior location and distinct anatomical configuration of the PF joint make it accessible and well visualised on lateral and skyline radiographic views, which are routinely obtained in clinical settings. This availability makes the PF joint an ideal candidate for targeted radiomics analysis using existing imaging data.

The rationale for developing a PF-specific radiomics model is thus twofold. First, it seeks to fill a critical gap in OA imaging by quantifying structural changes in a compartment frequently overlooked by traditional assessment tools. Second, it leverages the capacity of radiomics to identify imaging biomarkers that may precede symptomatic or clinical progression, enabling earlier risk stratification. This focus aligns with broader efforts in precision medicine to move beyond one-size-fits-all assessments and toward individualised predictions of disease trajectory.

In this chapter, we present a radiomics-based approach tailored specifically to the PF joint. Using handcrafted features derived from standard lateral knee radiographs, we aim to build a predictive model capable of estimating the future risk of undergoing knee replacement surgery. The development and validation of this model provide an opportunity to assess the prognostic value of PF-specific radiomic features and to evaluate whether such features offer additional insights beyond conventional radiographic grading systems. In doing so, this study lays the foundation for more compartment-aware, personalised approaches to OA management.

3.2 Methodology

3.2.1 Data sources and participants

This study utilised the MOST cohort. Participants were included if a valid baseline lateral radiograph of the left knee was available, as determined by a non-specialist quality review. Knees with a history of KR prior to baseline or missing baseline radiographs were excluded. Participants were also excluded if baseline KL grade were missing. Eligible participants were randomly divided into a training set and a testing set at a 2:1 ratio.

3.2.2 Exposures and Imaging Acquisition

Baseline lateral plain radiographs of the left knee were used as the primary imaging source for radiomic feature extraction. All X-rays were acquired using a standardised protocol in a semi-flexed, weight-bearing position. The knee was aligned parallel to the Bucky, with the foot placed against a plexiglass plate to ensure consistent visualisation of the PF joint.

3.2.3 Covariates and Baseline Measures

Baseline covariates included demographic information (age, sex, and body mass index [BMI]) and the KL grade and patellofemoral osteoarthritis (PFOA) of the left knee. Both KL grade and PFOA were performed by two experienced musculoskeletal radiologists through consensus scoring. The timing and occurrence of knee replacement were also recorded for each participant.

3.2.4 Outcomes

The primary outcome was the time to KR event. In the MOST cohort, it is estimated that 95% of recorded KR events represent total KR. However, this level of detail is not explicitly available in the public dataset.

3.2.5 Radiomics Analysis

Lateral knee radiographs from the MOST cohort were quantitatively analysed using radiomics techniques to evaluate their association with KR occurrence over a 60-month follow-up. The radiomics workflow included image preprocessing, ROI generation, feature extraction, model development, and risk stratification, as shown in Figure 3-1.

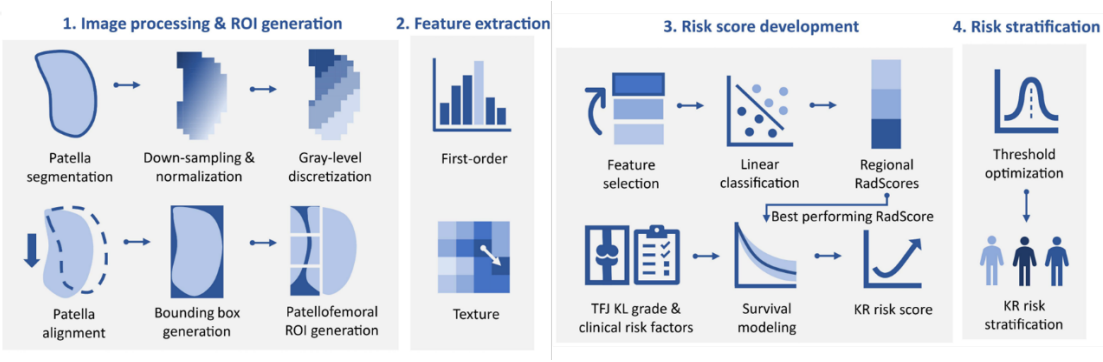


Figure 3-1 The workflow for knee replacement (KR) risk score development and risk stratification.

To enhance reproducibility and reduce noise, all images underwent standardised preprocessing, including resampling, intensity normalisation, and signal enhancement. Left knee lateral X-rays were manually segmented to delineate the patella using the Computer Vision Annotation Tool (CVAT). Image orientation was adjusted to achieve the maximal vertical-to-horizontal ratio of the patellar bounding box, ensuring consistent alignment. The images were then resampled to an isotropic resolution of 0.5 mm × 0.5 mm. Within the patellar region, contrast was normalised via Z-score

transformation and truncated at ±6 standard deviations, followed by grey-level reduction to 32 bins.

Three rectangular ROIs were automatically defined to capture structural information from distinct regions of the PF joint. These ROIs—ROI_{sup}, ROI_{mid}, and ROI_{inf}—were equal-sized partitions extending posteriorly from the lower third of the patellar bounding box towards the trochlear groove. This design ensured simplicity, clinical applicability, and robustness to inter-observer variability. All preprocessing steps and ROI definitions were implemented using the SimpleITK package (v2.2.1) (67).

Radiomic features were extracted from each ROI following the IBSI guidelines (60). A total of 93 first-order and texture features were calculated for both original and filtered versions of the image, resulting in 930 radiomic features per ROI. Filters included Laplacian of Gaussian and wavelet decompositions, which enriched the feature space by highlighting texture patterns. Feature extraction was performed using the PyRadiomics library (v3.0.0) (68). Full extraction parameters are detailed in Table 3-1.

Table 3-1 Image preprocessing and feature extraction parameters.

Parameter	Value
Normalization scale	100
Pixel value offset	600
Pixel value thresholding	0-1200
Resample pixel size (mm)	[0.5,0.5]
Image/mask interpolation algorithm	Nearest neighbour
Mask partial volume threshold	0.5
Interpolation grid alignment	Align grid origins

Gray-level discretization bin number	32
Image filters	Unfiltered, Laplacian-of-Gaussian, Wavelet
Kernel size of Laplacian-of-Gaussian filter (mm)	[1,2,3]
Wavelet filter type	Coilf1
Wavelet filter decompositions	[LL,HL,LH,HH]
Feature class	First-order, GLCM, GLRLM, GLSZM, GLDM, NGTDM

Note: For a detailed description and comprehensive list of all radiomic features analysed in this study, readers are directed to the PyRadiomics documentation available at https://pyradiomics.readthedocs.io/en/latest/features.html. This documentation provides extensive information on each feature's calculation and theoretical background, ensuring thorough understanding and facilitating the reproducibility of our analyses.

For each ROI, a Radiomics Score (RadScore) was constructed through feature selection and model training. The top five features were selected using the minimum Redundancy Maximum Relevance (mRMR) algorithm to balance relevancy with KR outcome and inter-feature independence (69). Ridge regression models were then trained to predict 60-month KR, with Z-score normalised inputs. Given the class imbalance in KR occurrence, an easy-ensemble strategy was adopted: 500 Ridge sub-models were trained on bootstrapped balanced samples and aggregated for final prediction. This was implemented using the imbalanced-learn package (v0.10.1) (70). The RadScore with the highest AUC was selected as the final predictor.

To construct a comprehensive KR risk score, the final RadScore was combined with TF joint KL grade and key demographic variables using a multivariate Cox regression model. This composite score quantified the individual risk of undergoing KR within 60 months.

Based on this score, a three-class stratification system was developed to categorise patients as:

- Non-progressors – no KR within 84 months
- Slow progressors – KR between 30 and 84 months
- Fast progressors – KR within 30 months

Two optimal thresholds were determined using Youden's index to maximise discrimination among the three groups. Importantly, the RadScore, composite KR risk score, and stratification cutoffs were developed exclusively in the training set and validated in the independent testing cohort.

3.2.6 Statistical analyses

To evaluate the independent prognostic value of the constructed RadScore, both univariate and multivariate Cox proportional hazards regression models were employed to estimate hazard ratios (HRs) and corresponding p-values for each risk factor associated with KR.

Subgroup analyses were conducted across different stages of tibiofemoral joint degeneration, categorised by baseline KL grade as early-stage (KL grade < 2), mid-stage (KL grade = 2), and late-stage (KL grade > 2). Within each subgroup, time-dependent receiver operating characteristic (ROC) curves and their associated areas under the ROC curve (AUCs) were calculated to assess discriminatory performance.

A comparator clinical model was also constructed using Cox regression by combining baseline KL grade and PFOA status. The predictive performance of individual predictors (KL grade, RadScore), the clinical model (KL grade + PFOA), and the

comprehensive KR risk score was assessed using the concordance index (C-index) and time-dependent AUCs at 30, 60, and 84 months.

To quantify statistical differences, 95% confidence intervals (CIs) for the C-index were obtained via 1,000-iteration bootstrapping. One-sided permutation tests with 1,000 label permutations were used to compare C-index values between models. All bootstrapping and permutation analyses were implemented using the Python package scikit-survival (71)

Finally, the KR risk stratification performance was assessed using Kaplan–Meier (KM) survival analysis, and its classification accuracy for non-, slow-, and fast-progressors was compared with that of KL grade using confusion matrices. KM analyses were performed using the Python package lifelines (version 0.27.4) (72).

3.3 Results

3.3.1 Participants characteristics

The study utilised data from 3,026 participants aged 50–79 years prospectively enrolled in the MOST cohort. After applying exclusion criteria, 2,943 individuals remained for analysis (Figure 3-2). These were randomly allocated into a training cohort (n = 1,962) and a testing cohort (n = 981).

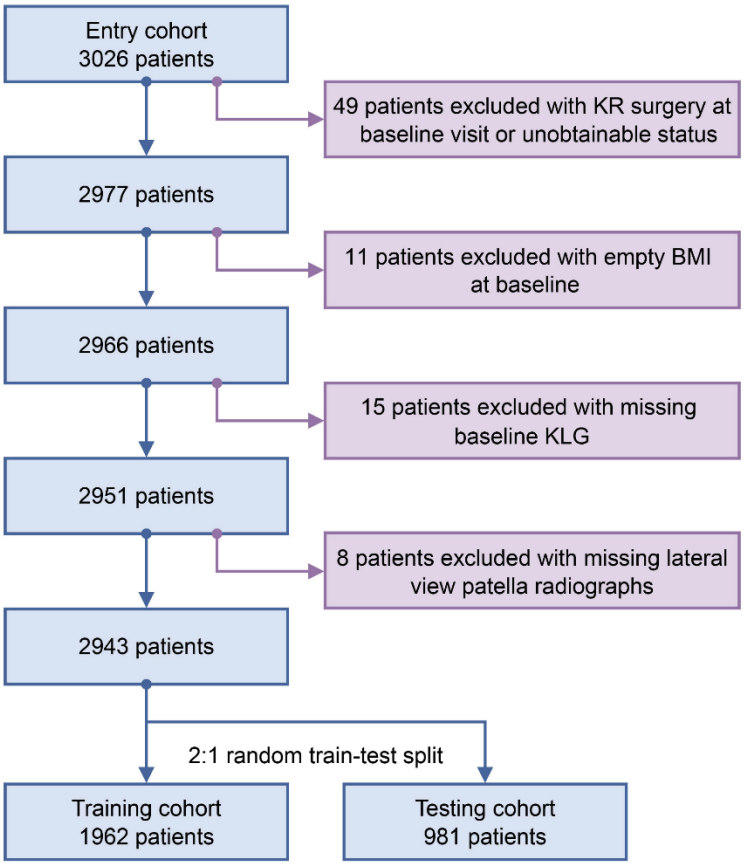


Figure 3-2 Cohort exclusion criteria

Abbreviations: KR = knee replacement, BMI = body mass index, KLG = Kellgren-Lawrence Grade

The distributions of key KR risk factors and the incidence of 84-month left KR are summarised in Table 3-2. The training cohort exhibited a statistically lower baseline BMI compared to the testing cohort (p = 0.020). Other baseline characteristics—including age, sex, KL grade, KR event occurrence, and follow-up duration—were comparable between the two cohorts.

Table 3-2 Distributions of the included knee replacement risk factors of the training and testing patients.

Parameter	Training	Testing	*p*-value
Patient No.	1962	981	

Age		62.42 (50-79)	62.40 (50-79)	0.951
Gender				
	Male	1177	587	0.968
	Female	785	394	
BMI		30.45 (16.72-57.83)	30.99 (18.50-71.91)	0.020
Baseline KL grade				
	0	882	417	0.251
	1	347	165	
	2	293	143	
	3	294	165	
	4	146	91	
Baseline PFOA				
	0	1546	755	0.519
	1	283	155	
	Missing	133	71	
84-month KR				
	0	1753	876	1.000
	1	209	105	
KR follow-up time				
	Medium	84	84	0.536
	Range	2-97	3-101	

Note: *p*-values were acquired by Student t-test for continuous variables and nominal variables with > 5 levels, including age, BMI, and KR follow-up time. The rest of the nominal and categorical variables were compared by the Chi-square test.

Abbreviations: KR = knee replacement, BMI = body mass index, KL = Kellgren and Lawrence, PFOA: patellofemoral osteoarthritis.

Among patients without recorded KR events, a proportion were lost to follow-up: 17/8 patients (training/testing) before 30 months, 42/18 before 60 months, and 79/35 before 84 months, respectively.

3.3.2 RadScore composition

Heterogeneous RadScore performances were found among the three ROIs in 60-month KR prediction (Figure 3-3). ROI_{inf}, covering the inferior PF joint area, reached the highest 60-month KR prediction performance with training and testing AUCs of 0.81 and 0.82, respectively. Therefore, ROI_{inf} was chosen as the ROI for the final RadScore for KR risk score development. Details of the final RadScore compositions can be found in Table 3-3.

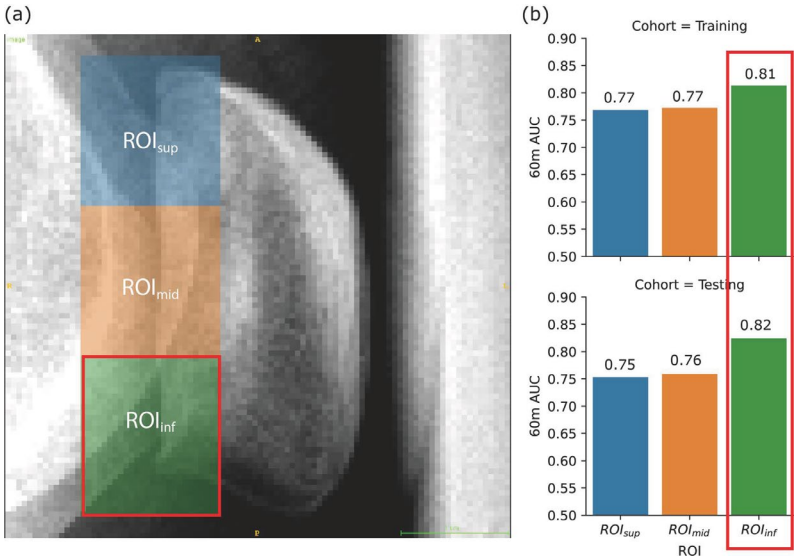


Figure 3-3 (a) Region-of-interest (ROI) segmentation of one example patient. (b) Bar plot of RadScore AUC of each ROI in prediction 60-month KR.

Note: ROI_{inf}, which is located at the inferior region of the patellofemoral joint and marked by green rectangles, achieved the best performance in both training and testing.

Table 3-3 Details of the selected radiomic features and model coefficients of the final patella

RadScore.

Alias	Image	Class	Name	Mean coefficient	Norm-mean	Norm-scale
F1	Original	First Order	Mean	0.199	555.43	65.32
F2	Original	NGTDM	Strength	-0.160	16.34	7.67
F3	LoG (sigma=2mm)	GLRLM	ShortRunHigh-GrayLevelEmphasis	-0.090	292.31	46.85
F4	Wavelet (LH)	GLCM	ClusterShade	-0.097	-35.24	33.18
F5	Wavelet (LL)	First Order	Mean	0.198	1110.87	130.63
Intercept				--0.24		

Note: RadScore can be calculated by the linear combination of the 10 radiomic features plus the intercept: $RadScore = \sum_i (f_i - m_i)/s_i \cdot c_i + a$, where f_i is value of the *i*th feature (Fi), m_i is the normalization mean, c_i is the normalization scale, c_i is the mean coefficient, and a is the intercept.

3.3.3 RadScore's independence and predictive value

The PF joint RadScore is an independent risk factor (p-value < 0.001) for KR in both univariate and multivariate settings, as reported in Table 3-4. During the univariate test, all the demographic and radiographic factors were significantly associated with KR in training and testing. However, only KL grade and RadScore persisted as independent prognostic factors in both training and testing. Notably, PFOA did not demonstrate independent prognostic value with the presence of RadScore.

Table 3-4 Univariate and multivariate survival analysis results of the final RadScore, baseline KL grade, and other knee replacement risk factors in training and testing

Cohort	Risk factor	Univariate	Multivariate

		HR (95CI)	*p*-value	HR (95CI)	*p*-value
Training	RadScore	2.49 (2.19-2.83)	**< 0.001**	1.45 (1.21-1.73)	**< 0.001**
	Age	1.04 (1.02-1.05)	**< 0.001**	1.01 (0.99-1.03)	0.270
	Gender	0.74 (0.56-0.99)	**0.046**	0.81 (0.60-1.09)	0.172
	BMI	1.06 (1.04-1.08)	**< 0.001**	0.99 (0.97-1.01)	0.409
	KL grade	2.53 (2.25-2.84)	**< 0.001**	1.88 (1.63-2.17)	**< 0.001**
	PFOA	4.72 (3.55-6.26)	**< 0.001**	1.19 (0.85-1.68)	0.304
Testing	RadScore	2.29 (1.91-2.75)	**< 0.001**	1.40(1.09-1.80)	**0.009**
	Age	1.04 (1.01-1.06)	**0.002**	1.00 (0.98-1.03)	0.745
	Gender	0.61 (0.40-0.93)	**0.022**	0.67 (0.43-1.04)	0.073
	BMI	1.05 (1.02-1.07)	**< 0.001**	0.99 (0.96-1.02)	0.682
	KL grade	2.39 (2.02-2.82)	**< 0.001**	2.02 (1.65-2.47)	**< 0.001**
	PFOA	4.17 (2.79-6.23)	**< 0.001**	1.12 (0.69-1.83)	0.635

Note: Univariate and multivariate survival analyses were performed by Cox regression. *P*-value less than 0.05 (bolded) was considered significant.

Abbreviations: RadScore, radiomics score; HR, hazard ratio; 95CI, 95% confidence interval; BMI, body mass index; KL grade, Kellgren and Lawrence grade; PFOA, patellofemoral osteoarthritis.

During subgroup analysis, RadScore demonstrated high predictive values for fast progressors (KR+ (30m)) within early-stage patients (KL grade < 2) with AUC = 0.94/0.75 (training/testing) (Figure 3-4). It also had high discriminative power for KR at all time points within mid-stage patients (KL grade = 2), and the highest AUCs (0.83/0.79) were also achieved in predicting 30-month KR occurrence.

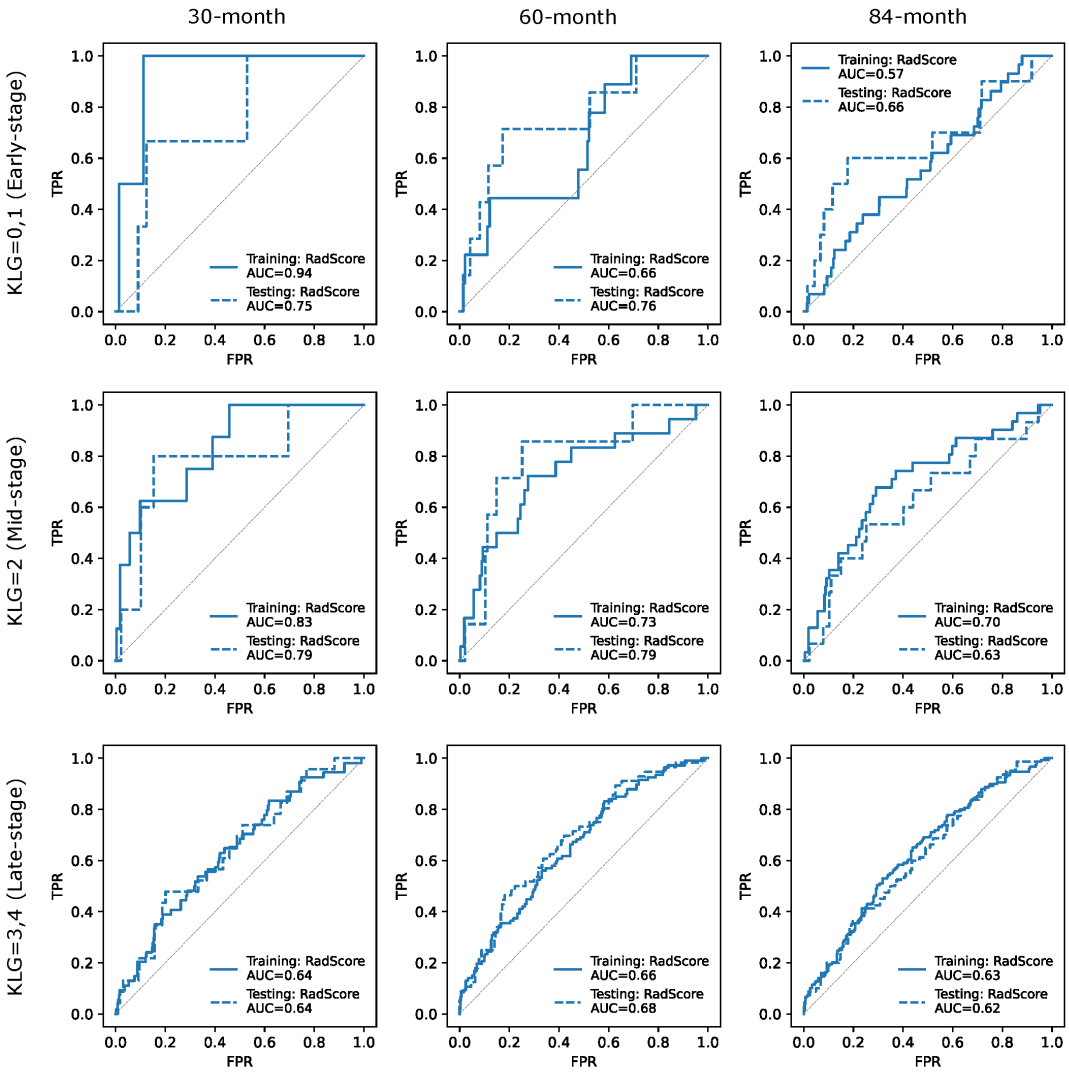


Figure 3-4 The receiver operating characteristic curves of RadScore in predicting 30-, 60-, and 84-month KR classification in training and testing under different disease stages at baseline visit.

