Title
Characterizing Phenotypes of Musculoskeletal Degeneration Using Medical Imaging and Deep Learning

Author
Iriondo, Claudia

Publication Date
2021

Peer reviewed|Thesis/dissertation

Characterizing Phenotypes of Musculoskeletal Degeneration Using Medical Imaging and Deep Learning

by

Claudia Iriondo

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
AND
UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:

Sharmila Majumdar — Sharmila Majumdar

Chair

Valentina Pedoia — Valentina Pedoia

Grace O'Connell — Grace O'Connell

_____

_____

Committee Members

Acknowledgements

I cannot begin to express how grateful I am to all the people who supported me during my PhD. First and foremost, I am incredibly thankful for the mentorship from my advisor, Sharmila Majumdar. From the moment I stepped into your office as a first-year student, I could tell you would be a fantastic mentor: you valued my input, nurtured my scientific curiosity, and kept me grounded and productive. Over the years, you encouraged me to participate in numerous conferences, grants, and collaborations. Through you, I've gained invaluable exposure to the field of musculoskeletal imaging and awareness of the clinical impact of musculoskeletal diseases, which kept me motivated and focused on my day-to-day research. I deeply appreciate your scientific guidance and thank you for providing all the opportunities and resources I needed to succeed.

I am also incredibly thankful for Valentina Pedoia's mentorship which was instrumental to my development as a scientist. You actively worked with me to shape 'out there' ideas into concrete research questions, experiments, and timelines. Even with a busy schedule, your office door was always open for a problem solving or brainstorming session, which, without exception, turned out to be fruitful and energizing. Thank you for establishing a collaborative research culture in our group. Sharmila and Valentina: I admire your deep technical expertise, tenacity, and boldness. Your passion for research is infectious and I could not have done this without the both of you.

My qualifying exam committee and dissertation committee members were crucial to the success of my projects. Thank you Peder Larson for your insightful feedback on early drafts of my qualifying exam proposal and helping me understand the technical

details underlying our quantitative pulse sequences. Richard Souza, thank you for patiently stress testing several of the assumptions underlying my research proposals and for your steadfast attention to detail in all our MQIR meetings. Helen Kim, thank you for introducing me to the world of causal DAGs in clinical research and encouraging me to outline every step of the data generating process as I formed my proposal. A big thank you to Grace O'Connell for contributing your tissue biomechanics and quantitative imaging expertise, as well as valuable career advice.

I have learned a tremendous amount from my talented colleagues in the Radiology and Biomedical Imaging Department at UCSF. Emma Bahroos thank you for patiently mentoring me in the practical aspects of MR and PET image acquisition as we optimized the imaging protocol for our study, and for helping me navigate the approvals to get new projects off the ground. You've been a constant positive influence throughout my entire PhD and for that I'm grateful. From the clinical side, thank you Thomas Link, Bill Dillon, Matthew Bucknor, Vinil Shah, Jason Talbott, and Cynthia Chin for providing clinical input to make sure our proposed research was impactful and our results were medically sound.

This experience wouldn't have been the same without my lab-mates, old and new: Jasmine Rossi-Devries, Hatef Mehrabian, Aldric Chau, Berk Norman, Michael Girard, Rutwik Shah, Victor Cheng, Radhika Tibrewala, Bruno Astuto, Sarah Mehany, Misung Han, Felix Liu, Alaleh Razmjoo, Alejandro Morales-Martinez, Eugene Ozinsky, Gaurav Inamdar, Jenny Lee, Kenneth Gao, Aniket Tolpadi, Io Flament, Francesco Caliva, Upasana Bharadwaj, Alex Beltran, Pablo Damasceno, and Madeline Hess from the Pedoia/Majumdar group. Alyssa Bird, Carla Kinnunen, and Koren Roach from the Souza group and Andrew Leynes from

the Larson group. I cannot express how much you all have individually contributed to my technical and personal growth, thank you for your constructive feedback, words of encouragement, and lighthearted moments. I'll miss our scientific chats over coffee, sporadic naps during lab meeting, and the birthday celebrations that never quite took off.

To my colleagues in the UC Berkeley/UCSF Bioengineering Graduate group, thank you for pushing me to become a better scientist and give back to the scientific community. I'm constantly humbled and amazed by all your scientific and personal accomplishments, and I cannot wait to see what your future holds. A special thank you to my graduate advisor, Tamara Alliston who was a great listener during our meetings and always had actionable advice.