3.3.4 Optimal prognostic performance by KR risk score

The KR risk score achieved the best KR prognosis performance by combining RadScore and KL grade. Table 3-5 reports the performance comparison by C-index and time-dependent AUCs among KL grade, RadScore, PFOA + KL grade, and KR risk score (RadScore + KL grade). KL grade itself had a C-index of 0.83 in training and 0.82 in testing, which were significantly improved by the addition of RadScore to 0.85/0.84 (*p*-value = 0.003/0.002). Similar trends were observed in time-dependent AUCs where

the highest was achieved by the KR risk score (30-month: 0.91/0.83, 60-month: 0.89/0.87, 84-month: 0.86/0.86). The KR risk score also achieved significantly higher C-index values than the PFOA + KL grade model in both the training (*p*-value = 0.035) and testing set (*p*-value = 0.011). Hazard ratios and *p*-values of the covariates of the KR risk score are presented in Table 3-6.

Table 3-5 Training and testing performance of three knee replacement risk prediction models.

Model	Training		Testing	
	C-index	***p*-value**	**C-index**	***p*-value**
KL grade	0.83 (0.81-0.86)	**0.003**	0.82 (0.78-0.86)	**0.002**
RadScore	0.78 (0.75-0.81)	**<0.001**	0.78 (0.74-0.82)	**0.018**
PFOA + KL grade	0.84 (0.81-0.86)	**0.035**	0.82 (0.78-0.86)	**0.011**
KR risk score	0.85 (0.82-0.87)	**-**	0.84 (0.80-0.87)	**-**
	30m AUC	***p*-value**	**30m AUC**	***p*-value**
KL grade	0.90 (0.86-0.92)	**0.007**	0.80 (0.73-0.87)	**0.020**
RadScore	0.83 (0.79-0.88)	**0.015**	0.81 (0.74-0.87)	0.290
PFOA + KL grade	0.90 (0.87-0.92)	**0.023**	0.81 (0.74-0.88)	0.191
KR risk score	0.91 (0.89-0.94)	**-**	0.83 (0.77-0.89)	**-**
	60m AUC	***p*-value**	**60m AUC**	***p*-value**
KL grade	0.87 (0.85-0.90)	**<0.001**	0.84 (0.79-0.89)	**<0.001**
RadScore	0.82 (0.78-0.85)	**<0.001**	0.83 (0.78-0.87)	0.123
PFOA + KL grade	0.88 (0.85-0.91)	**0.045**	0.85 (0.79-0.90)	**0.002**
KR risk score	0.89 (0.87-0.92)	**-**	0.87 (0.83-0.91)	**-**
	84m AUC	***p*-value**	**84m AUC**	***p*-value**
KL grade	0.84 (0.81-0.87)	**0.008**	0.84 (0.80-0.87)	**0.003**

RadScore	0.78 (0.74-0.81)	**<0.001**	0.79 (0.74-0.83)	**0.009**
PFOA + KL grade	0.85 (0.82-0.87)	0.093	0.84 (0.80-0.88)	**0.007**
KR risk score	0.86 (0.83-0.88)	-	0.86 (0.82-0.89)	-

Note: One-sided *p*-values were calculated by permutation test with 1000 iterations. *P*-value less than 0.05 (bolded) was considered significant. Significant performance improvements can be observed when combining RadScore with KL grade, compared to using KL grade alone.

Abbreviations: C-index = concordance index, RadScore = radiomics score, KL = Kellgren and Lawrence, 30m = 30-month, 60m = 60-month, 84m = 84-month.

Table 3-6 Coefficients of the KR risk score built by multivariate Cox regression.

Covariate	**Hazard ratio (95% confidence interval)**	***p*-value**
RadScore	1.49 (1.27-1.74)	< 0.001
KL grade	1.94 (1.70-1.23)	< 0.001

3.3.5 Risk stratification and survival analysis

Significant KR risk score differences were detected among the three follow-up time points, as shown in Figure 3-5. The KR risk scores of non-progressive patients at 84 months were the lowest, with an average value of 0.74. They were significantly higher (*p*-value < 0.001) for slow progressors (KR+ (30m-84m)) with an average value of 2.42. Fast progressors (KR+ (30m)) achieved the highest average KR risk score (2.96), which is significantly higher than slow progressors (*p*-value < 0.001).

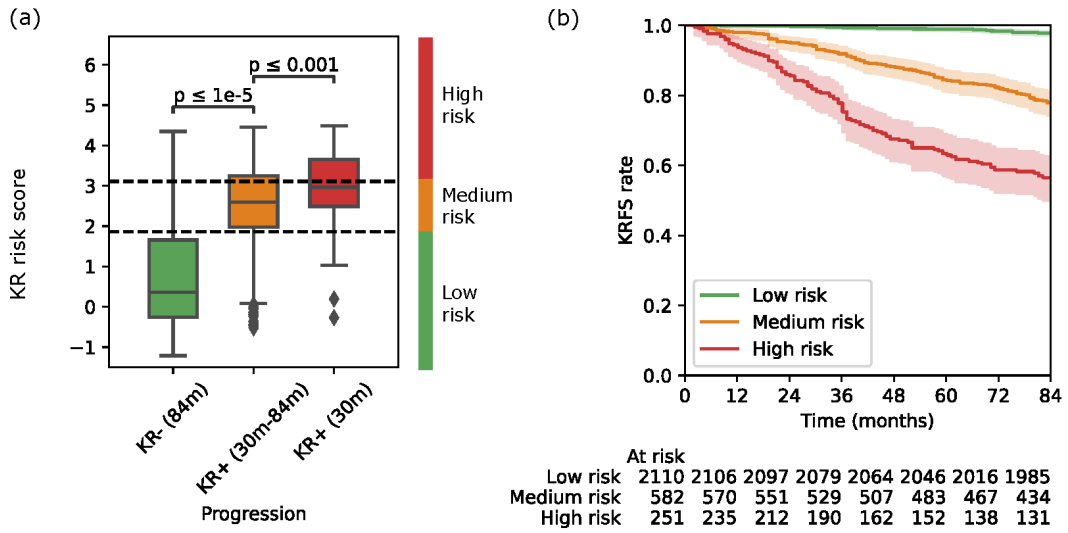


Figure 3-5 Distribution of KR Risk Score and Survival Outcomes by Risk Group

Note: (a) KR risk score (RadScore + KL grade) distribution comparisons among non-progressors within 84 months (KR- (84m)). (b) Knee replacement-free survival (KRFS) curves of the low risk (green), medium risk (orange), and high risk (red) patients from Kaplan-Meier analysis on the entire patient cohort.

Three risk groups were stratified based on the optimised KR risk score thresholds of 1.86 and 3.11 (Figure 3-5(a), dashed lines), and distinct survival patterns were observed among the three risk groups Figure 3-5(b). Patients with KR risk score less than 1.86 were classified as low risk (n = 2110) with minimum risk of KR progression within 84 months (6%), as drawn by the survival confusion matrix in Figure 3-6. Meanwhile, patients with the score of more than 1.86 but less than 3.11 were classified as medium risk (n = 582), showing a relatively higher risk of KR within 84 months (25%), but the fast progression (KR+ (30m)) rate remained as low as 7%. The high-risk group of patients (n = 251) who had the score greater than 3.11 demonstrated the highest risk of receiving KR within 84 months (48%) and 30 months (19%). In contrast, only 11% and 3% of KL grade = 2 patients were slow and fast progressors, respectively. Although similar rates of slow (49%) and fast (20%) progressors were achieved by the KL grade

of 4, more progressive patients were identified by the proposed high-risk criteria. Specifically, the positive predictive value (PPV) of our RadScore for predicting knee replacement (KR) was 46.41%, significantly outperforming the KL grade's PPV of 34.54%. Furthermore, in predicting KR within 30 months, the RadScore's PPV was 27.01%, compared to only 11.61% for the KL grade. These findings underscore the enhanced precision of RadScore over KL grade in identifying patients at high risk of both overall KR and rapid progression to KR within a shorter timeframe.

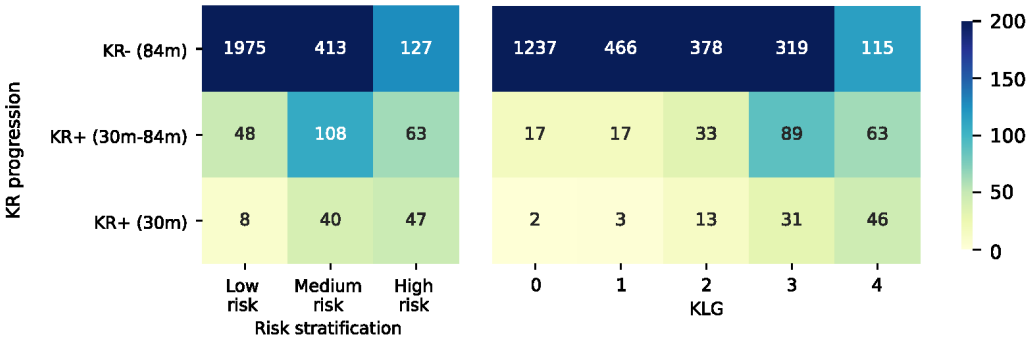


Figure 3-6 Confusion matrix of the proposed stratification system and KLG in predicting the three KR progression speeds.

3.4 Discussion

This study, for the first time, highlights the importance of quantitative analysis of PF joint from lateral knee radiographs in KR prediction. It also provides a comprehensive tool incorporating TF and PF joints radiographic information for assisting clinicians in stratifying patients based on disease progression speed. The developed PF joint RadScore was validated as an independent prognostic factor for KR and achieved better KR prognostic performance in early- and mid-stage. The comprehensive KR risk score achieved the highest performance based on the combination of two joints' radiographic information. Distinct KR-free survival patterns were delineated for the three stratified

risk groups, which could benefit precise rehabilitation therapy by prioritising higher risk patients with faster disease progression.

3.4.1 Clinical Implications

Despite the heterogeneous KR prediction performance, all three regional ROIs of PF joint demonstrated certain prognostic values. Those ROIs are located at the surface between the patella and the femoral notch, known as the trochlea, which is a key area of contact between these bones. According to Wolff's Law, bones adapt to the loads under which they are placed. Therefore, changes in this area can reflect the abnormal stresses on the knee, indicating early signs of OA. Previous research by Bayramoglu et al. once emphasised the importance of ROI location (30). It confirmed the PFOA diagnostic ability of two lateral patella ROIs at the PF joint margin (73), which was consistent with the ROI definitions in our study. The best-performing ROI was located on the inferior region of the PF joint, with a significant area outside the patella bone. Based on the distinctive patella shape differences observed from the three groups of patients, the final RadScore, built mainly from first-order radiomic features, may capture the patella morphological change due to the altered mechanical loading with knee joint deterioration. Similarly, previous studies have suggested that patella shape and alignment strongly correlate with PFOA, PF joint cartilage defect, and physical activity reduction (74-78). Such visually appreciable changes were effectively captured and quantified by radiomics, which might reduce the inter-observer variability and improve diagnostic consistency.

3.4.2 Independent predictors of KR

Results from multivariate analysis suggest that our radiomic characterisation of PF joint

on lateral radiographs (RadScore) was independently prognostic to the TF joint KL grade, and the integration of RadScore to KL grade could significantly boost the performance of KR prediction in the MOST dataset. Despite the limited increments in C-index and AUC values in the entire MOST cohort, our model revealed its unique advantages in predicting fast progressors among early- and mid-stage patients in the subgroup analysis. In contrast, the study that primarily focused on the TF joint demonstrated the highest performance for late-stage patients (79). This is consistent with previous research conclusions indicating that the PFOA manifests before the TFOA (80). Predicting fast progressors in the early- and mid-stage is crucial, as early intervention may alter the disease trajectory and lead to improved outcomes.

On the other hand, demographic information had limited independent prognostic values for KR prediction, and the current clinical diagnostic criteria for the PF joint (PFOA) did not achieve an independent prognostic value. This finding further underlines the importance of PF joint as well as its quantitative characterisation compared to the other risk factors. It may also suggest stronger correlations of PF joint with symptomatic presentations, which is consistent with previous clinical observations (81).

3.4.3 Limitations of this study

Several limitations in this study in data interpretation shall be fully aware of, which warrant further improvements in future investigations. First, only the initial visit radiographs were analysed for KR prediction. A dynamic risk assessment method using image sets from a time series may further improve prediction accuracy. Second, although the MOST dataset is combined by several cohorts, a comprehensive assessment of the proposed patella RadScore and KR risk score on various external datasets with different patient distributions is necessary to further demonstrate the

model's generalisability. We have investigated the OAI dataset, but it cannot fulfil our purposes. The dataset lacks sufficient subjects with lateral view X-rays; none of these cases had KR surgery records. Future research could explore alternative datasets or await updated data releases. Third, our machine learning analysis of lateral view radiographs requires patella segmentation, which was achieved by manually contouring. In addition, the KL grade of the PA view was acquired by manual reading. A fully automated risk assessment pipeline requires automatic lateral view patella segmentation and quantitative TF and PF joint assessments from the PA view radiographs, which will be conducted in the next stage of our research.

3.5 Chapter summary

In summary, we developed a PF joint RadScore on lateral knee radiographs, which was validated as an independent prognostic factor to predict KR risk among Knee OA patients. The KR risk score that incorporates TF and PF joints radiographic information achieved the best KR prognostic performance. Based on this score, the stratification system could triage knee OA patients into three distinct KR-free survival groups to reflect the progression speed. It would serve as a clinical reference to guide exercise or other physical therapy for secondary prevention of knee OA deterioration.

Chapter 4: Generalisability of a Patellofemoral Radiomics Model Across Multiple Cohorts Using Domain Adaptation

4.1 Chapter overview

Radiomics-based modelling has shown increasing promise in the assessment of knee OA, particularly in its ability to extract high-dimensional imaging biomarkers that are predictive of disease progression and surgical outcomes. In previous work, a PF-specific radiomics model developed using handcrafted features from lateral knee radiographs demonstrated strong performance in identifying individuals at elevated risk for total knee replacement. However, the translation of such models from research environments to broader clinical practice remains constrained by a critical limitation: generalizability.

Radiomics models are inherently sensitive to variations in image acquisition protocols, scanner types, population demographics, and data preprocessing pipelines. These factors can lead to significant shifts in feature distributions across datasets—a phenomenon referred to as “domain shift.” As a result, models trained on a single dataset may perform well in internal validation but exhibit reduced accuracy or calibration when applied to external populations. This issue poses a major barrier to the clinical implementation of radiomics, especially in multi-centre studies or across international cohorts where imaging conditions and patient profiles are not standardised.

In the context of knee OA, this challenge is especially pronounced. Existing large-scale cohorts often differ by region, race, healthcare system, and radiographic technique. For example, U.S.-based cohorts typically use a standardised fixed-flexion or weight-

bearing protocol, while cohorts from Asia or Europe may use alternative positioning methods. Even within the same modality, slight differences in resolution or contrast can affect the derived radiomic features. Furthermore, patient-specific characteristics—such as body mass index, skeletal morphology, or disease phenotype—may introduce additional variability in the appearance of joint structures on radiographs. Without appropriate adjustments, these factors can diminish the external validity of any trained model.

Domain adaptation has emerged as a viable solution to this problem. It refers to a class of machine learning techniques aimed at minimising the discrepancy between source and target data distributions, thereby enabling more consistent model performance across diverse datasets. In contrast to traditional model retraining, which requires large, annotated datasets from each new site, domain adaptation methods allow knowledge transfer using a limited number of samples from the target domain. This is especially advantageous in medical imaging, where data sharing is often restricted, and label availability may be limited.

In this chapter, we investigate the generalisability of the previously developed PF-specific radiomics model across multiple OA cohorts using domain adaptation strategies. Specifically, we explore the generalised patellofemoral radiomics (GPR) model's robustness when applied to international datasets differing in ethnicity, imaging protocol, and sample size. We evaluate both direct transfer performance and adaptation-enhanced performance, aiming to quantify the degree to which domain adaptation can mitigate performance degradation in new populations. By doing so, this work addresses a crucial translational barrier in OA radiomics research and contributes to the development of deployable, population-agnostic imaging biomarkers for clinical use.

4.2 Methodology

4.2.1 Data sources and participants

This study leveraged multicontinental cohorts for model training, refinement, and external testing. The MOST cohort (North America: US) was used for training, the HK cohort (Asia: HK) for refinement, and the Meniscal Tear and Osteoarthritis Risk (MenTOR) and Knee Injury Cohort at the Kennedy (KICK) cohorts (Europe: UK) for external testing.

In brief, we re-labelled the MOST dataset as the US cohort, the HADCL dataset as the HK cohort, and the combined MenTOR and KICK datasets as the UK cohorts. The UK cohorts were selected to represent individuals with early-stage radiographic OA changes drawn from high-risk populations of knee injury, allowing for the evaluation of model generalisability in younger individuals.

Eligibility criteria for this study: All cohort participants required a 'valid' baseline lateral knee X-rays on non-specialist review (see below). Participants were excluded if they had less than 60 months of follow-up without a recorded endpoint or sufficient interim data to definitively rule out reaching the study endpoint within 60 months. Knees with a history of KR before baseline or missing baseline radiographs (see below) were excluded. The knee with the higher KL grade or closer to the endpoint was designated as the index knee for analysis; if data for that knee were missing, the other knee was selected. The STROBE diagram can be found in Figure 4-1.

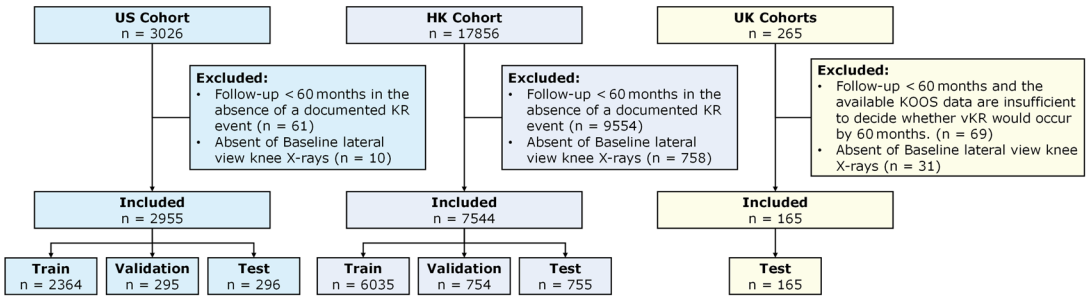


Figure 4-1 STROBE Diagram of Participant Selection and Cohort Inclusion Criteria.

Note: US = United States; HK = Hong Kong; UK = United Kingdom; KR = Knee Replacement; KOOS = Knee Injury and Osteoarthritis Outcome Score; vKR = Virtual Knee Replacement.

Subjects after inclusion and exclusion in the US and HK cohorts were randomly divided into training, validation, and testing sets in an 8:1:1 ratio. Subjects after inclusion and exclusion in the UK cohorts were combined and used as an external testing set.

4.2.2 Exposures and Imaging Acquisition

Lateral knee plain X-rays were used as the primary imaging source for radiomics feature extraction to be used as the exposure. The radiograph set was from the participant's baseline visit (or earliest visit after enrolment).

In the US cohort, X-rays were acquired using a standardised protocol, with participants in a semi-flexed, weight-bearing position. The knee was aligned parallel to the Bucky, and the foot was placed against a plexiglass plate, ensuring consistent visualisation of the PF joint.

In the HK cohort, X-rays were collected from routine clinical records across multiple hospitals. As no unified imaging protocol was applied, considerable heterogeneity in knee positioning, flexion angle, and exposure settings existed in real-world clinical

practice.

For the UK cohorts, X-rays were typically non-weight-bearing (participant lying on side and semi-flexed). Proper collimation, superimposition of femoral condyles, and standardised exposure settings were used to ensure consistent image quality.

These variations in imaging protocols between cohorts were one of the considerations when selecting cohorts, to enable typical variation in radiograph acquisition of a lateral view, as in clinical care, to apply domain adaptation techniques to improve model generalisability across diverse clinical environments.

For all cohort X-rays, a simple (non-specialist) quality control review at upload removed those where the whole PF joint was not present, or the index knee X-ray was not available/had been mislabelled. These formed the radiographic dataset for each cohort.

4.2.3 Covariates and Baseline Measures

Baseline covariates included age, sex, and BMI, all of which were recorded at the time of the radiograph imaging. Structural severity of knee OA was assessed using the KL grade of the TF joint (an ordinal grade, from 0-4). PFOA status was evaluated using lateral knee X-rays and recorded as a binary variable (presence or absence of radiographic OA features). Both KL grade and PFOA served as a comparator/reference with the model rather than a model input. They were determined by two trained musculoskeletal radiologists through consensus scoring within each of the studies. In the HK cohort, the data of BMI, KL grade, and PFOA were not recorded as discrete values.

In a subset of MOST study participants, WORMS was available to quantify cartilage morphology changes in the knee joint. WORMS was correlated with X-ray-derived parameters to understand the bone-cartilage pathological interrelationship.

4.2.4 Outcomes

The primary outcome of this study was the genuine KR (for US and HK cohorts) or virtual knee replacement (vKR, for UK cohorts) within a 60-month follow-up. In all cases, this was a binary variable (present/absent).

For the US cohort, an estimation of 95% of KR events recorded in the US cohort were total KR, although such information was unavailable in the public dataset. For the HK cohort, all KR events referred to total joint replacement surgery, labelled by ICD code 81.54.

The Knee Injury and Osteoarthritis Outcome Score (KOOS) is a 5-domain patient-reported survey and was the primary outcome measure for both the UK cohorts, so available for all participants at relevant follow-ups. In the UK cohorts, where KR events were limited due to the relatively young study, earlier stage populations, we applied a validated algorithm to define vKR based on longitudinal KOOS subscale scores to reflect the risk of disease progression (82). Full details of the vKR calculation and implementation are provided in the Appendix.

4.2.5 Radiomics Analysis

Radiomic features were extracted from lateral knee X-rays using our established protocol (83) via an open-source package named PyRadiomics (v3.0.1) (68). Radiomic features were combined with demographic features (baseline age, sex, and baseline

BMI) for feature selection. The workflow can be found in Figure 4-2. Feature selection was performed using the mRMR algorithm (69). Selected features were constrained to less than 10% of the test set size and normalised using Z-scores (11).

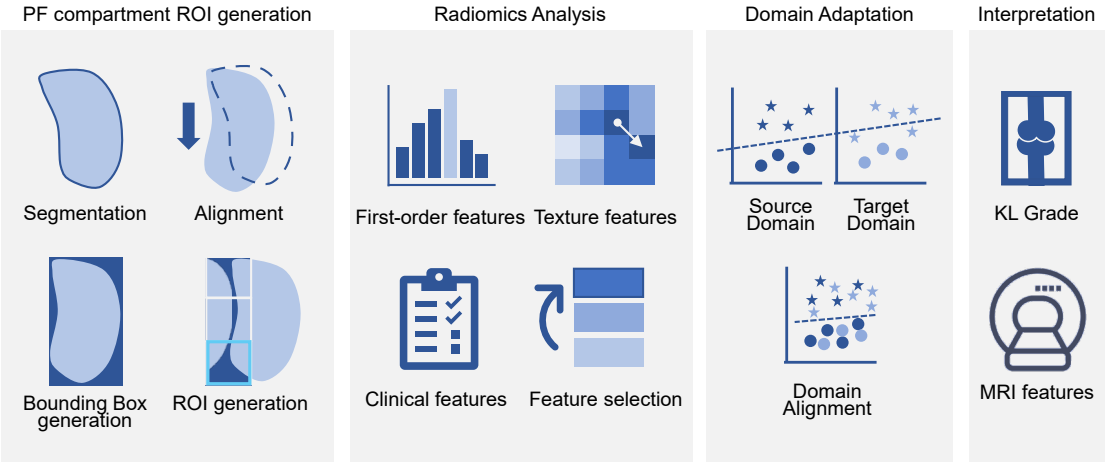


Figure 4-2 Methodological workflow for the study.

Note: ROI = Region of Interest; KL = Kellgren-Lawrence; MRI = Magnetic resonance imaging.

In our previous model (83), feature selection had been performed independently for the US and HK, as shown in Figure 4-3(A). In contrast, here in this domain adaptation approach, our GPR model constructed a shared feature space by using the intersection of the top-ranked features from the US and HK cohorts, retaining only features consistently informative across both groups for training the domain-adapted model.

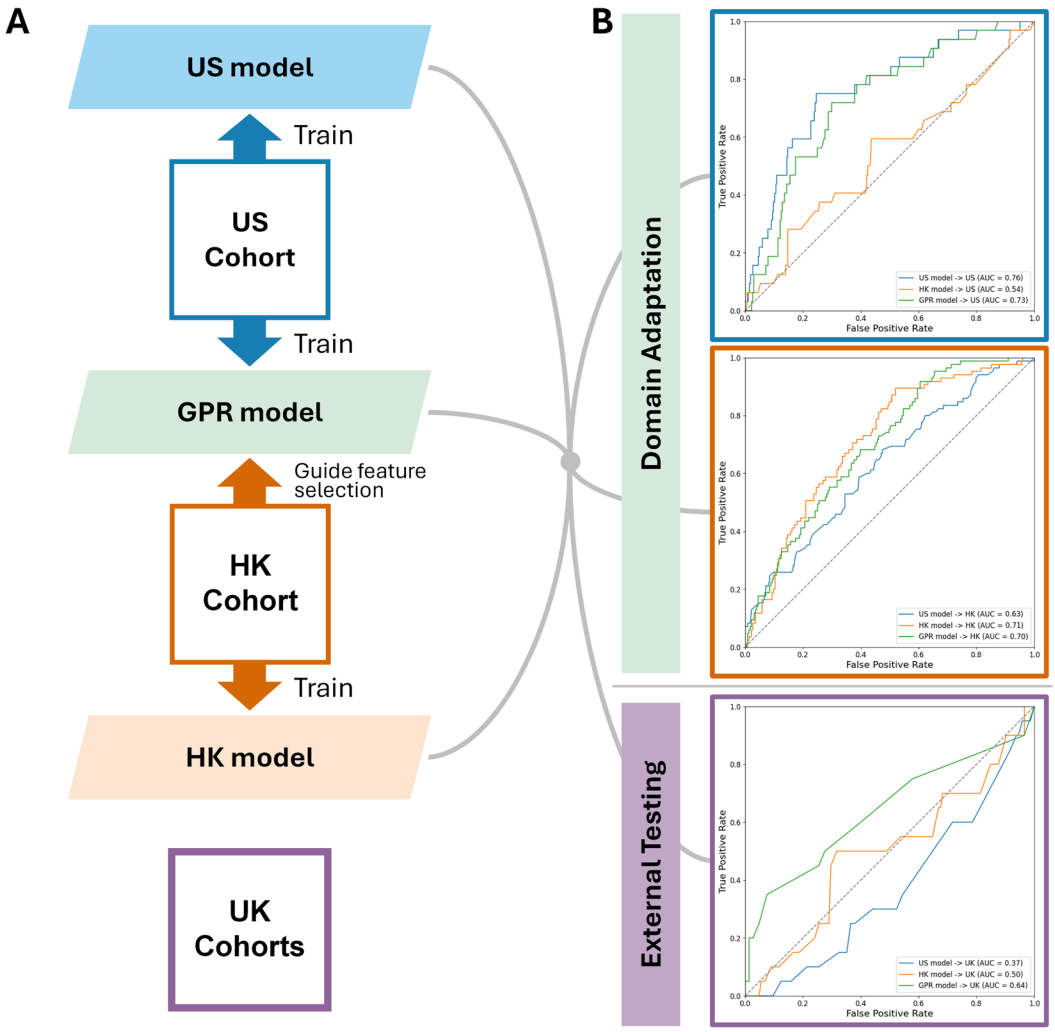


Figure 4-3 Model Development and Validation Across Cohorts.

Note: (A) Overview of model development for the US model, HK model, and GPR model across the cohort datasets.

(B) ROC curves comparing the performance of the three models (US model, HK model, and GPR model) in predicting KR across cohorts. Upper panel: Model performance in the US cohort. Middle panel: Model performance in the HK cohort. Lower panel: Model performance in the UK cohorts. Abbreviations: US = United States; HK = Hong Kong; UK = United Kingdom; GPR = Generalised Patellofemoral Radiomics; ROC = Receiver Operating Characteristic; KR = Knee Replacement; AUC = Area Under the Receiver Operating Characteristic Curve.

The Gradient Boosting Classifier, a machine learning algorithm that builds an ensemble of decision trees to improve prediction accuracy, was employed for the classification of genuine and virtual KR events. The feature set that achieved the highest average AUC across both the US cohort and HK cohort was selected as the best-performing set and used to construct the GPR Score, scaled from 0 to 1, which assessed the likelihood of receiving KR based on radiomic features.

To assess the performance and generalisability of the proposed models, we compared three different approaches: the US Model (trained on the US cohort), the HK Model (trained on the HK cohort), and the GPR Model (trained on intersecting features from the US and HK cohorts).

To further evaluate predictive performance, AUC values for GPR Score, PFOA status, KL Grade, and their combination (KL Grade + GPR Score) were compared in both the US and UK cohorts using Bayesian bootstrap resampling. Subgroup analyses were conducted in the US cohort, stratifying participants by OA severity (KL grade ≤ 2 vs. KL grade > 2), to determine the GPR Score's effectiveness across disease stages.

4.2.6 Statistical Analysis

Baseline characteristics were summarised as median (IQR; range) for continuous variables and n (%) for categorical variables. Differences between cohorts and the US cohort were assessed using the t-test for age and BMI, the Mann–Whitney U test for KL Grade, and the Chi-square test for sex, PFOA and KR/vKR event.

Model performance was assessed using AUC as the primary metric to assess classification accuracy in the GPR score for correctly classifying KR within 60 months.

Bayesian Bootstrap Resampling (84) with 10,000 iterations was used to compare AUC differences between models. In each iteration, normalised weights were sampled from an exponential distribution, and weighted AUCs were computed. The resulting distribution of AUC differences was used to estimate the two-sided probability ($Pr[AUC_1 > AUC_2]$) that one model outperformed another.

To understand the histopathology underlying the GPR score, the association between the GPR Score and joint structure was evaluated using Spearman correlation analyses with baseline MRI-based WORMS cartilage morphology scores in the US cohort. For comparison, correlations with PFOA status and KL Grade were also assessed. Spearman ρ and *p* value were calculated, and the results are shown in a correlation matrix.

4.3 Results

4.3.1 Participant characteristics

A total of 10664 participants, with 2,955 from the US cohort, 7,544 from the HK cohort, and 165 from the UK cohorts, were included in the study (Figure 4-1). Baseline characteristics (at the time of imaging), including age, sex, BMI, KL grade, and the frequency of PFOA, and the incidence of 60-month KR, were summarised in Table 4-1. Details for the MenTOR and KICK cohorts in the UK were listed in Table 4-2.

Table 4-1 Baseline Characteristics of the Studied Cohorts

	US Cohort	**HK Cohort**	***p*-value**	**UK Cohorts**	***p*-value**
Participants	2955	7544	/	165	/

Age*		62 (55–69; 50–79)	66 (57–77; 45–104)	<0.001	38 (26–49; 17–62)	<0.001
Sex	Male	1789 (60.5%)	2340 (31.0%)	<0.001	116 (70.3%)	0.015
	Female	1166 (39.5%)	5204 (69.0%)		49 (29.7%)	
BMI*		29.86 (26.66–33.71; 16.72–71.91)	/	/	26.63 (23.27–29.84; 18.94–49.90)	<0.001
KL Grade*	0	960 (32.5%)	/	/	84 (50.9%)	<0.001
	1	635 (21.5%)	/		25 (15.2%)	
	2	548 (18.5%)	/		34 (20.6%)	
	3	453 (15.3%)	/		17 (10.3%)	
	4	356 (12.0%)	/		0 (0.0%)	
	Missing	3 (0.1%)	/		5 (3.0%)	
PFOA*	Yes	535 (18.1%)	/	/	20 (12.1%)	0.237
	No	2237 (75.7%)	/		115 (69.7%)	
	Missing	183 (6.2%)	/		30 (18.2%)	
KR Event in 60m	Yes	325 (11.0%)	619 (8.2%)	<0.001	20[†] (12.1%)	0.749
	No	2630 (89.0%)	6925 (91.8%)		145 (87.9%)	

Note: Data are median n (IQR; range) or n (%). Percentages may not sum to 100% due to rounding. P-values were compared between this group and the US cohort. They were obtained using the t-test for continuous variables, including age and BMI. KL Grade was analysed using the Mann-Whitney U test. The remaining categorical variables were compared using the Chi-square test.

*** Baseline refers to the time of imaging.**

† KR Event for UK cohorts was assessed by vKR criteria.

Abbreviations: US = United States; HK = Hong Kong; UK = United Kingdom; BMI = Body Mass Index; KL = Kellgren–Lawrence Grade; PFOA = Patellofemoral Osteoarthritis; KR = Knee Replacement; vKR = Virtual Knee Replacement; IQR = Interquartile Range.

Table 4-2 Baseline Characteristics of the MenTOR & KICK

		UK Cohorts (MenTOR & KICK combined cases)	**MenTOR**	**KICK**
Patient No.		165	77	88
Age*		38 (26–49; 17–62)	50 (45–53; 31–62)	26 (23–34; 17–50)
Sex	Male	116 (70.3%)	48 (62.3%)	68 (77.3%)
	Female	49 (29.7%)	29 (37.7%)	20 (22.7%)
BMI*		26.63 (23.27–29.84; 18.94–49.90)	28.70 (24.83–32.38; 19.03–49.90)	25.09 (22.81–28.31; 18.94–38.94)
KL Grade*	0	84 (50.9%)	5 (6.5%)	79 (89.8%)
	1	25 (15.2%)	21 (27.2%)	4 (4.5%)
	2	34 (20.6%)	31 (40.3%)	3 (3.4%)
	3	17 (10.3%)	17 (22.1%)	0 (0.0%)
	4	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Missing	5 (3.0%)	3 (0.0%)	2 (2.3%)
PFOA*	Yes	20 (12.1%)	9 (11.7%)	11 (12.5%)

	No	115 (69.7%)	65 (84.4%)	50 (56.8%)
	Missing	30 (18.2%)	3 (3.9%)	27 (30.7%)
KR Event in 60m†	Yes	20 (12.1%)	19 (24.7%)	1 (1.1%)
	No	145 (87.9%)	58 (75.3%)	87 (98.9%)

Note: Note: Data are median n (IQR; range) or n (%). Percentages may not sum to 100% due to rounding.

*** Baseline refers to the time of imaging.**

† KR events were assessed by vKR criteria.

Abbreviations: UK = United Kingdom; BMI = Body Mass Index; KL = Kellgren–Lawrence Grade; PFOA = Patellofemoral Osteoarthritis; KR = Knee Replacement; vKR = Virtual Knee Replacement; IQR = Interquartile Range.

The HK cohort differed significantly from the US cohort in age, sex, and KR incidence. Compared to the US cohort, the UK cohorts were generally younger and had lower KL grades and BMI, as expected. However, the frequency of PFOA and KR/vKR was similar between the US and the UK.

4.3.2 Domain adaptation

Single-cohort-trained models showed reduced performance when applied to other cohorts (Figure 4-3(B), Table 4-3). The US-trained model achieved an AUC of 0.76 (95% CI: 0.68–0.84) in the US cohort but an AUC of 0.63 (0.57–0.69) in the HK cohort. The HK-trained model performed better in the HK cohort (AUC 0.71, 0.65–0.77) than the US-trained model ($p < 0.001$) and had lower performance in the US cohort (AUC 0.54, 0.45–0.64, $p < 0.001$).

Table 4-3 Model Performance Across Cohorts

Cohort	Model	AUC (95%CI)	*p*-value
US	US-trained	0.76 (0.68 – 0.84)	0.841
US	HK-trained	0.54 (0.45 – 0.64)	**0.005**
US	GPR	0.73 (0.65 – 0.81)	ref
HK	US-trained	0.63 (0.57 – 0.69)	**<0.001**
HK	HK-trained	0.71 (0.65 – 0.77)	0.823
HK	GPR	0.70 (0.64 – 0.76)	ref
UK	US-trained	0.37 (0.27 – 0.48)	**0.003**
UK	HK-trained	0.50 (0.37 – 0.62)	0.099
UK	GPR	0.64 (0.52 – 0.76)	ref

Note: AUC values are reported with 95% CI in parentheses. P-values represent comparisons of each model with the GPR model (reference, "ref") within each cohort and were calculated using Bayesian bootstrap resampling with 10,000 iterations.

Abbreviations: US = United States; HK = Hong Kong; UK = United Kingdom; GPR = Generalised Patellofemoral Radiomics; AUC = Area Under the Receiver Operating Characteristic Curve; CI = Confidence Interval.

To address the unmet need of generalisability of our model, we trained the GPR model on intersecting features from the US and HK cohorts. Selecting 9 features for the GPR model reached the best performance (Figure 4-4). These features were primarily first-order image intensity descriptors (Table 4-4). The GPR model achieved an AUC of 0.73 (0.65–0.81) in the US cohort, similar to the US-trained model (AUC 0.76, $p = 0.841$) and higher than the HK-trained model (AUC 0.54, $p = 0.005$). In the HK cohort, the GPR model achieved an AUC of 0.70 (0.64–0.76), comparable to the HK-trained model ($p = 0.823$) and higher than the US-trained model ($p < 0.001$). In the UK cohort, the

GPR model reached an AUC of 0.64, compared to 0.37 for the US-trained model (p = 0.003) and 0.50 for the HK-trained model (p = 0.099).

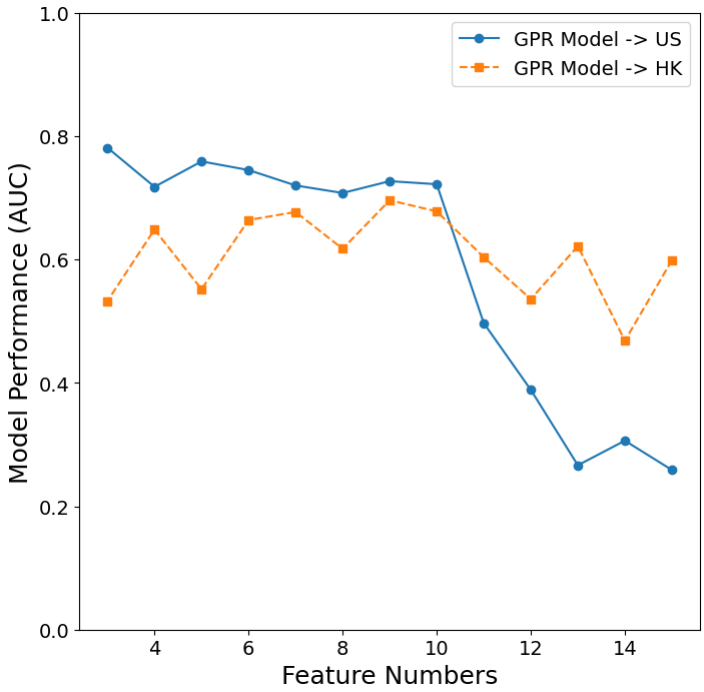


Figure 4-4 GPR model's performance on the US cohort and HK cohort using different feature sets. Abbreviations: US = United States; HK = Hong Kong; GPR = Generalised Patellofemoral Radiomics; AUC = Area Under the Receiver Operating Characteristic Curve.

Table 4-4 Radiomics Features Summary

Rank	US Cohort			HK Cohort			Intersection		
	Image	**Class**	**Name**	**Image**	**Class**	**Name**	**Image**	**Class**	**Name**
1	**Original**	**First order**	**Median**	Original	First order	90 percentile	Original	First order	Root mean squared
2	Wavelet (HL)	GLSZM	Small area high gray level emphasis	Wavelet (LH)	GLCM	Difference entropy	Wavelet (LL)	First order	Mean
3	**Wavelet (LL)**	**First order**	**Mean**	Wavelet (LL)	First order	90 percentile	Original	First order	10 percentile
4	**Wavelet (LL)**	**First order**	**Median**	**Wavelet (LL)**	**First order**	**Root mean squared**	Wavelet (LL)	First order	Root mean squared

5	Wavelet (HH)	GLRLM	Long run low gray level emphasis	Wavelet (LL)	First order	Maximum	Wavelet (LL)	First order	Median
6	**Original**	**First order**	**Mean**	**Original**	**First order**	**Root mean squared**	Wavelet (LL)	First order	10 percentile
7	**Original**	**First order**	**Root mean squared**	Wavelet (LH)	GLCM	Correlation	Original	First order	Mean
8	**Wavelet (LL)**	**First order**	**Root mean squared**	**Wavelet (LL)**	**First order**	**Mean**	Original	First order	Median
9	Original	First order	Energy	Original	First order	Maximum	Original	GLCM	Joint average
10	LoG (σ = 3mm)	First order	Median	**Original**	**First order**	**Mean**			
11	Original	GLSZM	Gray level variance	**Original**	**First order**	**Median**			
12	LoG (σ = 3mm)	First order	Mean	wavelet-HL	GLCM	Correlation			
13	Wavelet (LL)	First order	Energy	**Wavelet (LL)**	**First order**	**Median**			
14	LoG (σ = 2mm)	First order	Mean	LoG (σ = 1mm)	First order	Maximum			
15	Wavelet (LL)	First order	Total energy	Original	First order	Mean absolute deviation			
16	Wavelet (LL)	GLCM	Sum average	LoG (σ = 3mm)	First order	Minimum			
17	Original	First order	Total energy	Wavelet (LL)	GLSZM	Gray level non uniformity			
18	Wavelet (LL)	GLCM	Joint average	Wavelet (LL)	First order	Mean absolute deviation			
19	**Wavelet (LL)**	**First order**	**10 percentile**	LoG (σ = 3mm)	First order	Energy			
20	Original	GLCM	Sum average	**Wavelet (LL)**	**First order**	**10 percentile**			
21	Wavelet (LH)	GLCM	Cluster shade	**Original**	**GLCM**	**Joint average**			
22	Wavelet (LL)	GLCM	Auto correlation	**Original**	**First order**	**10 percentile**			
23	**Original**	**First order**	**10 percentile**	LoG (σ = 3mm)	First order	10 percentile			

24	**Original**	**GLCM**	**Joint average**	wavelet-HH	GLCM	Correlation	

Note: The radiomics features in bold are duplicated in the US and HK, which have been included in the intersection. For a detailed description and comprehensive list of all radiomic features analysed in this study, readers are directed to the PyRadiomics documentation available at https://pyradiomics.readthedocs.io/en/latest/features.html. This documentation provides extensive information on each feature's calculation and theoretical background, ensuring a thorough understanding and facilitating the reproducibility of our analyses (26).

Abbreviations: US = United States; HK = Hong Kong; LoG = Laplacian of Gaussian; GLCM = Gray Level Co-occurrence Matrix; GLSZM = Gray Level Size Zone Matrix; GLRLM = Gray Level Run Length Matrix; First order = First-order statistics; LL, LH, HL, HH = Low- and high-frequency components from wavelet decompositions; σ = Standard deviation (for LoG filter kernel).

4.3.3 Model interpretation

To develop the GPR Score, we applied the best-performing GPR model utilising 9 selected features. The GPR Score had the strongest correlations with lateral patellar cartilage ($\rho = 0.344$) and lateral anterior femoral cartilage ($\rho = 0.321$) on WORMS, indicating alignment with patellofemoral disease as a continuous measure (Figure 4-5(A), Figure 4-6(A)). In contrast, KL Grade correlated most strongly with medial femoral and tibial cartilage WORMS scores.