Finally, I am grateful to my close friends and family for keeping me motivated and sane. Thank you to my parents Alexander Iriondo and Pilar Aznar for your unconditional support, patience, and words of encouragement. Thank you to my little sister Inés: despite being the smallest member of the family, you've always been my biggest cheerleader. Os quiero.

I was blessed to be surrounded by amazing scientists and human beings throughout my PhD. Thank you all for making this possible.

Claudia Iriondo Aznar

Contributions

I would like to thank the coauthors who contributed to the work presented in this dissertation. Specifically, Felix Liu, Francesco Caliva, Sarthak Kumar, Valentina Pedoia, and Sharmila Majumdar for a manuscript titled “Towards Understanding Mechanistic Subgroups of Osteoarthritis: 8 Year Cartilage Thickness Trajectory Analysis” which was published in the Journal of Orthopaedic Research. Alaleh Razmjoo, Francesco Caliva, Jinhee Lee, Valentina Pedoia, and Sharmila Majumdar for their contributions to ISMRM abstracts titled “To-the-point: deep learning on dense point clouds for improved feature extraction” and “Learned knee cartilage and meniscus shape features are associated with osteoarthritis incidence”. Valentina Pedoia and Sharmila Majumdar for a manuscript titled “Lumbar intervertebral disc characterization through quantitative MRI analysis: An automatic voxel-based relaxometry approach” which appeared in Magnetic Resonance in Medicine. A big thank you to Sarah Mehany, Rutwik Shah, Upasana Bharadwaj, Emma Bahroos, Cynthia Chin, Mohammad Diab, Valentina Pedoia, and Sharmila Majumdar for a manuscript titled “Institution-wise shape analysis of spinal curvature and global alignment parameters” submitted to the Journal of Orthopaedic Research. Finally, thank you to my collaborators on projects that did not make it into this dissertation: Michael Girard for his contribution to the hip shape modeling project, Emma Bahroos and Vinil Shah for their contributions to the low back pain PET/MR study, and Victor Cheng for his work on adversarial image augmentation.

Characterizing Phenotypes of Musculoskeletal Degeneration Using Medical Imaging and Deep Learning

Claudia Iriondo

Abstract

Musculoskeletal disease is the leading cause of disability worldwide, with the 2019 Global Burden of Disease study reporting global disease prevalence of approximately 1.714 billion[1]. X-ray and magnetic resonance imaging (MRI) are routinely used for clinical diagnosis and monitoring of musculoskeletal disease, however, due to an increasing volume of acquired images and limited time, image assessments are mainly qualitative. This thesis aims to elevate the role of imaging in the assessment of musculoskeletal disease by developing fully automatic image analysis tools to improve image analysis sensitivity, speed, and/or precision. We target the two conditions with the highest prevalence and healthcare expenditure in the United States: knee osteoarthritis (OA) and back pain. We use deep learning to develop fully automatic tools for image analysis and demonstrate their utility in the assessment and analysis of research and clinical datasets. I will be presenting four main projects:

(1) A deep learning segmentation method for quantitative analysis of knee cartilage from structural MRI to conduct longitudinal analysis on cartilage thickness over 8 years

(2) A point cloud algorithm for feature learning from structural and compositional knee MRI to assess the importance of shape and composition features in predicting OA onset

(3) A registration pipeline for voxel-based analysis of MR imaging of the lumbar spine to examine local associations between $T_{1\rho}$, T_2, and patient reported outcomes

(4) A curve extraction algorithm for analysis of global spine shape from x-ray imaging to build a shape model that examines 3D spine shape variations in the UCSF patient population

Table of Contents

List of Figures

List of Tables

1 Overview

Chapters 2 and 3 serve as a general introduction to the anatomical systems and tissues examined in the remainder of the thesis, as well as the imaging methods used to study them. In Chapter 4, we introduce representation learning and discuss challenges in the application of machine learning for medical imaging. Lastly, we walk through the development of a software tool for training deep learning segmentation algorithms. Chapters 5-8 are presented as self-contained, experimental studies including study specific background, methods, results, discussion, and future directions.

2 Structure-function of musculoskeletal tissue in health and disease

Articular cartilage, mensici, and intervertebral discs are hydrated, load-bearing tissues integral to the stability and normal function of the musculoskeletal system. The beginning of the chapter will discuss the structure and function of these tissues in health, comparing their biochemical composition and biomechanical properties. Then, the pathophysiology of common knee and spine diseases will be presented, examining how structure is changed and function is impaired by physical or chemical disruptors. This will serve as a general introduction for Chapter 3 which focuses on imaging methods for disease diagnosis and monitoring.

2.1 Articular cartilage

Healthy articular cartilage has high compressive and shear strength which provides smooth articulation, and in the knee, transfers load between the surfaces of the femur, tibia, and patella bones. The mechanical properties of cartilage arise from compositional and structural gradients within the tissue[2; 3]. Cartilage has a layered structure and can be divided into zones: superficial, middle, deep, and calcified. The superficial zone is a thin layer of tightly packed type II collagen fibrils[4; 5]running parallel to the cartilage surface providing tensile strength and resistance to shear forces[6; 7]. This layer is in direct contact with the synovial fluid and functions as barrier,

isolating cartilage from the synovial immune system[8]. It is also the most hydrated layer containing 75-80% water[5]. The middle zone is a thicker layer with randomly oriented collagen fibrils and a higher density of proteoglycans (primarily aggrecan) which together provide high hydrostatic pressure to resist compressive loads[4; 9]. The deep zone is approximately 35% water and has the highest proteoglycan density and hydrostatic pressure[10]. Its collagen fibrils are larger in diameter and aligned perpendicular to the cartilage surface[4; 11]. A partially calcified region of the deep zone provides a physical barrier to prevent angiogenesis from the subchondral bone[12; 13]. The tidemark line is a thin, acellular region that separates the deep zone from the calcified cartilage zone. Large collagen fibrils in the calcified cartilage zone anchor into the underlying bone matrix, acting as an attachment between articular cartilage and subchondral bone[14]. The calcified zone is partially mineralized, has low proteoglycan density, and is semi-permeable to small solutes[15]. The highly organized collagen structure in cartilage can affect MR imaging through the 'magic angle effect' which will be described in Chapter 3. In healthy cartilage, the calcified zone is vascularized, but the remainder of the articular cartilage is avascular and chondrocytes rely on diffusion from the synovial fluid for nutrients[13; 14].

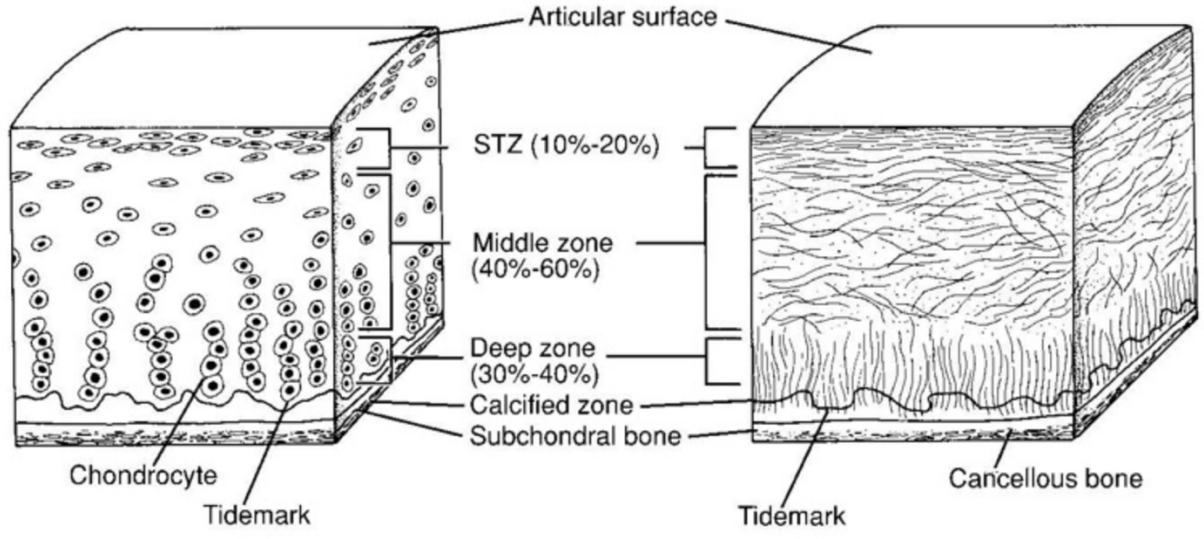


Figure 2.1 Schematic of layered structure in healthy articular cartilage. Left: Cell density, distribution, and shape. Right: Collagen fibril organization. Reproduced with permission from Wolters Kluwer Health, Inc: Buckwalter et al. 1994[16].