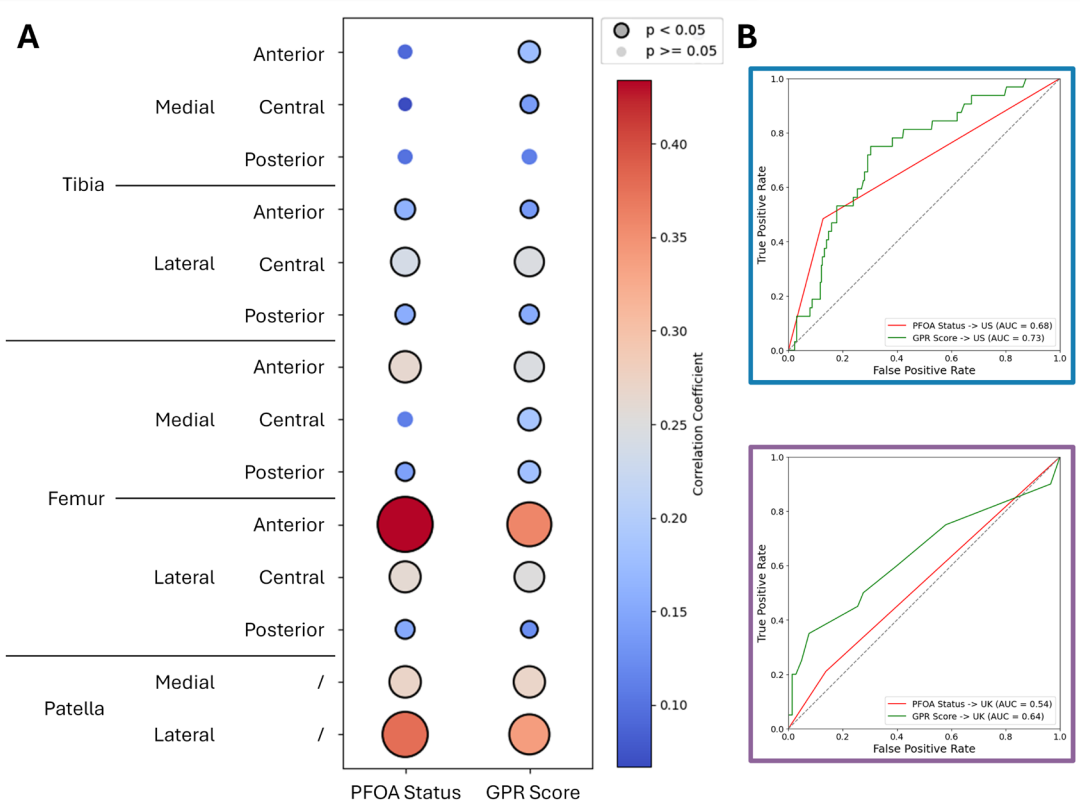


Figure 4-5 Comparison Between GPR Score and PFOA Status.

Note: (A) Coloured bubble plot illustrating Spearman correlations of GPR Score and PFOA status with MRI-based WORMS cartilage morphology subregions. Colour intensity indicates the strength of correlation (Spearman's ρ), and bubble size corresponds to statistical significance (p-value). Bubbles with black contours indicate statistically significant correlations ($p < 0.05$).

(B) ROC curves comparing the performance of GPR Score and PFOA status in predicting KR across cohorts. Upper panel: US cohort; Lower panel: UK cohorts.

Abbreviations: GPR = Generalised Patellofemoral Radiomics; PFOA = Patellofemoral Osteoarthritis; WORMS = Whole-Organ Magnetic Resonance Imaging Score; ROC = Receiver Operating Characteristic; AUC = Area Under the Receiver Operating Characteristic Curve; KR = Knee Replacement.

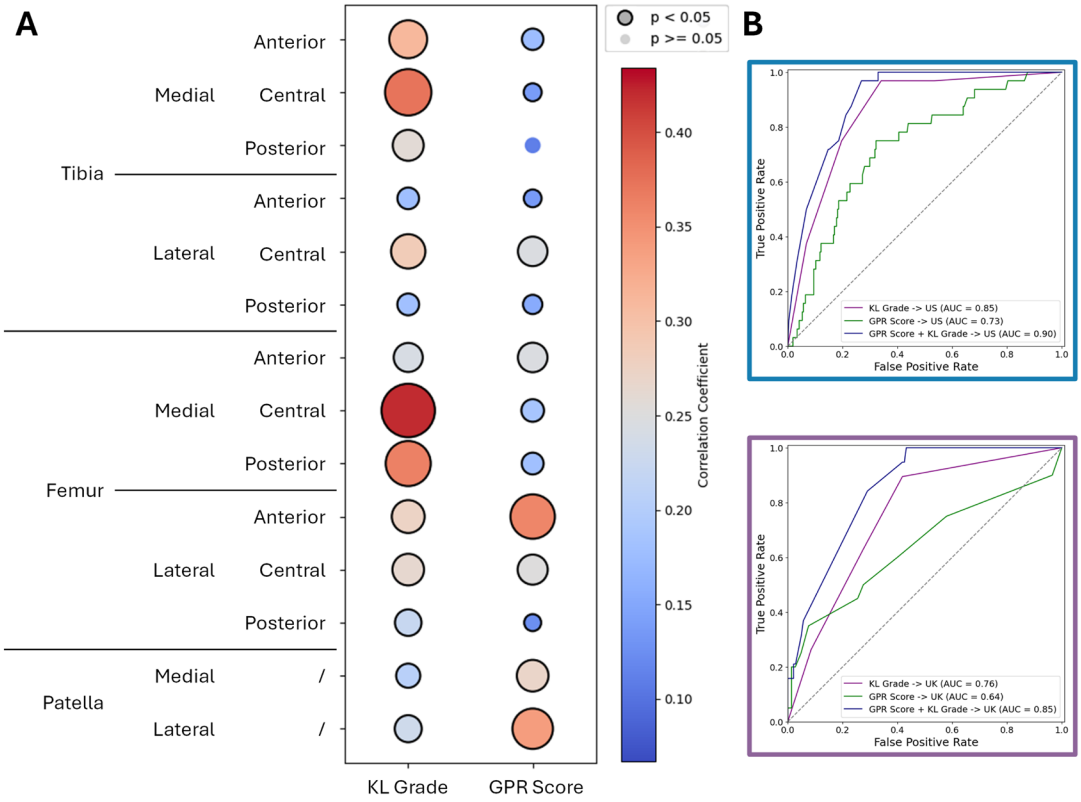


Figure 4-6 Relationship Between GPR Score and KL Grade.

Note: (A) Coloured bubble plot illustrating Spearman correlations of GPR Score and KL Grade with MRI-based WORMS cartilage morphology subregions. Colour intensity indicates the strength of correlation (Spearman's ρ), and bubble size corresponds to statistical significance (p-value). Bubbles with black contours indicate statistically significant correlations ($p < 0.05$).

(B) ROC curves comparing the performance of GPR Score, KL Grade, and their combination (GPR Score + KL Grade) in predicting KR across cohorts. Upper panel: US cohort; Lower panel: UK cohorts.

Abbreviations: GPR = Generalised Patellofemoral Radiomics; KL = Kellgren–Lawrence; WORMS = Whole-Organ Magnetic Resonance Imaging Score; ROC = Receiver Operating Characteristic; AUC = Area Under the Curve; KR = Knee Replacement.

Comparison of the GPR Score and PFOA status for predicting KR is shown in Figure 4-5(B). In the US cohort, the GPR Score had an AUC of 0.73, compared to 0.68 (0.59–

0.77) for PFOA status (p = 0.201). In the UK cohorts, the GPR Score had an AUC of 0.64, compared to 0.54 (0.44–0.64) for PFOA status (p = 0.074).

When combined with existing KL Grade information, it appeared that integrating information from both the PF and TF joints improved model performance (Figure 4-5(B)). The combined AUC increased to 0.90 (0.86–0.94) in the US cohort (vs. 0.73 for GPR Score, p = 0.003) and to 0.85 (0.78–0.91) in the UK cohort (vs. 0.64 for GPR Score, p = 0.013). The combined model and KL Grade alone yielded similar results (US: 0.90 vs. 0.85, 0.79–0.91, p = 0.159; UK: 0.85 vs. 0.76, 0.64–0.85, p = 0.175).

4.4 Discussion

In this study, we demonstrated the feasibility of a domain-adapted radiomics pipeline (the GPR model) to provide robust and generalisable predictions of knee OA progression across cohorts from the US, HK, and the UK. The GPR Score maintained high predictive accuracy in both internal and external testing, addressing domain gaps due to variations in population characteristics, clinical practice, and imaging protocols (85, 86). The straightforward and feasible nature of our approach—selecting shared, transferable radiomics features—makes it practical for wider adoption. Moreover, the use of ethnically and geographically diverse populations further enhances the external validity and potential real-world applicability of this approach.

4.4.1 Strength beyond the PFOA

Our GPR Score offers several potential advantages compared to traditional PFOA grading. First, it provides an automated method of assessment directly from standard X-rays, potentially reducing the reliance on specialist interpretation and easing

healthcare access, especially in settings with limited resources. Second, as a continuous numerical measure, the GPR Score may offer increased sensitivity and precision in capturing subtle structural changes over time, though this remains to be confirmed through further longitudinal validation studies. Third, its objective, quantitative nature, derived from radiomic features, might help reduce subjectivity and inter-reader variability common with manual grading systems, thus potentially enhancing reproducibility. Collectively, these characteristics suggest the GPR Score could be valuable, particularly in large-scale epidemiological research or clinical contexts where consistent expert evaluation is challenging.

4.4.2 Synergistic effect with KL grade

The synergy between the GPR Score and KL Grade suggests that combining the measures for both compartments of the knee joint provides a more comprehensive assessment than using either alone. The correlation between GPR Score and WORMS further supports its pathological relevance. Interestingly, KL Grade demonstrated stronger associations with medial joint structures, aligning with its historical emphasis on TFOA progression. The GPR Score showed a greater correlation with lateral cartilage changes, particularly in the patella and lateral anterior femur, reflecting its ability to capture distinct aspects of OA pathology. This lateral focus is clinically meaningful, as lateral PF degeneration is more prevalent and more strongly associated with pain and disability. It is commonly driven by modifiable factors such as maltracking, valgus alignment, lateral patellar tilt, and increased lateral joint loading (87, 88). When combined with KL Grade, the GPR Score significantly improved model performance, outperforming the GPR Score alone. Although differences compared to KL Grade alone were not statistically significant, likely due at least in part to limited

sample size, the combined model highlights the complementary value of integrating PF radiomics with TF grading for a more comprehensive OA risk assessment. It may represent an opportunity to add TF radiomics to such a model instead, to harness the advantages of a continuous, objective, automated measure.

4.4.3 Limitations of this study

This study has several limitations. First, while the study included geographically diverse cohorts, it did not capture South Asian, African or other populations, which may limit the generalisability of the findings across all ethnic groups. Second, the sample size in certain analyses, particularly in subgroup and external validation comparisons, was limited, which may reduce statistical power and increase uncertainty around model estimates. Moving forward, prospective studies are needed to validate this approach in larger, more diverse global cohorts and to evaluate its performance prospectively. Furthermore, while KL Grade provided a useful measure of TF involvement, a future radiomics-based pipeline integrating both PF and TF compartments could offer a more comprehensive evaluation, further enhancing clinical prediction and management strategies.

4.5 Chapter summary

In conclusion, by addressing the domain gaps prevalent in radiomics and machine learning research, our domain-adapted GPR model demonstrates feasibility and practicality for predicting knee OA progression in diverse, real-world clinical settings. This automated, objective radiograph-based biomarker offers a promising approach to stratify OA progression risk, potentially improving patient management and optimising healthcare resources globally.

Chapter 5: Deep-Learning Radiomics Analysis of the Tibiofemoral Joint: Improved Assessment of Knee OA Severity and Progression

5.1 Chapter overview

Radiomics has gained substantial traction in musculoskeletal imaging for its potential to extract high-dimensional features from standard medical images and to convert these features into clinically meaningful insights. Traditionally, radiomics pipelines have relied on handcrafted features—quantitative descriptors of image intensity, shape, and texture—paired with conventional machine learning models to perform classification or regression tasks. While effective in several contexts, this approach has limitations, especially when applied to anatomically complex and heterogeneous diseases such as knee OA.

The TF joint, the primary load-bearing compartment of the knee, presents a unique challenge for radiomics analysis. The morphological variations across patients, compounded by differences in OA phenotypes and radiographic presentation, demand an approach capable of capturing subtle, nonlinear relationships within the image data. Handcrafted features, although interpretable, may not fully encapsulate the joint's complex structural patterns. Moreover, the performance of such models is often constrained by the assumptions and simplifications inherent in manually defined features.

Deep learning, and specifically CNNs, offers a powerful alternative. Unlike traditional radiomics, deep learning models do not rely on predefined image descriptors. Instead,

they learn hierarchical representations of the image data directly from pixel-level inputs. This allows for the discovery of abstract and context-specific imaging biomarkers that may not be apparent to human observers or manually encoded. CNNs are particularly well-suited for image-based tasks due to their ability to preserve spatial relationships and recognise local patterns such as edges, textures, and anatomical contours.

In recent years, deep learning has demonstrated considerable success in medical image classification, segmentation, and prognosis prediction. Its application in OA, however, is still emerging. Studies have begun to explore CNNs for predicting OA presence or progression from knee radiographs, often showing superior performance compared to traditional methods. Nevertheless, many of these models are limited by a lack of interpretability, constrained generalizability, and a narrow focus on binary classification or KL grading replication.

In this chapter, we explore the use of a deep learning-based radiomics framework tailored to the assessment of TF joint severity and progression in knee OA. The model is designed to predict longitudinal outcomes rather than replicate existing grading systems, thereby shifting the focus from diagnostic labelling to personalised prognostication. Unlike handcrafted approaches, our method leverages the entire radiographic context of the joint, including implicit spatial patterns and morphological asymmetries that may be indicative of early or advancing disease.

Through this work, we aim to develop a deep-learning-based radiomics score (DR Score) and demonstrate that deep learning radiomics can serve as a complementary or superior alternative to traditional methods in capturing the full complexity of knee joint degeneration. Furthermore, we investigate the clinical implications of such a model, particularly its utility in stratifying patients at high risk of progression and informing

earlier, more targeted interventions.

5.2 Methodology

5.2.1 Data sources and participants

We analysed data from two major US-based, multicentre, longitudinal cohort studies—the MOST and the OAI—for model training and internal testing. Two UK-based studies, the MenTOR study and the KICK study, were used for external validation.

Participants were excluded if they had less than 60 months of follow-up, lacked baseline radiographs, presented with visible KR at baseline, or if the radiographs failed to meet quality standards necessary for accurate automatic segmentation.

Following exclusions, participants from the US datasets were randomly allocated to training and testing sets in an 8:2 ratio, ensuring that both knees from a single subject remained within the same subset. All eligible participants from the UK-based studies were used for external testing. Unlike the US-based studies, where both knees were analysed, only one index knee per subject was selected in the UK-based studies, based on the higher KL grade. This is because symptom progression in the UK-based studies was assessed at the subject level rather than separately for each knee. An overview of participant inclusion and exclusion across cohorts is summarised in Figure 5-1.

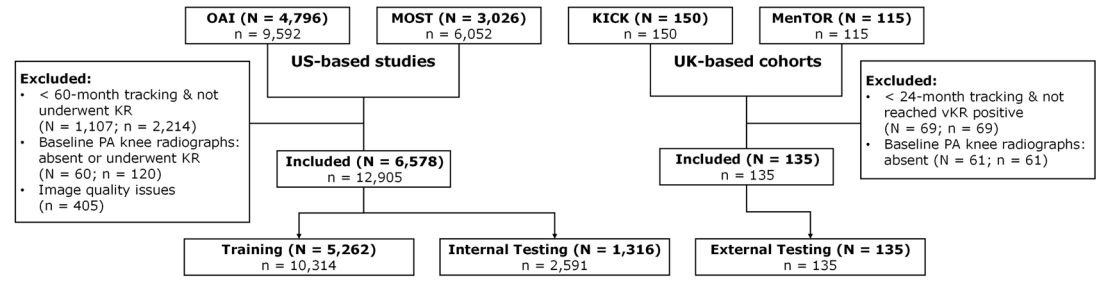


Figure 5-1 Flow diagram illustrating the inclusion and exclusion process of participants in the study.

Note: OAI = Osteoarthritis Initiative, MOST = Multicenter Osteoarthritis Study, KICK = Knee Injury Cohort at the Kennedy, MenTOR = Meniscal Tear and Osteoarthritis Risk, US = United States, UK = United Kingdom, PA = poster-anterior, BMI = Body Mass Index, KL = Kellgren-Lawrence, WORMS = Whole-Organ Magnetic Resonance Imaging Score, KR = knee replacement, KOOS = Knee Injury and Osteoarthritis Outcome Score, vKR = virtual knee replacement.

5.2.2 Exposures and imaging acquisition

Radiographs were obtained at the participant's baseline visit (or the earliest available visit after enrolment). All four cohorts (OAI, MOST, KICK, and MenTOR) acquired bilateral standing posteroanterior semi-flexed (Rosenberg) knee radiographs to assess structural osteoarthritis.

In OAI, MOST, and MenTOR, standardised positioning devices (SynaFlexer or plexiglass frame) were used to fix knee flexion and foot rotation while maintaining equal weight bearing. The x-ray tube was angled caudally (typically 10°, calibrated or adjusted when necessary) and centred at the tibiofemoral joint line to optimise tibial plateau alignment. MOST and MenTOR allowed minor angle adjustments (5°–15°) to improve tibial plateau superposition. In contrast, the KICK study used the Rosenberg view without a positioning frame or fluoroscopy. Radiographers relied on a foot

template and asked participants to flex comfortably, resulting in less standardised positioning across individuals.

Radiographs from all cohorts were used for KL grading. A non-specialist quality control review removed unavailable, mislabelled, or poor-quality images at upload. The remaining radiographs formed the final dataset for analysis.

5.2.3 Covariates and baseline measures

Baseline demographic variables (age, sex, and BMI) were collected for all participants. Structural severity of knee OA was assessed using the tibiofemoral KL grade (0–4) as previously assigned by each cohort (89). In a subset of MOST participants, additional clinical and imaging data were available, including ligament repair history and Whole-Organ Magnetic Resonance Imaging Score (WORMS) assessments at baseline (20).

5.2.4 Outcomes

The primary outcome was the time to a genuine KR event for US cohorts or a vKR for UK cohorts. Both total and partial knee replacements were considered. In the US-based studies, both KR events and time-to-event information were identified from medical records, postoperative imaging, and participant self-reports. In the UK-based studies, where few surgical events were recorded due to the younger age of participants, vKR status and corresponding event times were estimated based on symptom progression criteria (82). Further details regarding vKR criteria are provided in the Appendix.

5.2.5 Deep Learning Radiomics analysis

<u>Overview of the Framework</u>

The proposed radiomics pipeline is fully automated, from initial bounding box segmentation to the final output of the Deep-learning-based Radiomics Score (DR Score). The workflow consists of automatic detection of the knee joint, extraction of imaging features, attention-based feature aggregation, and DR Score generation (Figure 5-2).

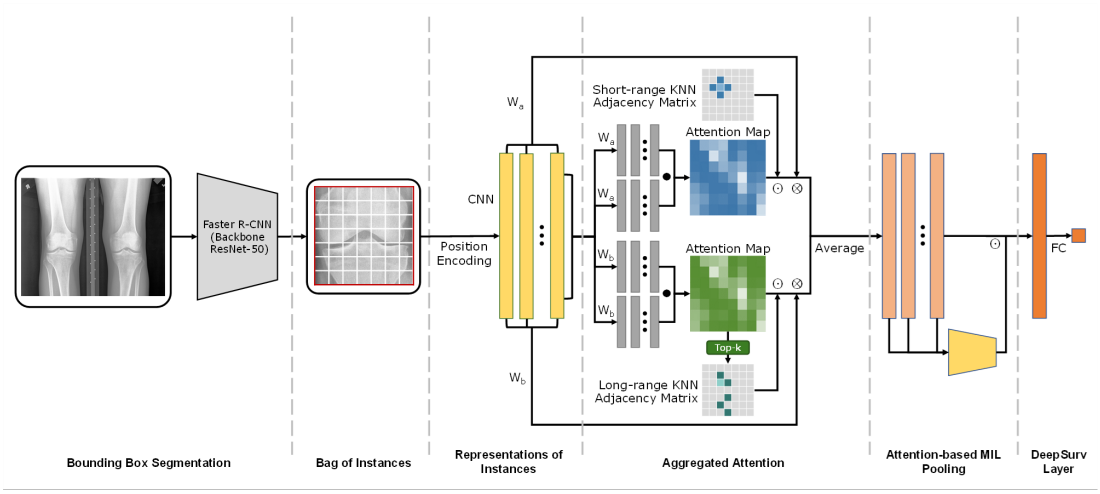


Figure 5-2 Deep learning radiomics framework pipeline.

Note: The process includes bounding box segmentation with Faster R-CNN, followed by dividing the region into a bag of instances. Features are extracted using a CNN with position encoding and processed through short-range and long-range aggregated attention. The outputs are pooled via traditional MIL attention pooling and passed through DeepSurv layers to generate the DR Score.

Abbreviations: Faster R-CNN = faster region-based convolutional neural network, CNN = convolutional neural network, KNN = top-k nearest neighbours, MIL = multiple instance learning, FC = fully connected.

Automatic Segmentation of Knee Joint (Pretraining)

To enable fully automated analysis, we fine-tuned a pretrained Faster Region-based Convolutional Neural Network (Faster R-CNN) model with a ResNet-50 backbone and

Feature Pyramid Network architecture (90) for knee joint localisation on radiographs. The original Faster R-CNN model, pretrained on a general object detection dataset, was adapted for knee imaging by additional training using 100 manually annotated bilateral knee radiographs (200 knees). Manual annotations were performed using ITK-SNAP (version 4.0.2) (91). We enforced standardised rectangular bounding boxes aligned with the knee's left and right contours to maintain consistency in ROI dimensions across images. After fine-tuning, the model was used to automatically detect and extract ROIs from all radiographs. Preprocessing included grayscale conversion, contrast enhancement (via histogram equalisation), and standardisation to left-knee orientation to ensure consistency prior to analysis.

Feature Extraction

Each extracted ROI was divided into non-overlapping patches. Each patch was independently processed using a ResNet-18 CNN (92), which extracted high-dimensional feature vectors capturing the local visual characteristics of each patch. To preserve spatial relationships, sinusoidal positional encodings were added to each feature vector, enabling the model to distinguish patches based on anatomical location as well as content (93).

Attention-based Aggregation

We applied both short- and long-range self-attention mechanisms to capture relevant local and global features.