2.2 Meniscus

There are two menisci in the knee: the lateral “O-shaped” meniscus and medial “C-shaped” meniscus. Menisci are fibrocartilage structures that provide shock absorption, lubrication, and stability to the knee joint[2; 17]. During loading, the meniscus transfers axial compressive loads into circumferential “hoop” stresses and experiences local compressive, shear, and tensile forces[2; 17]. In a similar way to articular cartilage, the meniscus is divided into distinct compositional and structural zones, each contributing to its load-bearing properties. In the outermost surface layer of the meniscus, a thin network of randomly oriented collagen type I fibrils are bridged by elastin fibers, providing structural support and a barrier from the synovial fluid. Right beneath, the lamellar layer is composed of collagen fiber bundles with a combination of radial and parallel alignments[17; 18]. In the meniscal core, or the deep layer, thick collagen type I fibers are oriented circumferentially and bridged by radially positioned “tie fibers”, creating a network with high tensile and shear resistance[18; 19]. Proteoglycan content in the deep layer provides compressive resistance to the tissue. Up to a third of the peripheral meniscal region is perfused while the rest is avascular, relying on nutrient diffusion from the synovial fluid[20].

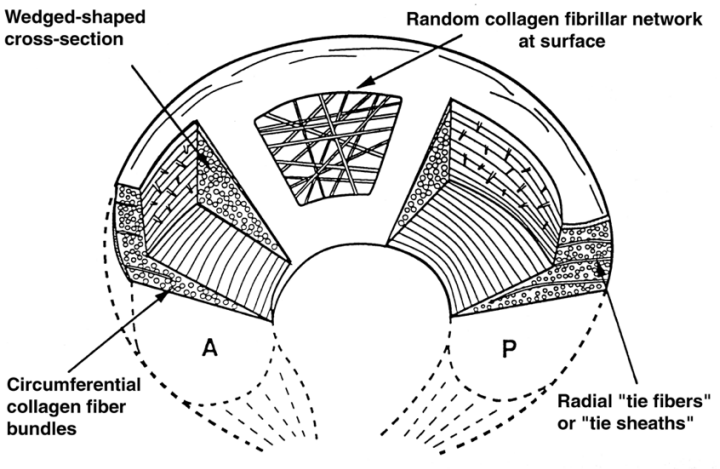


Figure 2.2 Schematic of zonal organization and collagen fiber arrangement in a healthy meniscus. A: Anterior, P: Posterior Reproduced with permission from Raven Press: Mow et al, Knee meniscus: basic and clinical foundations, 1992[21].

2.3 Intervertebral discs

Intervertebral discs are fibrocartilage structures that link the vertebra of the spine together and provide mechanical support by absorbing and transferring axial loads through the spinal column. The disc can be divided into three components: the annulus fibrosus (AF), the nucleus pulposus (NP), and the upper and lower cartilage endplates (CEP). The AF is a thick, fibrous ring surrounding a pressurized, gelatinous NP core. Approximately 15-25 concentric lamellae make up the AF, with each lamellar layer composed of thick type I collagen fiber bundles oriented at approximately 60˚ to the transverse plane, with alternating directionality between layers[22]. Like the structure of the deep layer in the meniscus, lamellae are bridged by a network of radially positioned collagen fibers[23], providing tensile strength and resistance to shear. Moving inward, type I collagen content decreases while type II collagen and proteoglycan content increase[24]. The AF is approximately 70% water, with a higher concentration of water in the inner AF[25; 26]. The NP is made of a loose meshwork of randomly oriented type II collagen fibers, a high density of proteoglycans, and is approximately 80% water[25]. The strong negative charge on proteoglycans, specifically aggrecan, maintains NP hydration. This results in high hydrostatic pressure that allows the NP to transfer axial compressive loads to “hoop” stresses in the AF and vertical loads to the CEP. The CEP is a thin 0.6mm layer that caps the upper and lower ends of the vertebral bodies[27], acting as an interface between the AF, NP and bone. Intact CEP and AF encapsulate the NP functioning as a passive immune barrier. Type II collagen fibers run parallel and perpendicular to the CEP surface, anchoring the AF lamellae and the NP mesh to the underlying bone[27]. Healthy adult intervertebral discs are largely avascular, as vascularization has only been observed in a small region of the outer AF[28]. Intervertebral disc cells rely on diffusion of nutrients from the adjacent vertebra through the CEP[29; 30].