Short-range attention focused on aggregating information from neighbouring patches within a predefined physical distance on the radiograph. For each patch, an adjacency

matrix was constructed to define connections to its immediate neighbours (94). Through a self-attention mechanism, the model learned to assign higher weights to neighbouring patches with similar features, allowing enhanced focus on continuous anatomical structures, such as joint space narrowing.

Long-range attention captured relationships between patches that were distant in physical space but exhibited similar visual features (95). This mechanism enabled the model to detect and relate structural abnormalities that may appear across different regions of the knee, such as widespread osteophyte formation or ligament changes.

The outputs were averaged to form unified feature vectors that integrated detailed and contextual imaging information (96). This dual-attention strategy enhanced the model's ability to understand complex anatomical relationships across the knee joint. An illustration of these attention mechanisms is provided in the first row of Figure 5-3.

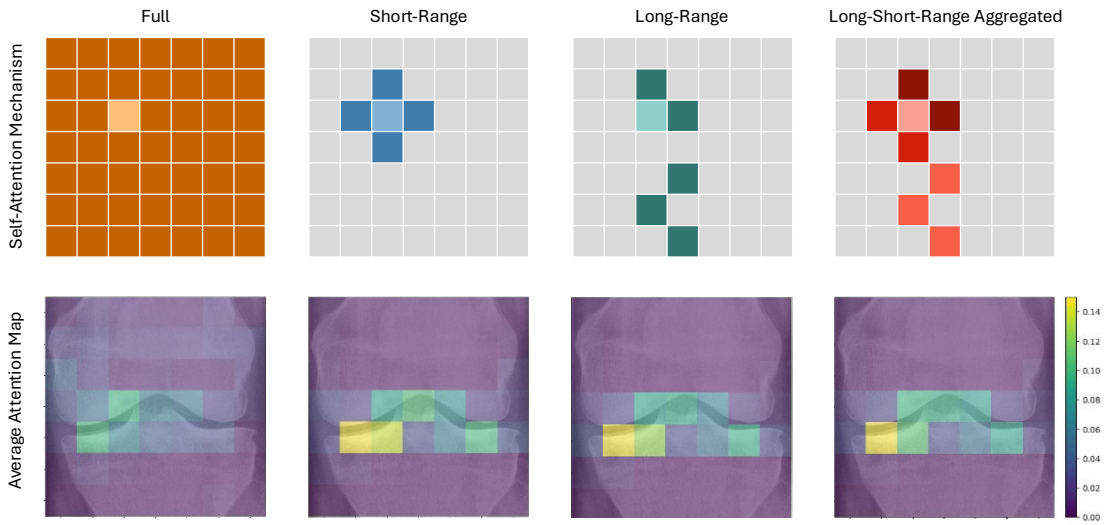


Figure 5-3 Comparison of different self-attention mechanisms.

Note: The figure illustrates traditional full self-attention, short-range self-attention, long-range self-attention, and long-short-range aggregated self-attention mechanisms. The bottom row shows the corresponding attention maps, overlaid on the knee joint, with weights highlighted. The

average attention map demonstrates that long-short-range aggregated self-attention focuses more on specific areas, particularly around the joint space, compared to full self-attention. This targeted attention enhances the model's ability to capture relevant features related to joint health.

DR Score Computation

Patch-level features were aggregated into a patient-level prediction using an attention-based multiple instance learning (MIL) framework (97). This allowed the model to assign greater weight to patches most relevant to KR risk.

The aggregated feature vector was then passed through a DeepSurv layer (98), which integrates a Cox proportional hazards model within a deep learning architecture. This enabled the direct prediction of time-to-event outcomes—specifically, the likelihood and timing of KR surgery over the study follow-up period.

The final output was defined as the DR Score. To facilitate interpretation and comparability, DR Scores were rescaled to a standard range from 0 to 4.

5.2.6 Statistical Analysis

The predictive performance of the DR Score was assessed using survival analysis metrics. The C-index measured the model's ability to correctly rank time-to-KR outcomes. To evaluate overall discrimination across follow-up, the average AUC was calculated by deriving AUC values at multiple fixed time points (3, 6, 9, 12, …, 108 months) and computing their mean. External validation in the two UK cohorts used the same evaluation framework. For comparison, we evaluated three models: (1) DR Score alone, (2) KL grade alone, and (3) a combined model including both.

To interpret model predictions, attention heatmaps were generated to visualise image regions contributing most strongly to the DR Score. Patch interrelationship analysis was also performed to assess distinctiveness of high-attention patches.

The relationship between DR Score and KL grade at baseline was examined using Spearman's rank correlation. We further assessed predictive relevance by calculating the proportion of knees undergoing KR during follow-up with 95% confidence intervals (CIs).

To determine independent prognostic value, univariate and multivariate Cox proportional hazards regression analyses were performed, and hazard ratios (HRs) with 95% CIs were reported for demographic and radiographic variables.

To explore clinical and structural relevance, Spearman's correlation was used to assess associations of DR Score and KL grade with ligament repair history and WORMS-defined joint damage. Additionally, the Mann–Whitney U test was applied to compare DR Score and KL grade distributions across WORMS-based lesion severity subgroups.

5.2.7 Risk Stratification

A piecewise linear regression model was fitted to the DR Score versus proportion curve to identify potential inflexion points indicating changes in the rate of risk increase. Based on the fitted model, it was planned that two cutoff points would be selected to separate all US subjects into three progression groups: slow progressors, intermediate progressors and rapid progressors. The stratification performance of the DR Score was further assessed using Kaplan-Meier survival analysis, comparing time-to-KR between risk groups.

5.3 Results

5.3.1 Participants characteristics

Figure 5-1 presents a flow diagram outlining the inclusion and exclusion process for participant selection. The US-based studies initially included 7,822 participants (15,644 knees) at baseline. A total of 120 knees (from 60 participants) were excluded due to missing baseline knee radiographs. An additional 405 knees were excluded because of inadequate image quality, and 2,214 knees (from 1,107 participants) were excluded due to death, withdrawal of consent, or loss to follow-up. Ultimately, 12,905 knees from 6,578 participants were included for analysis. Among these, 10,314 knees (5,262 participants) were allocated to the training set, and 2,591 knees (1,316 participants) to the internal testing set.

For the UK-based studies, a total of 265 participants were enrolled. Exclusions included 61 participants with missing baseline radiographs and 69 participants who withdrew consent or were lost to follow-up, resulting in 135 participants being included in the external testing set.

Table 5-1 summarises baseline demographic and clinical characteristics across the training, internal testing, and external testing sets, including age, sex, BMI, KL grade. The frequency of KR and vKR events is also reported. All variables were generally comparable between the training and internal testing sets. Variables were generally well balanced between the training and internal testing sets. As expected, external testing participants differed in age, BMI, and, to a lesser extent, sex due to demographic and clinical variation, but showed comparable distributions in KL grade and KR/vKR event frequency.

Table 5-1 Summary statistics for demographic variables of the study participants.

Parameters		Train	Internal Test	External Test	*p*-value (Train vs Test)	*p*-value (Train vs External)
Participants n (knees n)		5,262 (10,314, 79.9%)	1,316 (2,591, 20.1%)	135 (135)		
Age		61.4 (45 to 79)	61.1 (45 to 79)	35.2 (17 to 60)	0.231	**< 0.001**
Sex	Male	2,088 (39.7%)	537 (40.8%)	102 (75.6%)	0.456	0.072
	Female	3,174 (60.3%)	779 (59.2%)	33 (24.4%)		
BMI		29.42 (16.90 – 71.91)	29.37 (16.72 – 60.06)	26.74 (18.94 – 49.90)	0.761	**< 0.001**
KL grade (knees)	0	4,185 (40.5%)	1,045 (40.3%)	74 (54.8%)	0.219	0.382
	1	1,811 (17.6%)	420 (16.2%)	22 (16.2%)		
	2	2,237 (21.7%)	607 (23.4%)	28 (20.7%)		
	3	1,493 (14.5%)	357 (13.7%)	10 (7.4%)		
	4	513 (5.0%)	131 (5.1%)	0 (0%)		
	Missing	75 (0.7%)	31 (1.2%)	1 (0.7%)		
KR event (knees)	Yes	943 (9.1%)	247 (9.5%)	16 (11.9%, vKR)	0.540	0.565
	No	9,371 (90.9%)	2,344 (90.5%)	119 (88.1%, vKR)		

Note: Data are presented as mean (range) for continuous variables and frequency (percentage) for categorical variables. Statistical comparisons between groups were conducted using t-tests for continuous variables and chi-square tests for categorical variables. Significance was determined at

a p-value of < 0.050. BMI = body mass index, KL = Kellgren-Lawrence grade, KR = knee replacement, vKR = virtual knee replacement.

5.3.2 Development and evaluation of DR Score

During model optimisation, various attention mechanisms were compared. The proposed long–short range aggregation strategy demonstrated superior performance over conventional full self-attention, as summarised in Table 5-2 and Table 5-3. The DR Score demonstrated strong predictive performance for KR in the internal testing set derived from the US-based studies. It achieved a C-index of 0.849 and an AUC of 0.885 over a nine-year follow-up.

Table 5-2 Performance of full-attention model in different patch size

Patch Size	C-index (mean ± SD)	Average AUC (mean ± SD)
5 × 5	0.806 ± 0.003	0.847 ± 0.002
6 × 6	0.821 ± 0.002	0.855 ± 0.007
7 × 7	**0.825 ± 0.005**	**0.861 ± 0.006**
8 × 8	0.819 ± 0.012	0.863 ± 0.014
9 × 9	0.813 ± 0.013	0.853 ± 0.008

Table 5-3 Performance of long-short-range model in different settings

C-index		Long Range		
(mean ± SD)		3	**5**	7
Short Range	0	0.842 ± 0.003	0.844 ± 0.004	0.843 ± 0.003
	1	0.842 ± 0.003	**0.849 ± 0.001**	0.844 ± 0.001
	2	0.843 ± 0.003	0.848 ± 0.002	0.844 ± 0.004

Average AUC (mean ± SD)		Long Range		
		3	**5**	7
Short Range	0	0.876 ± 0.008	0.876 ± 0.003	0.878 ± 0.004
	1	0.878 ± 0.003	**0.885 ± 0.000**	0.878 ± 0.001
	2	0.877 ± 0.005	0.885 ± 0.002	0.881 ± 0.003

Visualisation through average heatmaps revealed that our model primarily focused on the medial joint space and moderately focused on the ligament region and lateral joint space, as shown in the second row of Figure 5-3. This pattern indicates that our model identified and emphasised the most relevant anatomical features in its assessments, particularly the joint space between the femur and tibia, as shown in Figure 5-4(A). We numbered the patches from 0 to 48, as illustrated in Figure 5-4(B), and analysed their interrelationships. We found that patches with high self-attention values were the most unique, exhibiting the lowest interrelationship with other patches, as shown in Figure 5-4(C).

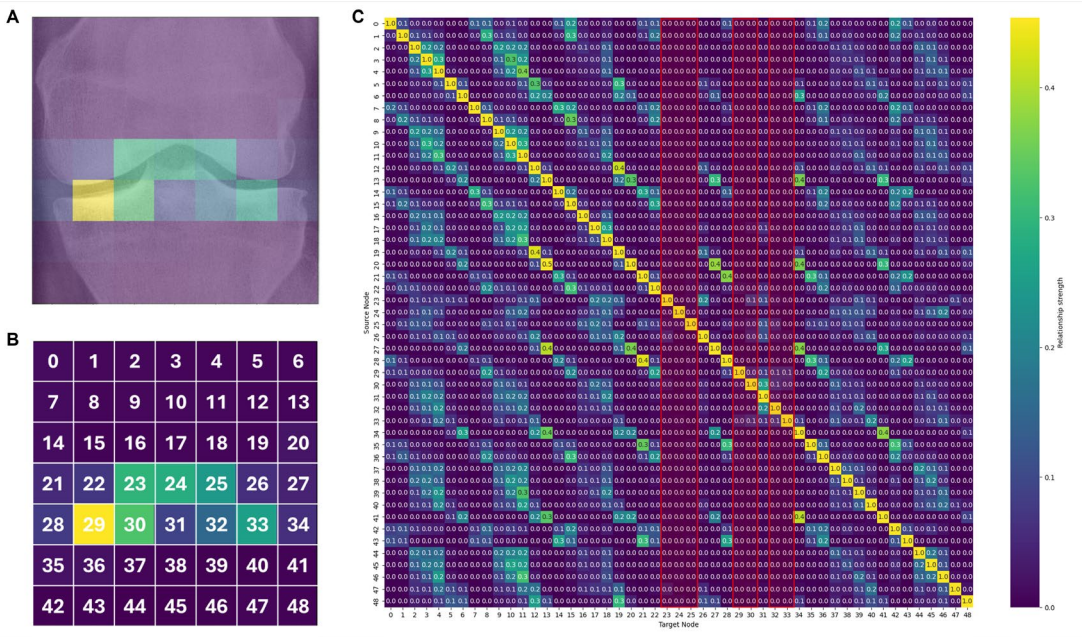


Figure 5-4 Analysis of Model Self-Attention and Patch Interrelationship.

Note: (A) Heatmap showing the model's focus (Bright yellow indicates greater attention, while dark blue represents lesser focus). (B) Numbered patch grid (0-48) used to analyse the interrelationships between patches. (C) Patches with high self-attention values (marked in red) were found to be the most unique, exhibiting the lowest interrelationship with other patches.

5.3.3 Relationship Between DR Score and KL Grade

Independence of DR Score

As shown in Table 5-4, all demographic and radiographic variables were significantly associated with the risk of knee replacement in univariate Cox regression analyses across both the training and testing sets. In multivariate analyses, however, only KL grade ($p < 0.001$) and the DR Score ($p < 0.001$) remained independent prognostic factors in both sets. Sex approached significance but did not reach statistical threshold (p = 0.060). Notably, the HR associated with the DR Score was substantially higher than that of KL grade (4.53 vs. 1.36), suggesting superior predictive strength for progression to knee replacement surgery.

Table 5-4 Results of univariate and multivariate analyses examining the association between predictor variables and the outcome of interest.

	Univariate		**Multivariate**	
	HR (95CI)	***p*-value**	**HR (95CI)**	***p*-value**
Train				
Sex	0.65 (0.54 – 0.78)	**< 0.001**	0.64 (0.53 – 0.77)	**< 0.001**
Age	1.04 (1.03 – 1.05)	**< 0.001**	1.01 (1.00 – 1.02)	0.084
BMI	1.06 (1.05 – 1.07)	**< 0.001**	1.01 (1.00 – 1.02)	0.097

KL Grade	1.66 (1.61 – 1.72)	**< 0.001**	1.30 (1.23 – 1.38)	**< 0.001**
DR Score	5.04 (4.32 – 5.89)	**< 0.001**	3.34 (2.84 – 3.94)	**< 0.001**
Test				
Sex	0.61 (0.42 – 0.89)	**0.010**	0.69 (0.47 – 1.02)	0.060
Age	1.04 (1.01 – 1.06)	**0.001**	0.99 (0.97 – 1.01)	0.428
BMI	1.05 (1.03 – 1.08)	**< 0.001**	1.00 (0.98 – 1.03)	0.991
KL Grade	1.71 (1.60 – 1.81)	**< 0.001**	1.36 (1.20 – 1.56)	**< 0.001**
DR Score	7.42 (5.08 – 10.80)	**< 0.001**	4.53 (3.06 – 6.70)	**< 0.001**

Note: Univariate and multivariate survival analyses were performed by Cox regression and results are presented as HR with 95% CI. Significance was determined at a p-value of < 0.050 (bolded). HR = hazard ratio, 95CI = 95% confidence interval, BMI = body mass index, KL Grade = Kellgren-Lawrence grade, DR Score = deep learning-based radiomics score

Agreement Between DR Score and KL Grade

As shown in Figure 5-5(A), Spearman's rank correlation between baseline DR Score and KL grade demonstrated a moderate positive association ($\rho = 0.696$, $p < 0.001$), indicating partial overlap in the structural severity captured by the two measures.

The proportion of knees undergoing knee replacement was plotted across increasing strata of DR Score and KL grade (Figure 5-5(B) and 6-5(C)). Both measures showed a positive association with surgical outcomes; however, the DR Score exhibited a more continuous and gradually increasing risk gradient, suggesting superior discriminative capacity across a broader spectrum of disease severity.

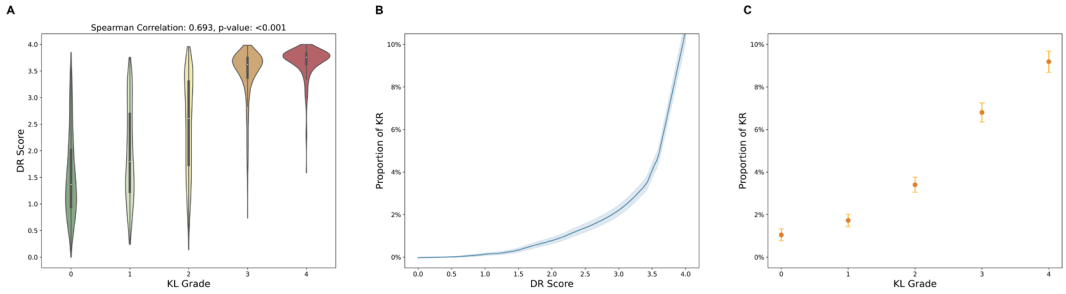


Figure 5-5 Comparison between DR Score and KL Grade in relation to knee replacement outcomes.

Note: (A) Spearman's rank correlation between baseline DR Score and KL grade. (B) Proportion of knees undergoing knee replacement across increasing DR Score strata. Shaded regions show 95% CIs. (C) Proportion of knees undergoing knee replacement across KL grade strata.

Abbreviations: DR Score = deep learning-based radiomics score, KL = Kellgren-Lawrence, KR = knee replacement.

Association with Clinical and MRI Features

As shown in Figure 5-6(A), the DR Score showed a stronger correlation than KL grade with ligament repair history (Spearman's $\rho = 0.517$, $p = 0.050$), suggesting greater sensitivity to post-traumatic changes.

Among participants with radiographic knee OA (KL grade ≥ 2), both the DR Score and KL grade exhibited moderate to weak correlations with WORMS cartilage and osteophyte features. Figure 5-6(B) presents the average and standard deviation of correlation coefficients across anterior, central, and posterior subregions of each compartment (20), revealing that the DR Score was more strongly associated with lateral compartment features compared to KL grade.

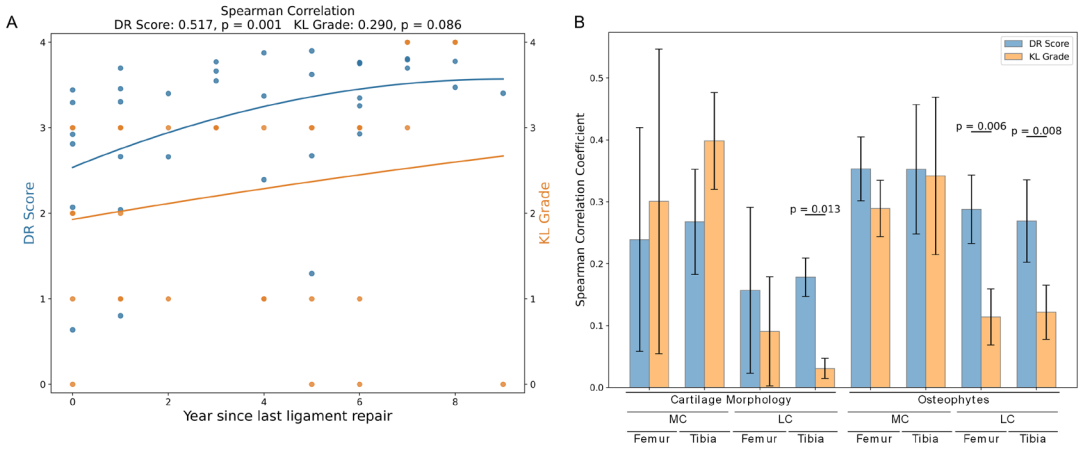


Figure 5-6 Associations between DR Score, KL grade, and clinical/MRI-based structural features.

Note: (A) Spearman's correlation between DR Score and ligament repair history. (B) Comparison of Spearman's correlation coefficients between DR Score and KL grade with WORMS-derived cartilage thickness and osteophyte features across anterior, central, and posterior subregions of each knee compartment.

Abbreviations: DR Score = deep learning-based radiomics score, KL Grade = Kellgren-Lawrence grade, MC = Medial Compartment, LC = Lateral Compartment

Additive Value of DR Score over KL Grade

As shown in Table 5-5, KL grade alone achieved a C-index of 0.821 (95% CI: 0.795–0.851) and an average AUC of 0.856 (0.830–0.880), whereas the DR Score alone significantly improved performance to a C-index of 0.849 (0.826–0.871) and an average AUC of 0.885 (0.861–0.907). The combined model yielded the highest discrimination, with a C-index of 0.863 (0.838–0.889) and an average AUC of 0.900 (0.876–0.923), also statistically superior to KL grade. It demonstrated the potentially synergistic value of combining both measures.

5.3.4 External Validation in UK-based studies

In external testing (Table 5-5), conducted in younger and demographically distinct UK-based cohorts using vKR as the outcome, the DR Score preserved clinically useful predictive accuracy (C-index 0.751; average AUC 0.791), comparable to KL grade (C-index 0.751; average AUC 0.776). Importantly, combining both measures further enhanced performance, achieving a C-index of 0.806 and an average AUC of 0.850, indicating complementary prognostic value beyond either measure alone.

Table 5-5 Comparative Performance Metrics of DR Score, KL Grade, and Combined Model

	Internal Test				External Test			
Model	**C-index (95% CI)**	**p-value**	**Average AUC (95% CI)**	**p-value**	**C-index (95% CI)**	**p-value**	**Average AUC (95% CI)**	**p-value**
KL Grade	0.821 (0.795 – 0.851)	ref	0.856 (0.830 – 0.880)	ref	0.751 (0.646 – 0.850)	ref	0.776 (0.682 – 0.865)	ref
DR Score	0.849 (0.826 – 0.871)	**0.037**	0.885 (0.861 – 0.907)	**0.015**	0.751 (0.606 – 0.877)	0.999	0.791 (0.664 – 0.903)	0.893
Combined Model	0.863 (0.838 – 0.889)	**<0.001**	0.900 (0.876 – 0.923)	**<0.001**	0.806 (0.681 – 0.918)	0.384	0.850 (0.740 – 0.951)	0.173

Note: Combined model composed of KL Grade and DR Score. C-index = concordance index, AUC = area under the receiver operating characteristic curve, KL = Kellgren-Lawrence, DR Score = deep learning-based radiomics score

5.3.5 Risk stratification and survival analysis

As shown in Figure 5-7(A), piecewise linear regression applied to the relationship between DR Score and the proportion of KR (Figure 3B) identified two inflexion points at DR Scores of 2.31 and 3.49. These cut-offs were used to stratify all US participants into slow (DR Score < 2.31), intermediate (2.31 ≤ DR Score < 3.49), and rapid (DR Score ≥ 3.49) progressors. Kaplan-Meier survival analysis (Figure 5-7(B)) demonstrated clearly separated risk trajectories across the three groups. Over the 108-

month (nine-year) follow-up, slow progressors (n = 7,447) had a cumulative KR incidence of 2.0%, intermediate progressors (n = 3,557) had 9.8%, and rapid progressors (n = 2,438) exhibited a markedly higher incidence of 36.3%.