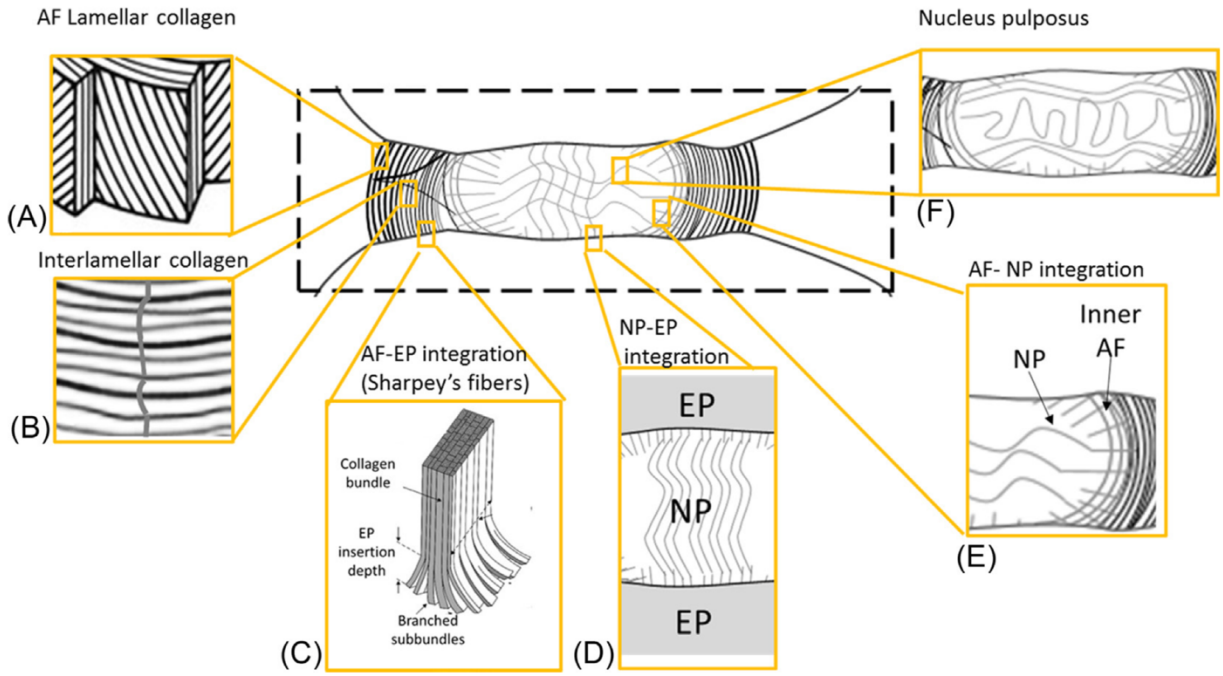


Figure 2.3 Schematic of intervertebral disc structure by disc regions. Reproduced from Sharabi et al., The Mechanical Role of Collagen Fibers in the Intervertebral Disc. 2018 with permission[31]. NP: nucleus pulposus, AF: annulus fibrosus, EP: endplate

2.4 Bones

In contrast to the soft tissues described above, high cell density and rich vascularization provide bones with strong regenerative potential. Modeling refers to the formation of new bone (osteogenesis) while remodeling describes changes in internal tissue organization and composition, both of which are regulated by metabolic and mechanical factors. There are two bone compartments: cortical and trabecular. Cortical bone lines the bone exterior, has low porosity and is responsible for providing mechanical support in the appendicular skeleton[32; 33]. It is composed of highly oriented, densely packed type I collagen fibrils, hydroxyapatite, and 23% water[33]. A 4-14 mm thick layer of cortical bone surrounds the tibial and femoral diaphyses[34; 35] while a thinner layer exists at epiphyses, including at the interface with the calcified cartilage layer. Similarly, a thin 0.244-0.290 mm shell of cortical bone surrounds the vertebrae[36]. Trabecular bone (also called cancellous bone) is localized in the bone interior and is a highly porous, complex meshwork of rodlike and platelike structures called trabeculae. Trabeculae are

similarly composed of hydroxyapatite, collagen type I, and 27% water[33]. However, in-vivo, trabecular bone is filled with a mixture of fat, water, and proteins from bone marrow which has important implications for the imaging methods discussed in Chapter 3. In vertebrae, trabecular bone bears most of the compressive load while in the appendicular skeleton, it provides secondary support to the cortical bone[32; 33]. During bone remodeling, it has been observed that trabecular bone microstructure aligns itself with the principal stress axes[37].

There are three main bones in the knee joint: femur, patella, and tibia. Chapters 5 and 6 will focus on the cartilage compartments of these bones as well as the meniscus. The spine is made of up of approximately 33 fused and non-fused vertebrae, divided into 5 regions: coccyx (4), sacral (5), lumbar (5), thoracic (12), and cervical (7). A 3D render of the tissues discussed in this dissertation is visualized in Figure 2.5. Chapter 7 studies the lumbar intervertebral discs, while Chapter 8 examines global spine shape dictated by the 3D position of the thoracic and lumbar vertebrae.

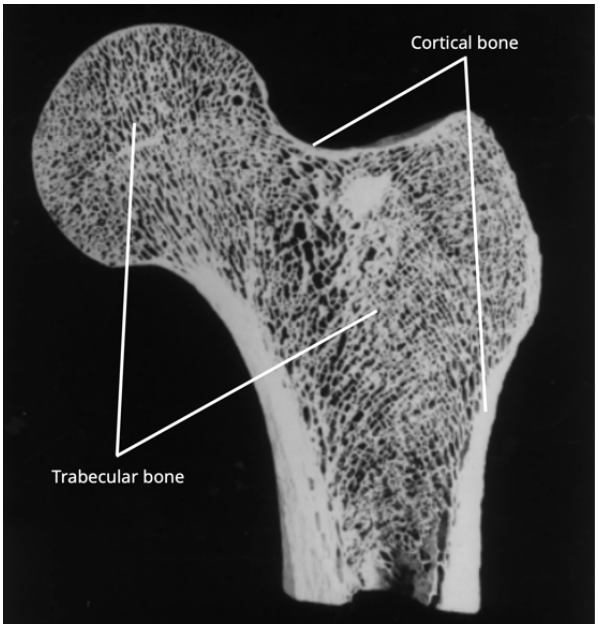


Figure 2.4 Axial cross section of a proximal femur highlighting regions with cortical and trabecular bone structures. Patterns in alignment of trabeculae follow principal loading directions. Reproduced with permission from Lovejoy et al. The maka femur and its bearing on the antiquity of human walking 2002[38].

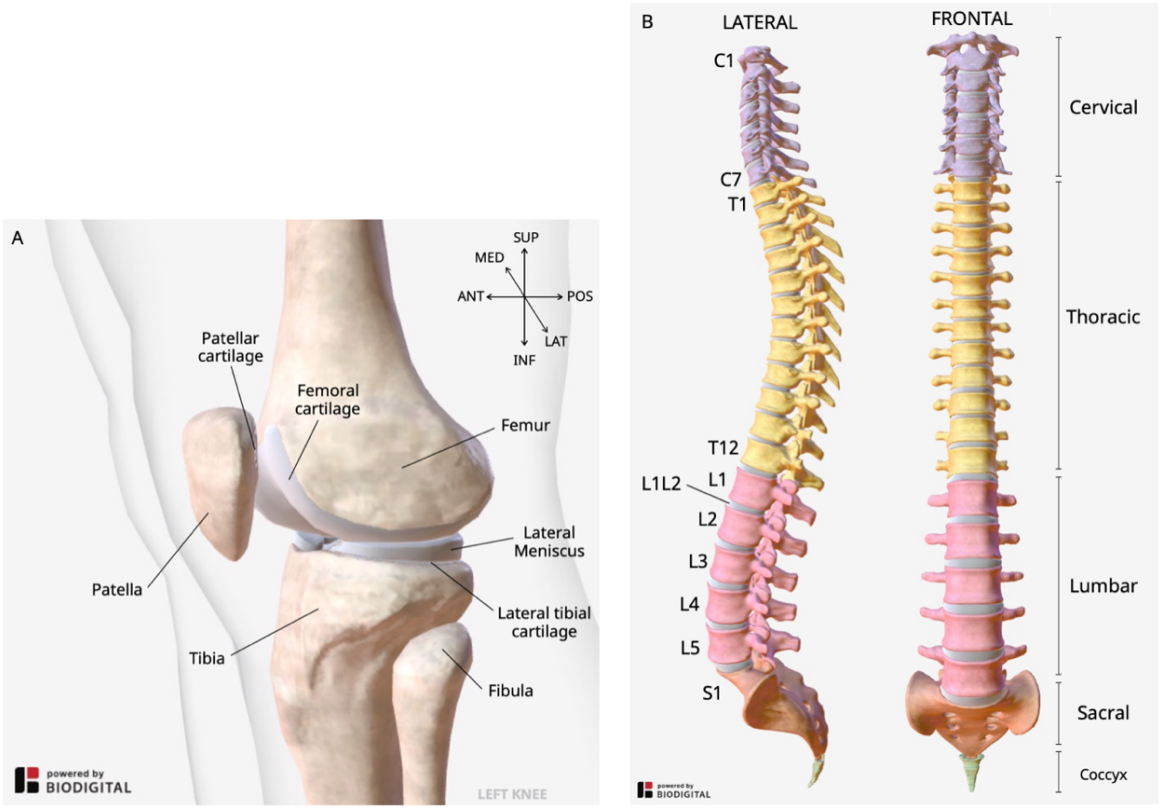


Figure 2.5 Volumetric rendering of the relevant knee and spine tissues. A) Oblique sagittal view of the left knee with bones, cartilage, and menisci visualized; synovium, ligaments, tendons, muscles not shown. B) Lateral and frontal views of a spine with anatomically normal, color coded by spine region. Intervertebral discs and vertebra are visualized; spinal cord, ligaments, and muscles excluded for clarity. Visualizations reproduced with permission from BIODIGITAL.