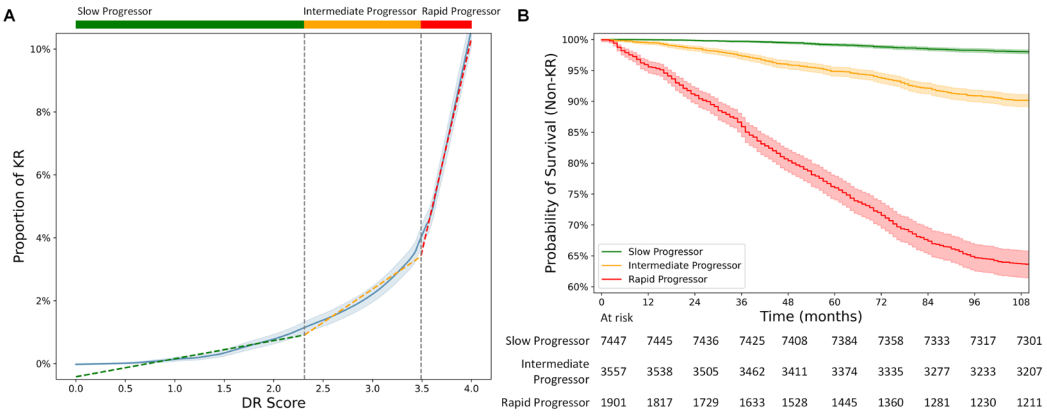


Figure 5-7 Risk stratification using the DR Score and corresponding survival outcomes.

Note: (A) Piecewise linear regression identified inflexion points at a DR Score of 2.31 and 3.49. (B) Kaplan–Meier survival curves for the slow, intermediate and rapid progressors.

Abbreviations: DR Score = deep learning-based radiomics score, KR = knee replacement

5.4 Discussion

This study developed and validated a DR Score to predict the risk of knee replacement using plain radiographs. While its performance was comparable to the established KL grade, the DR Score offers additional advantages as a continuous, fully automated measure that may capture subtle variation in disease severity and progression risk beyond categorical grading. This approach has the potential to enhance risk profiling, facilitate early identification of high-risk patients, and complement existing clinical assessments.

5.4.1 Comparative Performance and Generalisability

When compared with previous deep learning efforts, our model appears to demonstrate both better discrimination and broader generalisability. Earlier deep learning approaches for knee OA radiographs reported AUCs of 0.79 in population-based cohorts (13). In contrast, our model achieved an AUC of 0.885 in the OAI and MOST cohorts and maintained robust performance in two independent UK cohorts (AUC = 0.791). These findings highlight the potential additional value of the DR Score, which would suggest improved performance despite large, unselected cohorts with diverse imaging protocols and demonstrates consistent external validation despite distinct clinical settings, countries and patient populations.

5.4.2 Comprehensive Feature Representation Beyond KL Grading

The DR Score captures joint features beyond the medial tibiofemoral compartment targeted by KL grading, including the lateral compartment and intercondylar/cruciate ligament region, offering a more comprehensive view of pathology. Previous deep learning studies also noted attention to these regions—beyond what KL grading captures—but did not investigate their clinical relevance (99, 100). This broader focus could be beneficial because important structural changes in OA are not always confined to the medial side. For example, the lateral compartment may show greater heterogeneity in early or post-traumatic OA and is often affected by lateral meniscal injury or valgus/pivot-shift instability (101, 102). Similarly, cruciate ligament damage may reflect not only prior injury but also gradual attritional or degenerative changes that are not captured by KL grading or consistently documented in clinical history (103, 104). These may help explain why the DR Score maintained strong performance in younger UK cohorts. Overall, by integrating signals from multiple joint compartments,

the DR Score may capture clinically relevant aspects of disease that conventional grading or self-reported history could overlook.

5.4.3 Clinical Utility and Personalised Risk Stratification

From a clinical perspective, the continuous nature of the DR Score, unlike the categorical KL grade, allows more precise risk stratification and supports KR risk-aligned triage, enabling personalised treatment planning. In practice, individuals at higher risk could receive earlier interventions such as education, weight management, and exercise-based physiotherapy, along with closer monitoring and timely referral to orthopaedics or rheumatology when needed. Moreover, identifying distinct progression groups can further optimise surgical prioritisation, which is especially valuable in healthcare systems with long waiting times for KR, such as Hong Kong.

5.4.4 Implementation and Integration into Clinical Workflow

The DR Score would also be readily accessible for integration into routine clinical path. Knee radiographs are already widely ordered and archived in Picture Archiving and Communication Systems (PACS) systems, and our automated pipeline requires only the standard Digital Imaging and Communications in Medicine (DICOM) image, generating an interpretable output that could be automatically appended to radiology reports alongside KL grading. The addition of a continuous, reproducible score which may relate to clinical risk may improve reporting consistency across centres and reduce dependence on reader expertise. In regions where radiologist reporting is costly or delayed, the DR Score could further support timely decision-making. Importantly, the DR Score could be accompanied by attention heatmaps that highlight lesion-relevant regions, providing visual interpretability (much like outputs for vertebrae in a bone

density scan report) that can enhance clinician trust and further support appropriate clinical use, the implementation of which could be explored further with relevant stakeholder groups.

5.4.5 Limitations of this study

First, variability in image acquisition, positioning, and quality may affect model performance; however, training on heterogeneous US cohorts and validation in UK cohorts was intended to improve robustness across real-world settings. Ongoing monitoring and local calibration will still be important in implementation. Second, vKR was used in the UK cohorts due to the younger population and low surgical event rates. Although vKR is not identical to actual KR, it reflects clinically meaningful symptom progression and may capture appropriate KR need. Nevertheless, longer follow-up with actual KR events will be valuable to confirm these findings. Finally, the DR Score was derived from posteroanterior radiographs only. This modality was chosen because it is the most widely used screening view in clinical practice, supporting scalability; however, future work integrating lateral or multimodal imaging (e.g. MRI) may further enhance predictive performance.

5.5 Chapter summary

In conclusion, the DR Score provides a potentially clinically valuable, automated, and reproducible approach to predicting KR risk from radiographs. By offering a continuous output, broader anatomical sensitivity, and potential for integration into routine practice, it complements KL grading and may enable robust stratification across diverse populations. These findings highlight its potential as a scalable tool to support more personalised and risk-stratified care in knee OA.

Chapter 6: Multi-view radiomics: integration of patellofemoral and tibiofemoral features enhances prediction of knee OA progression

6.1 Chapter overview

Knee OA is a multifactorial and structurally heterogeneous disease that rarely manifests uniformly across the joint. Instead, degeneration typically affects specific anatomical compartments—most commonly the TF and PF joints—either independently or concurrently. Each compartment has distinct biomechanical functions, loading patterns, and disease trajectories. The TF joint bears the majority of axial load during gait and is often the primary focus of radiographic grading and clinical decision-making. In contrast, the PF joint is crucial for activities involving knee flexion, such as stair climbing or rising from a chair, and is increasingly recognized as a frequent site of early OA-related changes and pain.

Despite these compartment-specific characteristics, most existing radiographic OA assessments—such as the KL grading scale—provide a single global score, failing to capture the nuanced spatial distribution of joint degeneration. As a result, early signs of localized OA, particularly in the PF joint, may be overlooked or underestimated. Furthermore, radiomics-based studies to date have largely adhered to this compartmental separation, developing predictive models either for the TF or PF joint independently. While such approaches have demonstrated value in their respective contexts, they inherently assume that each compartment functions as an isolated predictive domain. This assumption overlooks potential synergistic interactions and structural interdependencies between compartments that may collectively influence OA

progression and patient outcomes.

The rationale for adopting a multi-compartment, or multi-view, radiomics approach lies in the hypothesis that integrating information from both the TF and PF joints provides a more comprehensive and biologically relevant representation of joint health. Structural abnormalities in one compartment may alter biomechanics and loading patterns in the other, potentially accelerating degeneration. For instance, cartilage loss or osteophyte formation in the PF joint may affect patellar tracking and increase stress on the medial TF compartment. Conversely, varus or valgus malalignment in the TF joint may secondarily alter PF joint mechanics. Capturing such inter-compartmental relationships requires a modelling framework capable of synthesizing features from multiple views and anatomical regions.

In addition to improving biological fidelity, multi-view radiomics may also enhance predictive performance. Independent PF and TF models may each capture unique features relevant to OA progression, but their predictive power may be incomplete when considered in isolation. A combined model has the potential to increase sensitivity and specificity by incorporating a broader array of imaging biomarkers, thus better reflecting the complex and multifaceted nature of the disease. This is particularly important for patients at intermediate KL grades or those with asymmetric disease patterns, where clinical decision-making is often uncertain.

Therefore, in this chapter, we propose a multi-view radiomics framework that integrates features extracted from both the PF and TF compartments to improve prediction of knee OA progression. This approach builds upon prior compartment-specific models developed in earlier chapters and addresses a critical methodological gap in OA imaging research. By evaluating whether integration leads to measurable improvements

in prognostic accuracy and clinical utility, this work aims to lay the groundwork for more holistic, patient-centred risk stratification strategies in knee OA management.

6.2 Methodology

6.2.1 Data Sources and Participants

This preliminary study utilised the MOST cohort, referred to here as the US cohort, and a combination of the MenTOR and KICK cohorts, collectively designated as the UK cohorts.

Participants were included if they had valid baseline radiographs from both PA and lateral views of the knee, enabling radiomics analysis of both the PF and TF compartments. To ensure reliable outcome classification, participants were required to have at least 60 months of follow-up or clear interim clinical documentation that allowed for definitive assessment of osteoarthritis progression status. Individuals with a history of knee replacement prior to baseline were excluded.

For each participant, the most affected knee—defined as the knee with the higher KL grade or the one more proximally related to the study endpoint—was selected as the index knee for analysis. If data for the initially selected knee were incomplete, the contralateral knee was used instead.

In this study, the US cohort was used solely for internal model evaluation, as both the GPR Score and the DR Score had already been independently developed and validated using this population in prior work. The UK cohort served as an external test set to further assess the generalisability and predictive value of the multi-view combined model, allowing direct comparison of individual and integrated radiomics scores across

different clinical populations.

6.2.2 Exposures and Imaging Acquisition

Both PA and lateral knee radiographs were used as the primary imaging sources for radiomics feature extraction. These two views provided complementary information on the TF and PF compartments, respectively, and were jointly analysed in the multi-view radiomics framework. All radiographs were taken at the baseline visit (or earliest available visit post-enrolment) and served as the exposures for predictive modelling.

In the US cohort, PA and lateral knee X-rays were acquired using standardized protocols. For the PA view, the Rosenberg technique was employed, with participants in a semi-flexed, weight-bearing position and knees positioned for optimal joint space visualization. The lateral view was also acquired in a semi-flexed, weight-bearing position with the knee aligned parallel to the Bucky and the foot placed against a plexiglass plate, ensuring clear visualization of the PF joint. These imaging protocols provided consistent and high-quality radiographs suitable for radiomics analysis.

In the UK cohorts, both PA and lateral radiographs were collected as part of routine clinical imaging. PA views were typically acquired using standing protocols, though weight-bearing status may have varied across clinical sites. Lateral views were generally non-weight-bearing, with participants lying on their side in a semi-flexed position. Standardized exposure settings, proper collimation, and superimposition of femoral condyles were applied to maintain image consistency across subjects. Nonetheless, some heterogeneity in positioning and acquisition technique remained, reflecting the variability encountered in real-world clinical environments.

For all cohorts, a non-specialist quality control procedure was applied. X-rays were excluded if the index knee was missing, mislabelled, or not clearly visible, or if the PF joint was not fully captured in the lateral view. Only knees with valid PA and lateral radiographs at baseline were included in the final radiographic dataset for analysis.

6.2.3 Covariates and baseline measures

Baseline covariates in this study included KL grade and PFOA status, both of which served as reference standards for evaluating the performance of the radiomics models. These covariates were assessed at the same baseline visit when the PA and lateral radiographs were obtained.

KL grade, an ordinal measure ranging from 0 to 4, was used to evaluate the structural severity of tibiofemoral joint osteoarthritis based on PA view radiographs. PFOA status was assessed from lateral view radiographs and recorded as a binary variable indicating the presence or absence of radiographic features of patellofemoral OA. Both metrics were determined through consensus readings by two experienced musculoskeletal radiologists as part of the original cohort studies.

Importantly, neither KL grade nor PFOA status was used as model input. Instead, they served as comparators in downstream analysis to benchmark the predictive performance of the radiomics-based scores (GPR and DR Scores) and their combination. This design allowed for an objective evaluation of whether imaging-derived quantitative features could outperform conventional expert-based grading systems in predicting disease progression.

6.2.4 Outcomes

The primary outcome of this study was the occurrence of KR or vKR within a 60-month follow-up period, treated as a binary variable (yes/no).

In the US cohort, KR events primarily referred to total knee replacements, as estimated by clinical collaborators. In the UK cohorts, where actual KR events were infrequent due to the younger age and earlier disease stage of participants, a validated algorithm was applied to derive vKR from longitudinal KOOS data (82). This approach allowed for the identification of knees with significant symptomatic deterioration indicative of eventual surgical need.

Full details on the vKR algorithm are provided in Appendix. No outcome-related data were used in model training or score calculation.

6.2.5 Radiomics Score Computation and Model Comparison Strategy

This study aimed to assess the added predictive value of integrating radiomic features from both PF and TF compartments in forecasting knee replacement risk. Two previously developed and independently validated radiomics-based scores were used:

GPR Score, derived from features extracted from lateral knee radiographs and trained on PF joint morphology.

DR Score, generated using deep learning feature representations from PA view radiographs targeting TF joint structure.

Both scores were computed for each index knee using baseline radiographs from the

US and UK cohorts. No model retraining or recalibration was performed; the outputs from the original models were directly used.

To evaluate the predictive utility of combining these two views, the following comparisons were conducted:

GPR Score vs. Combined Score, and DR Score vs. Combined Score: assessing whether the integrated model (GPR + DR) outperforms either score alone, thus supporting the value of multi-compartment analysis.

KL Grade vs. KL Grade + PFOA, and PFOA vs. KL Grade + PFOA: evaluating whether combining conventional clinical indicators improves prediction.

Combined GPR + DR Score vs. KL Grade, and Combined GPR + DR Score vs. KL Grade + PFOA: benchmarking the multi-view radiomics model against expert-derived structural assessments.

All comparisons were performed in the external test set (UK cohorts), using standard metrics of predictive discrimination including AUC, sensitivity, specificity, and calibration plots where applicable.

This analysis was designed to determine whether multi-view radiomics provides incremental prognostic value beyond single-compartment models or clinical grading systems, offering a preliminary but important step toward comprehensive image-based OA risk stratification.

6.2.6 Statistical Analysis

All statistical analyses were conducted using Python (version 3.9), with commonly used packages such as scikit-learn, pandas, numpy, and pROC for performance evaluation and visualization.

The primary evaluation metric was the AUC, used to compare the discriminative performance of different predictive models. AUC values were computed for each of the following configurations in the external test set (UK cohorts):

- GPR Score alone
- DR Score alone
- Combined GPR + DR Score
- KL Grade alone
- PFOA alone
- Combined KL Grade + PFOA
- Combined GPR + DR Score vs. KL Grade
- Combined GPR + DR Score vs. KL Grade + PFOA

DeLong's test was used for pairwise AUC comparisons to determine whether differences between models were statistically significant. Additionally, sensitivity, specificity, and accuracy were calculated at the optimal cutoff point (defined by Youden's index) to aid interpretation.

Missing data in the outcome (vKR) definition were handled according to the validated algorithm previously developed. Only knees with clearly defined outcomes and complete radiographic data (PA and lateral views) were included in the final analysis.

All statistical tests were two-sided, and a p-value of less than 0.05 was considered statistically significant unless otherwise specified.

6.3 Results

6.3.1 Participants Characteristics

After merging the subjects having GPR Score and DR Score, we get 2,659 participants from US cohort and 135 participants from UK cohorts. Baseline characteristics (at the time of imaging), including age, sex, BMI, KL grade, and the frequency of PFOA, and the incidence of 60-month KR, were summarised in Table 6-1.

Table 6-1 Baseline Characteristics of the Studied Cohorts

		US Cohort	**UK Cohorts**	***p*-value**
Participants		2659	135	/
Age*		62 (55–68; 50–79)	33 (24–47; 17–60)	<0.001
Sex	Male	1628 (61.2%)	101 (74.8%)	<0.001
	Female	1031 (38.8%)	34 (25.2%)	
BMI*		29.72 (26.58–33.56; 16.72–71.91)	26.28 (22.96–29.42; 18.94–49.90)	<0.001
KL Grade*	0	1194 (44.9%)	79 (58.5%)	<0.001
	1	454 (17.1%)	16 (11.9%)	
	2	402 (15.2%)	25 (18.5%)	
	3	405 (15.1%)	8 (5.9%)	

	4	204 (7.7%)	0 (0.0%)	
	Missing	0 (0.0%)	7 (5.2%)	
PFOA*	Yes	403 (15.2%)	17 (12.6%)	1.000
	No	2142 (80.6%)	88 (65.2%)	
	Missing	114 (4.2%)	30 (22.2%)	
KR Event in 60m	Yes	312 (11.7%)	15† (11.1%)	0.934
	No	2347 (88.3%)	120 (88.9%)	

Note: Note: Data are median n (IQR; range) or n (%). Percentages may not sum to 100% due to rounding. P-values were compared between this group and the US cohort. They were obtained using the t-test for continuous variables, including age and BMI. KL Grade was analysed using the Mann-Whitney U test. The remaining categorical variables were compared using the Chi-square test.

*** Baseline refers to the time of imaging.**

† KR Event for UK cohorts was assessed by vKR criteria.

Abbreviations:

US = United States; HK = Hong Kong; UK = United Kingdom; BMI = Body Mass Index; KL = Kellgren–Lawrence Grade; PFOA = Patellofemoral Osteoarthritis; KR = Knee Replacement; vKR = Virtual Knee Replacement; IQR = Interquartile Range.

6.3.2 Model Performance in the US Cohort (MOST)

In the MOST cohort, which provided a well-controlled training and validation environment, the combined radiomics model (DR + GPR Score) demonstrated the highest discriminative ability in predicting KR, with a C-index of 0.907 and an average

AUC of 0.930, both serving as the reference standard for comparison (Table 6-2).

When evaluated individually, both the DR Score and GPR Score showed strong performance. The DR Score achieved a C-index of 0.851 and an average AUC of 0.888, while the GPR Score yielded a C-index of 0.879 and AUC of 0.886. However, neither surpassed the combined model; the improvements in both C-index and AUC were not statistically significant when compared to the individual scores ($p > 0.050$), but the trend consistently favoured integration.

Traditional radiographic assessments performed notably worse. The KL grade alone had a C-index of 0.831 and AUC of 0.867, while PFOA status alone produced substantially lower values (C-index: 0.652, AUC: 0.681), both with statistically significant differences compared to the combined model ($p < 0.050$). Adding PFOA to KL grade (C-index: 0.835, AUC: 0.871) offered only marginal improvement and remained inferior to the multi-view radiomics model.

Table 6-2 Comparison between different score and their combination in US cohort

	C-index	**P-value**	**Average AUC**	***p*-value**
DR Score	0.851	**0.011**	0.888	0.099
GPR Score	0.879	0.108	0.886	0.092
DR + GPR Score	**0.907**	***ref***	**0.930**	***ref***
PFOA	0.652	**0.001**	0.681	**0.001**
KL Grade	0.831	**0.001**	0.867	**0.029**
PFOA + KL Grade	0.835	**0.001**	0.871	**0.001**

Abbreviations: C-index = Concordance Index, AUC = Area Under the Curve, PFOA = Patellofemoral Osteoarthritis, KL Grade = Kellgren–Lawrence Grade, DR Score = Deep-Learning-based Radiomics Score, GPR Score = Generalised Patellofemoral Radiomics Score.

These findings suggest that radiomics-derived scores, especially when combined, outperform conventional clinical gradings in prognosticating knee OA progression in a well-structured cohort.

6.3.3 Model Performance in the UK Cohorts (External Testing)

In the UK cohorts, which served as an external test set representing a younger and more heterogeneous population, the combined DR + GPR Score again outperformed all other models, achieving a C-index of 0.779 and an average AUC of 0.847 (Table 6-3).

Among the individual models, the DR Score maintained relatively good performance (C-index: 0.734, AUC: 0.799) and was not significantly inferior to the combined model ($p = 0.331$ for C-index; $p = 0.361$ for AUC). In contrast, the GPR Score alone showed markedly reduced performance in this setting (C-index: 0.594, AUC: 0.654), likely reflecting the higher variability in lateral radiographs or differences in PF joint pathology prevalence in this population.

Traditional clinical assessments again underperformed. The KL grade yielded a C-index of 0.802 and AUC of 0.821, which were numerically close to the radiomics model but not statistically superior ($p = 0.652$ and $p = 0.582$, respectively). PFOA status alone was the weakest predictor (C-index: 0.531, AUC: 0.572), and even when combined with KL grade (C-index: 0.782, AUC: 0.794), the model did not outperform the multi-view radiomics approach. Notably, adding PFOA to KL grade in the UK cohort resulted in a statistically significant AUC difference ($p = 0.027$), though it remained lower than the radiomics model.