2.5 Pathophysiology of common musculoskeletal disorders

Low oxygen tension and lack of vascularity in cartilage, meniscus, and intervertebral discs contribute to their limited regenerative potential[39; 40]. It follows that maintaining tissue homeostasis and immune privilege is integral to preserving tissue structure and biomechanical function. Contrary to the “wear and tear” hypothesis, tissues dynamically respond to their environment. Tissue health depends on a delicate balance of cellular anabolism and catabolism which can be disrupted by trauma, abnormal mechanical loading, metabolic changes, or cell senescence[41-45]. Perturbations in tissue homeostasis can initiate a degenerative cycle: a positive feedback loop between cells, extracellular matrix, and biomechanics. This disease

model was proposed for degenerative disc disease (DDD)[46] but is generally applicable to other degenerative musculoskeletal diseases such as osteoarthritis (OA)[47] and adult spinal deformity (ASD)[48].

2.5.1 Osteoarthritis

Knee osteoarthritis is considered a heterogeneous disease with different mechanistic phenotypes based on the cause of initial tissue disruption, from mechanical overloading to low-grade systemic inflammation[42; 47; 49]. Etiology aside, osteoarthritis presents similarly in the joint as cartilage loses its structural integrity in stages. First, collagen in the superficial zone becomes disorganized which results in a decrease in proteoglycan content and increase in water content, effectively increasing tissue permeability and decreasing hydrostatic pressure[14]. Fraying of the superficial layer compromises its ability to withstand shear forces which leads to flaking of the cartilage surface and eventually fissures extending into the transitional zone[50]. Abnormal mechanical loading is propagated through all the cartilage layers up to the subchondral bone, activating a cellular response that attempts to stabilize the tissue initiating a vascular invasion of the cartilage[14; 51]. Eventually, further disruption of collagen organization and loss of proteoglycan and water content lead to thin cartilage and areas of exposed bone. Menisci play an interesting role in OA, as meniscal damage can lead to the development of OA in otherwise healthy knees while knee OA can degenerate a healthy meniscus[52], this is likely explained by abnormal biomechanical loading in both scenarios. Disorganization of collagen fibers causes fraying of the meniscus, starting at the deep layer then moving towards the surface[53], resulting in a loss of tissue tensile strength. Proteoglycan content increases, type I and II collagen content decrease, and water content is unchanged. In later stages of degeneration, meniscal tears in the avascular region are frequent[53].

2.5.2 Degenerative Disc Disease

Changes in the NP are typically the first to occur in intervertebral disc degeneration in response to biomechanical or chemical disruption. NP cells shift from producing type II collagen to type I collagen. Aggrecan becomes fragmented, losing its ability to bind water[26], causing a decrease in hydrostatic pressure and an increase in shear stress to the NP and AF[54; 55]. Loss of disc height reduces axial tension in the AF causing bulging of the inner and outer lamellae and lead to instability of the vertebra-disc motion segment[56]. The NP is unable to evenly distribute loads and increased localized stresses on the AF and CEP[57] lead to annular tears or endplate fractures[22]. In turn, damage to the annulus and endplate can compromise NP cells' immune privilege[58], triggering a strong inflammatory response. In late-stage degeneration, the disc is dehydrated, severely narrowed, and disc regions are indistinguishable as the AF and NP are entirely fibrotic[26; 54]. There is significant vascular ingrowth to the disc and mineralization of the CEP.

2.5.3 Adult Spinal Deformity

Degenerative changes in the intervertebral discs and vertebrae along several segments of the spine can cause global spinal instability. Adult spinal deformity is a broad spectrum of conditions related to the abnormal curvature of the thoracic and lumbar spine. Degenerative deformities are the most common type of adult spinal deformity[59]. As discussed, local degenerative changes due to aging occur in the intervertebral disc[54], which change the distribution of mechanical loads[55]. A portion of the mechanical load is shifted to the facet joints and anterior vertebral bodies, causing local bone remodeling and a slight change in global balance[60]. These maladaptive loading patterns exacerbate tissue degeneration, as paraspinal muscles compensate for this change in balance by opposing forward motion[61] which exerts additional compressive stress on the weakened intervertebral discs. Changes in vertebra shape and bone marrow are also observed[62]. Severe spinal deformities interfere with normal

biomechanical function: adults with deformities report lower physical function scores in health-related quality of life metrics, with the anatomical location of the deformity (lumbar, thoracic) and time of onset influencing these metrics[63].