Table 6-3 Comparison between different score and their combination in UK cohorts

	C-index	P-value	Average AUC	*p*-value
DR Score	0.734	0.331	0.799	0.361
GPR Score	0.594	**0.043**	0.654	0.089
DR + GPR Score	**0.779**	***ref***	**0.847**	***ref***
PFOA	0.531	**0.015**	0.572	**0.031**
KL Grade	0.802	0.637	0.821	0.398
PFOA + KL Grade	0.782	0.553	0.794	0.366

Abbreviations: C-index = Concordance Index, AUC = Area Under the Curve, PFOA = Patellofemoral Osteoarthritis, KL Grade = Kellgren–Lawrence Grade, DR Score = Deep-Learning-based Radiomics Score, GPR Score = Generalised Patellofemoral Radiomics Score.

Overall, these results demonstrate the robustness and added value of integrating TF and PF radiomic features for predicting structural OA progression across populations with differing clinical characteristics and imaging quality.

6.4 Discussion

This study explored the preliminary integration of radiomics features extracted from both PF and TF compartments using standard clinical knee radiographs. By combining two previously validated scores—the GPR Score and the DR Score—we aimed to investigate whether a multi-compartment, multi-view radiomics approach could improve the prediction of knee OA progression.

Across both the MOST cohort (training/validation) and UK cohorts (external test), the combined model consistently outperformed or matched the performance of single-compartment radiomics scores and traditional clinical grading systems (KL grade and

PFOA). Notably, in the MOST cohort, the combined model achieved the highest C-index and AUC values, suggesting strong discriminative ability in predicting future knee replacement. Even in the UK cohorts, which represented a younger and more heterogeneous population, the combined radiomics model maintained robust performance, demonstrating its generalizability.

Interestingly, while the DR Score showed relatively stable performance across both datasets, the GPR Score's discriminative power declined in the UK cohorts. This discrepancy may reflect differences in PF joint pathology prevalence, imaging protocols, or cohort characteristics, highlighting the potential impact of anatomical region and radiographic view on radiomics model performance. Nevertheless, the additive value of PF features—when combined with TF features—remained evident in both cohorts.

Comparison with conventional metrics such as KL grade and PFOA status reaffirmed the limitations of expert-derived categorical assessments. KL grade alone performed reasonably well but lacked the granularity and prognostic strength observed in radiomics-based scores. Moreover, while adding PFOA to KL grade slightly improved prediction, the combined radiomics model consistently outperformed both, underscoring the benefits of automated, quantitative imaging biomarkers.

These findings support the hypothesis that a multi-compartment radiomics framework captures complementary information from both joint compartments and may provide a more holistic assessment of OA-related structural changes. Although this study represents a preliminary attempt at combining two independent models, it sets the stage for more integrated learning frameworks in the future.

6.5 Chapter summary

In this chapter, we demonstrated that combining patellofemoral and tibiofemoral radiomic features using standard knee radiographs improves the prediction of knee OA progression over single-compartment models and traditional grading systems. The combined GPR + DR Score showed superior or comparable performance across both training and external validation cohorts, confirming the additive prognostic value of a multi-view radiomics strategy.

While this approach was based on independent pre-trained models, the results offer a compelling rationale for developing unified, end-to-end multi-view learning architectures that more effectively integrate anatomical information. Future work should further investigate joint-level interaction modelling, cross-modality integration, and real-world clinical implementation to support precision risk stratification in OA management.

Chapter 7: Discussion, Conclusion, and Future Directions

7.1 Overview and Synthesis of Findings

This thesis explored the development, validation, and integration of radiomics-based approaches for knee OA assessment and prediction, with a particular focus on compartment-specific and multi-view modelling. Through four interrelated studies, the work addressed critical methodological and translational gaps in current OA imaging research, demonstrating how radiomics and deep learning techniques can be strategically applied to improve clinical understanding and risk stratification.

The first study established the feasibility and value of radiomics analysis specifically focused on the PF joint. Using lateral radiographs, a set of handcrafted features was extracted to train a model that predicted the risk of future KR. The study showed that radiomic features derived from the PF compartment contained meaningful prognostic information that extended beyond traditional KL grading. This finding underscored the potential of compartment-specific modelling and highlighted the clinical relevance of PF joint degeneration, which has often been underrepresented in OA imaging research.

Building on this foundation, the second study addressed the generalizability of the PF-specific model by applying it across multiple international cohorts. A domain adaptation framework was employed to mitigate distributional differences between datasets originating from different imaging protocols and populations. The adapted model maintained stable performance across external cohorts, demonstrating that radiomics can be rendered robust and transferable through appropriate methodological strategies. This work addressed a fundamental challenge in radiomics research—the risk of overfitting to local datasets—and illustrated a viable path for broader clinical

deployment of predictive models.

In the third study, the focus shifted to the TF compartment and the use of deep learning methods. A CNN was developed to automatically learn image-based features from the TF joint for the prediction of OA severity and structural progression. Unlike traditional radiomics models, the deep learning framework did not rely on handcrafted feature engineering but instead leveraged data-driven representations, capturing subtle, nonlinear patterns within the joint structure. This approach yielded performance gains over KL grading and highlighted the complementary value of deep learning radiomics in modelling complex structural variation.

The fourth study sought to integrate these two lines of work—compartment-specific modelling and deep learning—into a unified multi-view framework. Radiomic features from both the PF and TF compartments were combined to form a joint representation of the knee, with the goal of enhancing predictive accuracy and capturing inter-compartmental interactions. The results demonstrated that multi-compartment models outperformed single-view approaches, particularly in cases with mild-to-moderate disease severity, where prognostication is most challenging. This integrative strategy reflects the complex anatomical and functional interplay of knee joint compartments and moves toward a more holistic and individualised risk assessment paradigm.

Taken together, the four studies presented in this thesis collectively advance the field of OA imaging by introducing a structured, evidence-based framework for radiomics application. Each study contributes a critical layer to the overarching narrative—from localised modelling and external validation to deep learning enhancement and compartmental integration. Importantly, the findings support the broader vision that radiomics, when applied with methodological rigour and clinical insight, can evolve

into a robust decision-support tool for early diagnosis, progression prediction, and personalised management of knee OA.

7.2 How the Work Addresses the Thesis Objectives

The overarching aim of this thesis was to explore and enhance radiomics-based approaches for the assessment and prediction of knee osteoarthritis (OA), with a focus on compartment-specific modelling, generalizability, and integrative analysis. The four studies presented in this thesis were strategically designed to address specific objectives derived from this aim, and together they form a coherent and progressive body of work that advances the methodological and clinical frontiers of OA imaging.

The first objective was to investigate whether radiomic features extracted from the PF joint could provide additional prognostic value beyond conventional radiographic grading systems. This was addressed in Chapter 3, which demonstrated that PF-specific radiomics features, derived from standard lateral knee radiographs, could effectively predict future knee replacement risk. The study filled a critical gap in OA imaging by highlighting the clinical importance of the PF joint and showing that this underappreciated compartment contains unique structural signals relevant to disease progression.

The second objective involved evaluating the generalizability of radiomics models across diverse populations and imaging protocols. Chapter 4 directly addressed this need through the application of domain adaptation techniques to improve model performance in external cohorts. By testing the PF radiomics model on multiple datasets from different geographical and demographic backgrounds, this study demonstrated that radiomics-based tools can be adapted to operate reliably across real-world clinical

settings, thereby supporting their potential for broader clinical translation.

The third objective was to develop and evaluate a deep learning-based radiomics framework centred on the TF joint. In Chapter 5, a CNN was implemented to extract latent imaging features directly from radiographs, allowing for a data-driven understanding of joint morphology and its relationship to OA severity and progression. This study expanded the methodological scope of the thesis, demonstrating that deep learning could serve as a powerful complement to traditional radiomics approaches, particularly in its ability to model complex anatomical patterns without manual feature engineering.

The fourth and final objective was to explore whether integrating radiomic features from both the PF and TF compartments would yield improved predictions of OA progression. This was accomplished in Chapter 6, where a multi-view radiomics model was proposed and validated. The study showed that combining compartment-specific information captured a more comprehensive representation of joint health and improved the accuracy of progression prediction, particularly in clinically ambiguous cases. This integrative approach reflects a move toward holistic and personalised modelling strategies in OA research.

Collectively, these four studies operationalise the thesis objectives and contribute incrementally toward the development of a robust, generalizable, and clinically meaningful radiomics-based framework for knee OA. Each article not only addresses a specific research question but also builds upon the insights and limitations of the preceding studies, resulting in a coherent research trajectory. Together, they validate the feasibility of using radiomics and deep learning to derive meaningful imaging biomarkers, demonstrate methods to extend these models across domains, and

culminate in an integrative strategy that mirrors the complex, multicompartmental nature of knee OA.

In doing so, the thesis advances the scientific understanding of how quantitative imaging can be used to inform clinical decision-making, stratify patients based on progression risk, and ultimately support more personalised approaches to knee OA management.

7.3 Theoretical and Clinical Connections Among Studies

Although each study in this thesis addressed a distinct research question, they are closely interwoven through shared theoretical foundations and convergent clinical implications. At the theoretical level, all four studies are grounded in the central premise that quantitative image analysis can uncover latent structural information predictive of disease severity and progression in knee OA. Radiomics and deep learning serve as complementary strategies to operationalise this premise—radiomics through engineered feature extraction, and deep learning through automated representation learning. These methods, though distinct in technical approach, are unified in their capacity to transform traditional two-dimensional radiographs into high-dimensional data sources for risk stratification.

The clinical motivation underlying the studies is similarly consistent: to overcome the limitations of conventional radiographic grading systems, such as the KL grading, which are coarse, semi-quantitative, and often fail to detect early or compartment-specific degeneration (10, 11). This shared objective provides a common framework for the four studies, which collectively propose a shift from global, qualitative scoring toward localized, data-driven modelling of OA risk and trajectory.

The compartment-specific focus in Chapters 3 and 5—targeting the PF and TF joints, respectively—highlights the heterogeneity of OA manifestation and the need to model anatomical substructures independently. While the PF and TF compartments differ in their biomechanical roles and disease progression pathways, the studies demonstrate that both regions harbour predictive imaging biomarkers that are undervalued by global assessment tools. The complementary nature of these compartment-focused models forms the conceptual foundation for Chapter 6, which integrates features from both compartments. This multi-view approach aligns with theoretical frameworks in systems biology and personalised medicine that emphasise integrative, multi-dimensional modelling of complex disease processes.

Furthermore, the issue of generalizability explored in Chapter 4 links directly to the broader challenge of translating image-based biomarkers into real-world clinical practice. Domain adaptation techniques applied in this study not only address dataset-specific bias but also reinforce the broader theoretical notion that effective clinical models must be robust across heterogeneous patient populations and imaging protocols. This reinforces the clinical value of the models developed in other chapters and underscores the necessity of external validation in health data science.

From a translational perspective, all four studies converge on the clinical goal of early identification of patients at risk for rapid OA progression or knee replacement surgery. This objective is particularly relevant in health systems with limited orthopaedic resources and long surgical wait times, where accurate triage can improve patient outcomes and resource allocation. The studies propose practical frameworks—whether through radiomics scores, deep learning outputs, or combined risk indices—that can be integrated into clinical decision-support systems.

In summary, the included studies are conceptually linked through a shared commitment to quantitative, individualised OA modelling, and clinically aligned through their focus on improving diagnosis, prognostication, and patient management. Together, they form a cohesive narrative that bridges methodological innovation with real-world applicability, advancing the field toward more precise and equitable OA care.

7.4 Contributions to Knowledge and Clinical Impact

This thesis contributes novel insights to the field of knee OA by introducing and validating radiomics-based approaches that enhance the assessment of disease severity, predict progression, and improve the personalisation of clinical care. Through a structured series of studies, it bridges methodological innovation with practical application, advancing both the scientific understanding and clinical utility of imaging biomarkers in OA.

One of the primary contributions of this work is the demonstration that radiomics analysis, when targeted to specific anatomical compartments of the knee, can yield predictive markers that outperform or complement traditional grading systems such as the KL grade. The introduction of PF-specific radiomics in Chapter 3 fills a notable gap in OA research, where the PF joint has often been overlooked despite its clinical significance, particularly in patients with anterior knee pain or isolated PF degeneration. By showing that PF radiomic features can independently predict the risk of knee replacement, the study offers a refined tool for identifying high-risk patients who may not be flagged by conventional assessments.

The thesis also contributes to the critical discourse on model generalizability in medical imaging. Chapter 4 highlights the limitations of models trained in a single population

and presents domain adaptation as a solution for bridging distributional gaps across diverse datasets. This addresses a persistent challenge in radiomics—the sensitivity of features to imaging settings, scanners, and demographic differences—and provides a pathway for ensuring equitable and reliable application of machine learning models in multi-centre or global settings. Such work contributes to the growing literature on translational machine learning and strengthens the feasibility of deploying OA prediction tools in heterogeneous clinical environments.

A further methodological advance is offered through the incorporation of deep learning, specifically CNNs, to model TF joint structures in Chapter 5. This marks a shift from reliance on predefined features to data-driven feature learning, enabling the capture of complex and non-linear morphological patterns that may elude traditional radiomic pipelines. The study not only expands the modelling toolbox available to OA researchers but also provides evidence that deep learning-based models can offer superior granularity and predictive performance for structural progression.

Lastly, the integration of PF and TF features in Chapter 6 presents a holistic view of the knee joint by leveraging a multi-compartment approach. This integrative model aligns with the clinical reality that OA is a multifaceted disease affecting the joint as a whole, often with compartmental interplay. By demonstrating that multi-view radiomics improves predictive accuracy—particularly in early or ambiguous cases—this work contributes a clinically meaningful strategy for more precise patient stratification and earlier intervention.

The potential impact of these contributions on OA management is multi-fold. First, they support the development of radiograph-based decision-support tools that can assist clinicians in identifying patients at risk of rapid progression, enabling more timely and

tailored therapeutic planning. Second, they pave the way for automated, scalable imaging solutions that reduce subjectivity and improve consistency in OA evaluation. Third, by enhancing prediction and stratification, these models may reduce unnecessary interventions for low-risk individuals while prioritising care for those most likely to benefit—ultimately contributing to more efficient resource use in strained healthcare systems.

In sum, this thesis advances the field of OA research by developing and validating a suite of radiomics-based models that not only improve the accuracy of disease assessment but also move toward clinical translation. These contributions lay the foundation for a new generation of precision tools in musculoskeletal radiology, with the potential to reshape how OA is diagnosed, monitored, and managed in practice.

7.5 Methodological Strengths, Innovations, and Limitations

The four studies presented in this thesis collectively represent a methodologically rigorous and forward-looking approach to the application of radiomics in knee OA. Each study incorporates key innovations that strengthen the overall contribution, while also revealing common methodological challenges that merit careful reflection. This section synthesises the methodological strengths, novel elements, and shared limitations that characterise the body of work.

One of the core methodological strengths lies in the compartment-specific design. By focusing independently on the PF and tibiofemoral TF joints, the studies recognise the anatomical heterogeneity of OA and avoid the oversimplification inherent in global joint scoring systems. This design enables targeted feature extraction and improves the interpretability of findings, allowing for a more nuanced understanding of how disease

evolves within different parts of the joint.

Another strength is the staged research design, which progresses from foundational handcrafted radiomics to deep learning and ultimately to integrative modelling. This stepwise approach provides clarity on the added value of each methodology and avoids overcomplication at the early stages. The use of CNNs in the deep-learning-based radiomics model represents a key methodological innovation, offering automated feature learning from raw image data and mitigating the need for extensive manual feature engineering. This approach enhances scalability and adaptability in future applications.

A further innovation is the incorporation of domain adaptation strategies to improve model generalizability. Most radiomics studies are limited to internal validation and suffer from performance degradation when applied to external cohorts. In contrast, this thesis explicitly addresses domain shift by introducing transfer learning methods that adapt the model to new populations without full retraining. This advancement not only improves the translational potential of the models but also contributes methodologically to the field of generalizable machine learning in medical imaging.

Despite these strengths, several common limitations are acknowledged. First, although the use of radiographs as input data offers broad clinical accessibility, they are inherently two-dimensional and may not capture subtle cartilage or soft-tissue changes visible on MRI. As a result, radiograph-based radiomics models may miss structural nuances relevant to early OA or biomechanical alterations, limiting their sensitivity in certain contexts.

Second, while the models were evaluated on multiple datasets, their performance

remains dependent on image quality, preprocessing choices, and scanner variability. Even with domain adaptation, full harmonisation across sites remains challenging, and residual bias cannot be entirely excluded. Additionally, the interpretability of deep learning models—despite their strong performance—remains a challenge. The features extracted by CNNs are high-dimensional and lack direct anatomical or physiological correspondence, which may hinder clinical trust and regulatory approval.

Another shared limitation relates to the observational design of the included cohorts. Most datasets (MOST, OAI, MenTOR, and KICK) are existing prospective cohorts, whereas the Hong Kong dataset is retrospective. Although progression and surgical outcomes were used as endpoints, causal relationships cannot be inferred, and model predictions may be affected by unmeasured confounding. In addition, some subgroups (e.g., knees with isolated PFOA or early-stage disease) are underrepresented, which may limit the generalizability of the findings to broader clinical populations.

Lastly, while the studies show clear potential for clinical application, none of the models has yet been prospectively validated or integrated into routine clinical workflows. Real-world feasibility, usability by non-technical clinicians, and cost-effectiveness analyses remain as future steps before clinical deployment.

In conclusion, the methodological rigour and innovation of this thesis lie in its thoughtful progression from handcrafted to deep-learning radiomics, its compartment-specific modelling strategy, and its attention to cross-cohort generalizability. These strengths are balanced by limitations related to data scope, imaging modality, interpretability, and validation stage. Recognising these constraints is critical for guiding future improvements and ensuring that radiomics models evolve in a clinically responsible and scientifically robust manner.

7.6 Recommendations for future research

While the present thesis demonstrates the feasibility and clinical value of radiomics-based analysis in knee OA, several avenues remain for future research to extend, refine, and translate these findings into clinical practice. The following recommendations highlight potential directions for methodological enhancement, dataset expansion, and integrative modelling.

First, the current multi-view radiomics model integrates information from the PF and TF compartments using relatively straightforward feature-level concatenation. Future studies may explore more advanced architectures, such as attention-based models or transformer networks, which are capable of capturing inter-compartment relationships more dynamically and adaptively. These methods may be better suited to model the complex spatial and functional interplay between knee compartments. Additionally, radiographs such as the skyline view, which provides an axial perspective of the PF joint, may offer complementary information and should be considered for inclusion in multi-view frameworks.

Second, all predictive models developed in this thesis rely solely on baseline radiographs, representing a single time point in the disease course. However, knee OA is a chronic and progressive condition, and the trajectory of structural change may carry more prognostic value than any single image. Future research should investigate the use of longitudinal imaging data, incorporating temporal features derived from serial radiographs. Changes over time may reflect the influence of lifestyle, comorbidities, or interventions, and could allow for the modelling of progression velocity—potentially improving the accuracy and clinical relevance of predictions.

Third, the domain adaptation strategies explored in this thesis were applied only to the PF-specific radiomics model. While initial results indicate the feasibility of transferring models across diverse imaging sources and populations, further work is needed to test domain adaptation in more complex frameworks, including multi-compartment and deep-learning-based models. Validation across additional datasets from varied clinical environments will be essential to confirm robustness and identify limitations in generalizability.

Fourth, the current radiomics models focus exclusively on features extracted from medical imaging, which reflect morphological and textural properties of joint structures. However, OA is a multifactorial disease influenced by biomechanical, biochemical, and systemic factors. Future models should consider integrating clinical variables (e.g., age, BMI, physical activity), patient-reported outcomes, and molecular biomarkers (e.g., inflammatory cytokines or cartilage degradation products). A multimodal approach could offer a more comprehensive understanding of disease mechanisms and further personalise risk stratification.

Finally, while this thesis demonstrates proof-of-concept for radiomics as a decision-support tool, prospective studies are needed to evaluate its real-world utility. Future research should focus on the development of user-friendly platforms for radiomics-based prediction, conduct clinical trials to assess decision impact, and explore how such tools might be incorporated into personalised treatment pathways. These efforts are aligned with the broader goals of precision medicine, aiming to move beyond "one-size-fits-all" approaches and deliver targeted interventions based on individual risk profiles.

In summary, the future of radiomics in knee OA lies in methodological innovation,

longitudinal and multimodal integration, rigorous external validation, and translational research that bridges the gap between technical development and clinical impact.

7.7 Concluding Remarks

This thesis set out to explore the potential of radiomics and deep learning techniques in advancing the assessment and prediction of knee OA progression, with a focus on compartment-specific modelling and clinical applicability. Through a sequence of four interlinked studies, the work has contributed to a more precise, data-driven understanding of knee joint degeneration using routinely acquired radiographic images.

The findings collectively demonstrate that quantitative imaging features from the patellofemoral and tibiofemoral compartments contain prognostically relevant information that is not captured by traditional grading systems. By applying advanced machine learning models—including handcrafted radiomics, convolutional neural networks, and domain adaptation techniques—this thesis offers novel insights into OA heterogeneity and progression risk. Furthermore, the integration of multi-compartment features has shown that comprehensive joint analysis can enhance predictive accuracy and support individualised patient management.

While limitations remain, the work lays a robust foundation for future developments in OA imaging research. It calls for greater integration of temporal data, incorporation of multimodal inputs, and prospective evaluation in clinical environments. The long-term vision is clear: to transition from descriptive and delayed OA assessment toward predictive and personalised care.

Ultimately, this thesis contributes to a growing movement in musculoskeletal

medicine—one that seeks to transform diagnostic radiology from a qualitative interpretive tool into a quantitative, predictive platform. As radiomics and AI technologies continue to mature, their incorporation into routine OA management has the potential to improve clinical outcomes, reduce healthcare burdens, and guide more timely and targeted interventions for patients worldwide.

Appendix

Virtual Knee Replacement (vKR)

vKR is a standardised, data-driven surrogate for end-stage knee osteoarthritis (OA), designed to reduce the influence of non-OA factors on knee replacement (KR) decisions. It has demonstrated strong predictive ability for KR in the Osteoarthritis Initiative (OAI), a large observational OA cohort study (82).

A vKR event is recorded when:

$$KOOS\ Knee\ Pain + 0.5 \times KOOS\ Quality\ of\ Life - (KOOS\ Knee\ Pain\ change\ in\ last\ year) < 88$$

where the KOOS Knee Pain change term is included only if pain has worsened over the past year.

In practice, KOOS scores often fluctuate, so we refined the criteria to improve reliability:

- Sustained decline: vKR Score < 88 at three separate follow-ups, with the third occurrence recorded as the vKR time.

- Rapid deterioration: vKR Score < 88 in two consecutive follow-ups, with the lowest score assigned as the vKR time.

References

1. Wong AY, Samartzis D, Maher C. The global burden of osteoarthritis: past and future perspectives. The Lancet Rheumatology. 2023;5(9):e496-e7.

2. Prieto-Alhambra D, Judge A, Javaid MK, Cooper C, Diez-Perez A, Arden NK. Incidence and risk factors for clinically diagnosed knee, hip and hand osteoarthritis: influences of age, gender and osteoarthritis affecting other joints. Annals of the Rheumatic Diseases. 2014;73(9):1659.

3. James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. 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 2017. The Lancet. 2018;392(10159):1789-858.

4. Cui A, Li H, Wang D, Zhong J, Chen Y, Lu H. Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies. EClinicalMedicine. 2020;29-30:100587.

5. Wallace IJ, Worthington S, Felson DT, Jurmain RD, Wren KT, Maijanen H, et al. Knee osteoarthritis has doubled in prevalence since the mid-20th century. Proceedings of the National Academy of Sciences. 2017;114(35):9332-6.

6. Price AJ, Alvand A, Troelsen A, Katz JN, Hooper G, Gray A, et al. Knee replacement. The Lancet. 2018;392(10158):1672-82.

7. Elective Total Joint Replacement Surgery [March 15, 2025]. Available from: https://www.ha.org.hk/visitor/ha_visitor_index.asp?Content_ID=221223&Lang=ENG&Dimension=1.

8. Tang Sa, Zhang C, Oo WM, Fu K, Risberg MA, Bierma-Zeinstra SM, et al. Osteoarthritis. Nature Reviews Disease Primers. 2025;11(1):10.

9. Dieppe P, Basler H, Chard J, Croft P, Dixon J, Hurley M, et al. Knee replacement surgery for osteoarthritis: effectiveness, practice variations, indications and possible determinants of utilization. Rheumatology (Oxford, England). 1999;38(1):73-83.

10. Mahmoudian A, Lohmander LS, Mobasheri A, Englund M, Luyten FP. Early-stage symptomatic osteoarthritis of the knee — time for action. Nature Reviews Rheumatology. 2021;17(10):621-32.

11. Jiang T, Lau S-H, Zhang J, Chan L-C, Wang W, Chan P-K, et al. Radiomics signature of osteoarthritis: Current status and perspective. Journal of Orthopaedic Translation. 2024;45:100-6.

12. Bowes MA, Kacena K, Alabas OA, Brett AD, Dube B, Bodick N, et al. Machine-learning, MRI bone shape and important clinical outcomes in osteoarthritis: data from the Osteoarthritis Initiative. Annals of the rheumatic diseases. 2021;80(4):502-8.

13. Emery CA, Whittaker JL, Mahmoudian A, Lohmander LS, Roos EM, Bennell KL, et al. Establishing outcome measures in early knee osteoarthritis. Nature Reviews Rheumatology. 2019;15(7):438-48.

14. Roemer FW, Demehri S, Omoumi P, Link TM, Kijowski R, Saarakkala S, et al. State of the Art: Imaging of Osteoarthritis—Revisited 2020. Radiology. 2020;296(1):5-21.

15. Kohn MD, Sassoon AA, Fernando ND. Classifications in Brief: Kellgren-Lawrence Classification of Osteoarthritis. Clinical Orthopaedics and Related Research®. 2016;474(8):1886-93.

16. Antony J, McGuinness K, Moran K, O'Connor NE, editors. Automatic Detection of Knee Joints and Quantification of Knee Osteoarthritis Severity Using Convolutional Neural Networks2017; Cham: Springer International Publishing.

17. Hannan MT, Felson DT, Pincus T. Analysis of the discordance between radiographic changes and knee pain in osteoarthritis of the knee. J Rheumatol. 2000;27(6):1513-7.

18. Dieppe PA, Lohmander LS. Pathogenesis and management of pain in osteoarthritis. Lancet. 2005;365(9463):965-73.

19. Hunter DJ, McDougall JJ, Keefe FJ. The Symptoms of Osteoarthritis and the Genesis of Pain. Rheumatic Disease Clinics of North America. 2008;34(3):623-43.

20. Peterfy CG, Guermazi A, Zaim S, Tirman PFJ, Miaux Y, White D, et al. Whole-Organ Magnetic Resonance Imaging Score (WORMS) of the knee in osteoarthritis. Osteoarthritis and Cartilage. 2004;12(3):177-90.

21. Hunter DJ, Guermazi A, Lo GH, Grainger AJ, Conaghan PG, Boudreau RM, et al. Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis Knee Score). Osteoarthritis and cartilage. 2011;19(8):990-1002.

22. Frye BM, Najim AA, Adams JB, Berend KR, Lombardi Jr AV. MRI is more accurate than CT for patient-specific total knee arthroplasty. The knee. 2015;22(6):609-12.

23. Shan J, Alam SK, Garra B, Zhang Y, Ahmed T. Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods. Ultrasound in medicine & biology. 2016;42(4):980-8.

24. Tiulpin A, Saarakkala S, Mathiessen A, Hammer H, Furnes O, Fenstad A, et al. Predicting total knee replacement from ultrasound using machine learning. Osteoarthritis and Cartilage. 2019;27:S360-S1.

25. Oishi Y, Ishige Y, Takemura H, Kurokawa H, Tanaka Y, Kosugi S, et al., editors. Three-Dimensional Shape Statistical Analysis of Tibial Plafond Deformed by Ankle Osteoarthritis. 2021 6th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS); 2021 25-27 Nov. 2021.

26. Minciullo L, Bromiley PA, Felson DT, Cootes TF. Indecisive Trees for Classification and Prediction of Knee Osteoarthritis. Springer, Cham. 2017.

27. Bayramoglu N, Nieminen MT, Saarakkala S. Machine learning based texture analysis of patella from X-rays for detecting patellofemoral osteoarthritis. International journal of medical informatics. 2022;157:104627.

28. Navale DI, Hegadi RS, Mendgudli N, editors. Block based texture analysis approach for knee osteoarthritis identification using SVM. IEEE International Wie Conference on Electrical & Computer Engineering; 2015.

29. Zhao J, Jiang T, Lin Y, Chan LC, Chan PK, Wen C, et al. Adaptive Fusion of Deep Learning with Statistical Anatomical Knowledge for Robust Patella Segmentation

from CT Images. IEEE Journal of Biomedical and Health Informatics. 2024:1-12.

30. Bayramoglu N, Tiulpin A, Hirvasniemi J, Nieminen MT, Saarakkala S. Adaptive segmentation of knee radiographs for selecting the optimal ROI in texture analysis. Osteoarthritis and cartilage. 2020;28(7):941-52.

31. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48(4):441-6.

32. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. Insights into Imaging. 2020;11(1):91.

33. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2016;278(2):563-77.

34. Bianchi J, de Oliveira Ruellas AC, Gonçalves JR, Paniagua B, Prieto JC, Styner M, et al. Osteoarthritis of the Temporomandibular Joint can be diagnosed earlier using biomarkers and machine learning. Scientific Reports. 2020;10(1):8012.

35. Tenório APM, Faleiros MC, Junior JRF, Dalto VF, Assad RL, Louzada-Junior P, et al. A study of MRI-based radiomics biomarkers for sacroiliitis and spondyloarthritis. International Journal of Computer Assisted Radiology and Surgery. 2020;15(10):1737-48.

36. Xie Y, Dan Y, Tao H, Wang C, Zhang C, Wang Y, et al. Radiomics Feature Analysis of Cartilage and Subchondral Bone in Differentiating Knees Predisposed to Posttraumatic Osteoarthritis after Anterior Cruciate Ligament Reconstruction from Healthy Knees. BioMed Research International. 2021;2021:4351499.

37. Hirvasniemi J, Klein S, Bierma-Zeinstra S, Vernooij MW, Schiphof D, Oei EHG. A machine learning approach to distinguish between knees without and with osteoarthritis using MRI-based radiomic features from tibial bone. European Radiology. 2021;31(11):8513-21.

38. Lin T, Fu S, Zeng D, Lu S, Zhou M, Li J, et al. Predicting response to vitamin D treatment on osteoarthritis-A radiomics nomogram study based on magnetic

resonance imaging. Osteoarthritis and Cartilage. 2021;29:S347-S8.

39. Zhang J, Wang JZ, Yuan Z, Sobel ES, Jiang H. Computer-aided classification of optical images for diagnosis of osteoarthritis in the finger joints. Journal of X-ray science and technology. 2011;19(4):531-44.

40. Stachowiak GW, Wolski M, Woloszynski T, Podsiadlo P. Detection and prediction of osteoarthritis in knee and hand joints based on the X-ray image analysis. Biosurface & Biotribology. 2016;2(4):162-72.

41. Paniagua B, Ruellas A, Benavides E, Marron S, Wolford L, Cevidanes L. Validation of CBCT for the computation of textural biomarkers: SPIE; 2015.

42. Bianchi J, Gonçalves JR, de Oliveira Ruellas AC, Ashman LM, Vimort JB, Yatabe M, et al. Quantitative bone imaging biomarkers to diagnose temporomandibular joint osteoarthritis. International Journal of Oral and Maxillofacial Surgery. 2021;50(2):227-35.

43. Ribera NT, de Dumast P, Yatabe M, Ruellas A, Ioshida M, Paniagua B, et al. Shape variation analyzer: a classifier for temporomandibular joint damaged by osteoarthritis: SPIE; 2019.

44. Halilaj E, Moore DC, Laidlaw DH, Got CJ, Weiss A, Ladd AL, et al. The morphology of the thumb carpometacarpal joint does not differ between men and women, but changes with aging and early osteoarthritis. Journal of Biomechanics. 2014;47(11):2709-14.

45. Tycowicz CV, editor Towards Shape-Based Knee Osteoarthritis Classification Using Graph Convolutional Networks. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI); 2020.

46. Kubkaddi S, Ravikumar KM. Early detection of Knee Osteoarthritis using SVM Classifier. 2017.

47. Kumar VA, Jayanthy AK, editors. Classification of MRI images in 2D coronal view and measurement of articular cartilage thickness for early detection of knee osteoarthritis. 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT); 2016.

48. Peuna A, Thevenot J, Saarakkala S, Nieminen MT, Lammentausta E. Machine learning classification on texture analyzed T2 maps of osteoarthritic cartilage: oulu knee osteoarthritis study. Osteoarthritis and Cartilage. 2021;29(6):859-69.

49. Yu HJ, Horiuch S, Luk A, Rudd A, Ton J, Kuoy E, et al. Texture features from T2 mapping of talar dome cartilage in normal volunteers and dancers. Osteoarthritis and Cartilage. 2018;26:S72-S3.

50. Thomson J, O'Neill T, Felson D, Cootes T, editors. Automated shape and texture analysis for detection of osteoarthritis from radiographs of the knee. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2015: Springer.

51. Wolski, M., Podsiadlo, P., Stachowiak, G., et al. Trabecular bone texture detected by plain radiography is associated with MRI-defined osteophytes in finger joints of women without radiographic osteoarthritis. Osteoarthritis and cartilage. 2018;26(7):924-8.

52. Nelson AE, Golightly YM, Lateef S, Renner JB, Jordan JM, Aspden RM, et al. Cross-sectional associations between variations in ankle shape by statistical shape modeling, injury history, and race: the Johnston County Osteoarthritis Project. Journal of foot and ankle research. 2017;10(1):1-7.

53. Kvarda P, Heisler L, Krhenbühl N, Steiner CS, Hintermann B. 3D Assessment in Posttraumatic Ankle Osteoarthritis. Foot & Ankle International. 2020;42(199):107110072096131.

54. Marques J, Genant HK, Lillholm M, Dam EB. Diagnosis of osteoarthritis and prognosis of tibial cartilage loss by quantification of tibia trabecular bone from MRI. Magnetic Resonance in Medicine. 2013;70(2):568-75.

55. Almhdie-Imjabbar A, Nguyen K-L, Toumi H, Jennane R, Lespessailles E. Prediction of knee osteoarthritis progression using radiological descriptors obtained from bone texture analysis and Siamese neural networks: data from OAI and MOST cohorts. Arthritis Research & Therapy. 2022;24(1):66.

56. Gielis W, Weinans H, Welsing PM, van Spil W, Agricola R, Cootes T, et al. An automated workflow based on hip shape improves personalized risk prediction for hip

osteoarthritis in the CHECK study. Osteoarthritis and cartilage. 2020;28(1):62-70.

57. Hirvasniemi J, Gielis WP, Arbabi S, Agricola R, Willem EVS, Arbabi V, et al. Bone Texture Analysis for Prediction of Incident Radio-graphic Hip Osteoarthritis Using Machine Learning: Data from the Cohort Hip and Cohort Knee (CHECK) study. Osteoarthritis & Cartilage. 2019.

58. Neogi T, Bowes MA, Niu J, Souza K, Fe Lson DT. Magnetic Resonance Imaging-Based Three-Dimensional Bone Shape of the Knee Predicts Onset of Knee Osteoarthritis: Data From the Osteoarthritis Initiative. Arthritis & Rheumatology. 2013;65(8).

59. Li J, Fu S, Gong Z, Zhu Z, Zeng D, Cao P, et al. MRI-based Texture Analysis of Infrapatellar Fat Pad to Predict Knee Osteoarthritis Incidence. Radiology. 2022;0(0):212009.

60. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology. 2020;295(2):328-38.

61. Segal NAMD, Nevitt MCP, Gross KDMD, Hietpas JMSW, Glass NAMA, Lewis CEMD, et al. The Multicenter Osteoarthritis Study: Opportunities for Rehabilitation Research. PM & R. 2013;5(8):647-54.

62. Eckstein F, Wirth W, Nevitt MC. Recent advances in osteoarthritis imaging—the Osteoarthritis Initiative. Nature Reviews Rheumatology. 2012;8(10):622-30.

63. Duncan RC, Hay EM, Saklatvala J, Croft PR. Prevalence of radiographic osteoarthritis—it all depends on your point of view. Rheumatology (Oxford, England). 2006;45(6):757-60.

64. Hart HF, Stefanik JJ, Wyndow N, Machotka Z, Crossley KM. The prevalence of radiographic and MRI-defined patellofemoral osteoarthritis and structural pathology: a systematic review and meta-analysis. British journal of sports medicine. 2017;51(16):1195-208.

65. de Lange-Brokaar BJE, Bijsterbosch J, Kornaat PR, Yusuf E, Ioan-Facsinay A,

Zuurmond AM, et al. Radiographic progression of knee osteoarthritis is associated with MRI abnormalities in both the patellofemoral and tibiofemoral joint. Osteoarthritis and cartilage. 2016;24(3):473-9.

66. Macri EM, van Middelkoop M, Damen J, Bos PK, Bierma-Zeinstra SMA. Higher risk of knee arthroplasty during ten-year follow-up if baseline radiographic osteoarthritis involves the patellofemoral joint: a CHECK Cohort Study. BMC musculoskeletal disorders. 2022;23(1):1-600.

67. Lowekamp BC, Chen DT, Ibáñez L, Blezek D. The design of simpleITK. Frontiers in neuroinformatics. 2013;7:45-.

68. Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer research (Chicago, Ill). 2017;77(21):e104-e7.

69. Ding C, Peng H. MINIMUM REDUNDANCY FEATURE SELECTION FROM MICROARRAY GENE EXPRESSION DATA. Journal of bioinformatics and computational biology. 2005;3(2):185-205.

70. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of machine learning research. 2017;18:1-5.

71. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods. 2020;17(3):261-72.

72. Davidson-Pilon C. lifelines: survival analysis in Python. Journal of open source software. 2019;4(40):1317.

73. Bayramoglu N, Nieminen MT, Saarakkala S. Machine learning based texture analysis of patella from X-rays for detecting patellofemoral osteoarthritis. International journal of medical informatics (Shannon, Ireland). 2022;157:104627-.

74. Kalichman L, Zhang Y, Niu J, Goggins J, Gale D, Felson DT, et al. The association between patellar alignment and patellofemoral joint osteoarthritis features—an MRI study. Rheumatology (Oxford, England). 2007;46(8):1303-8.

75. Eijkenboom JJ, Waarsing JH, Lankhorst NE, Bierma-Zeinstra SM, van Middelkoop M. Statistical shape modelling of the patella: Patients with patellofemoral pain and patellofemoral osteoarthritis share similar aberrant shape aspects compared to healthy controls. Osteoarthritis and cartilage. 2016;24:S243-S4.

76. Liao TC, Jergas H, Tibrewala R, Bahroos E, Link TM, Majumdar S, et al. Longitudinal analysis of the contribution of 3D patella and trochlear bone shape on patellofemoral joint osteoarthritic features. Journal of orthopaedic research. 2021;39(3):506-15.

77. Dai Y, Yin H, Xu C, Zhang H, Guo A, Diao N. Association of patellofemoral morphology and alignment with the radiographic severity of patellofemoral osteoarthritis. Journal of orthopaedic surgery and research. 2021;16(1):1-548.

78. Gudas R, Šiupšinskas L, Gudaitė A, Vansevičius V, Stankevičius E, Smailys A, et al. The patello-femoral joint degeneration and the shape of the patella in the population needing an arthroscopic procedure. Medicina (Kaunas, Lithuania). 2018;54(2):21.

79. Rajamohan HR, Wang T, Leung K, Chang G, Cho K, Kijowski R, et al. Prediction of total knee replacement using deep learning analysis of knee MRI. Scientific reports. 2023;13(1):6922-.

80. Duncan R, Peat G, Thomas E, Hay EM, Croft P. Incidence, progression and sequence of development of radiographic knee osteoarthritis in a symptomatic population. Annals of the rheumatic diseases. 2011;70(11):1944-8.

81. Culvenor AG, Lai CCH, Gabbe BJ, Makdissi M, Collins NJ, Vicenzino B, et al. Patellofemoral osteoarthritis is prevalent and associated with worse symptoms and function after hamstring tendon autograft ACL reconstruction. British journal of sports medicine. 2014;48(6):435-9.

82. Kwoh CK, Boudreau RM, Eckstein F, Roemer F, Hannon MJ, Guermazi A, et al. A virtual knee replacement (vKR) multi-component endpoint for knee osteoarthritis based on patient-reported PROs: Data from the Osteoarthritis Initiative. Osteoarthritis and Cartilage.

83. Zhang J, Jiang T, Chan L-C, Lau S-H, Wang W, Teng X, et al. Radiomics

analysis of patellofemoral joint improves knee replacement risk prediction: Data from the Multicenter Osteoarthritis Study (MOST). Osteoarthritis and Cartilage Open. 2024:100448.

84. Rubin DB. The bayesian bootstrap. The annals of statistics. 1981:130-4.

85. Valsamis EM, Jensen ML, Coward G, Sayers A, Pinedo-Villanueva R, Rasmussen JV, et al. Risk of serious adverse events after primary shoulder replacement: development and external validation of a prediction model using linked national data from England and Denmark. The Lancet Rheumatology. 2024;6(9):e607-e14.

86. McCabe PG, Lisboa P, Baltzopoulos B, Olier I. Externally validated models for first diagnosis and risk of progression of knee osteoarthritis. PLOS ONE. 2022;17(7):e0270652.

87. Ukachukwu V, Duncan R, Belcher J, Marshall M, Stefanik J, Crossley K, et al. Clinical Significance of Medial Versus Lateral Compartment Patellofemoral Osteoarthritis: Cross-Sectional Analyses in an Adult Population With Knee Pain. Arthritis Care & Research. 2017;69(7):943-51.

88. Stefanik JJ, Gross KD, Guermazi A, Felson DT, Roemer FW, Zhang Y, et al. The relation of MRI-detected structural damage in the medial and lateral patellofemoral joint to knee pain: the Multicenter and Framingham Osteoarthritis Studies. Osteoarthritis and Cartilage. 2015;23(4):565-70.

89. Kellgren JH, Lawrence JS. Radiological Assessment of Osteo-Arthrosis. Annals of the Rheumatic Diseases. 1957;16(4):494-502.

90. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence. 2016;39(6):1137-49.

91. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage. 2006;31(3):1116-28.

92. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern

recognition; 2016.

93. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

94. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv preprint arXiv:171010903. 2017.

95. Wang X, Girshick R, Gupta A, He K, editors. Non-local neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2018.

96. Shao Z, Bian H, Chen Y, Wang Y, Zhang J, Ji X. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in Neural Information Processing Systems. 2021;34.

97. Rymarczyk D, Borowa A, Tabor J, Zielinski B, editors. Kernel self-attention for weakly-supervised image classification using deep multiple instance learning. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2021.

98. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC medical research methodology. 2018;18:1-12.

99. Tiulpin A, Klein S, Bierma-Zeinstra SMA, Thevenot J, Rahtu E, Meurs Jv, et al. Multimodal Machine Learning-based Knee Osteoarthritis Progression Prediction from Plain Radiographs and Clinical Data. Scientific Reports. 2019;9(1):20038.

100. Leung K, Zhang B, Tan J, Shen Y, Geras KJ, Babb JS, et al. Prediction of Total Knee Replacement and Diagnosis of Osteoarthritis by Using Deep Learning on Knee Radiographs: Data from the Osteoarthritis Initiative. Radiology. 2020;296(3):584-93.

101. Wei J, Gross D, Lane NE, Lu N, Wang M, Zeng C, et al. Risk factor heterogeneity for medial and lateral compartment knee osteoarthritis: analysis of two prospective cohorts. Osteoarthritis and Cartilage. 2019;27(4):603-10.

102. Davidson EJ, Figgie C, Nguyen J, Pedoia V, Majumdar S, Potter HG, et al. Chondral Injury Associated With ACL Injury: Assessing Progressive Chondral

Degeneration With Morphologic and Quantitative MRI Techniques. Sports Health. 2024;16(5):722-34.

103. Hunter DJ, Lohmander LS, Makovey J, Tamez-Peña J, Totterman S, Schreyer E, et al. The effect of anterior cruciate ligament injury on bone curvature: exploratory analysis in the KANON trial. Osteoarthritis and Cartilage. 2014;22(7):959-68.

104. Namiri NK, Flament I, Astuto B, Shah R, Tibrewala R, Caliva F, et al. Deep Learning for Hierarchical Severity Staging of Anterior Cruciate Ligament Injuries from MRI. Radiology: Artificial Intelligence. 2020;2(4):e190207.