2.5.4 Adolescent Idiopathic Scoliosis

Spinal deformities can also present in the pediatric population although the underlying pathophysiology is often different from adults. For example, in adolescent idiopathic scoliosis (AIS), while the exact etiology is not known, the early stages of the deformity are not characterized by intervertebral disc or vertebral degeneration. However, mechanical loading does play a part[64], as evidence suggests a mismatch between skeletal growth and muscular development reduces axial loading which triggers cells to produce more proteoglycans. Disc pressure and disc height increase, placing large amounts of tension on the paraspinal ligaments, limiting their ability to grow[65] and vertically locking the spine into place causing torsion and bending (differential growth hypothesis)[66]. However, the Disc and vertebra degeneration have been observed in AIS patients[67], but are believed to be secondary to the deformity rather than a cause of the deformity. Calcification of the CEPs along with lower water and proteoglycan content in the NP have been observed in scoliotic discs[67; 68] resulting in reduced cell viability[67]. AF collagen and elastic fiber organization is also disrupted[23], particularly at the curve apex on the convex side of the curve[68].

Osteoarthritis, degenerative disc disease, and spinal deformity are common causes of knee and back pain. Early to mid-stage therapeutic interventions for these diseases are largely unsuccessful at halting or reversing degeneration but can provide pain relief[47; 69; 70]. Surgery – knee arthroplasty or spinal fusion– are effective procedures to treat end-stage disease but carry significant risk of complications. Better understanding of disease subtypes and progression could assist in the development of disease-modifying drugs. The next chapter will introduce imaging methods for the characterization of musculoskeletal tissue in health and disease.

3 Clinical and quantitative imaging

Osteoarthritis (OA) and back pain are typically assessed via non-invasive imaging where clinicians search for pain generating structures by examining anatomy for signs of degeneration. The basic physics of x-ray and magnetic resonance (MR) imaging will be presented, with particular emphasis on the appearance of the tissues introduced in Chapter 2 and how they are clinically evaluated. Then, quantitative MR imaging will be discussed, highlighting its benefits, and identifying key roadblocks preventing its integration into the clinical workflow. This will set the stage for Chapter 4, where convolutional neural networks for image analysis are introduced.

3.1 X-ray physics

There are three components to clinical x-ray imaging: x-ray generation, attenuation, and detection[71]. An x-ray source is made of a vacuum chamber with a cathode, which supplies electrons, and an anode, which serves as electron target, held at a potential difference. Current is run through the cathode filament such that it heats up and releases electrons. These electrons gain kinetic energy as they accelerate towards the anode target. Upon colliding against the metallic anode, the electrons' kinetic energy is converted into electromagnetic radiation, specifically heat and Bremsstrahlung radiation (x-ray beam). The energy of the beam is modulated by the potential difference of the anode and cathode, the current supplied to the

cathode filament, and the material composition of the target anode. The x-ray beam characteristics– energy, beam shape, intensity distribution– can be further modified through collimation and beam filtration. The x-ray beam then interacts with the patient, as photons pass through, scatter, or are absorbed by the body. A screen-film or digital detector is placed on the other side to detect outgoing photons and capture a 2D projection image of patient anatomy. Photon interactions with tissue are determined by the tissue's attenuation coefficient and geometry. Moreover, a portion of these interactions produce energetic electrons that cause damage to surrounding tissue though ionization[72]. High intensity values on x-ray images correspond to anatomical regions of high attenuation, low intensity to regions of low attenuation. In the 10-80keV energy range used for x-ray imaging, bone is well visualized since bone mass attenuation coefficients are approximately 0.20-29 cm^2/g, which is higher than the surrounding tissues. For comparison: water 0.18-5.3 cm^2/g, fat 0.18-3.2 cm^2/g, and air 0.17-5.1 cm^2/g[73]. Soft tissues such as cartilage, menisci, and discs have very low attenuation (0.18-5.3 cm^2/g) which result in extremely poor soft tissue contrast on x-rays.

3.2 Clinical x-ray imaging

X-ray imaging is frequently used in clinical settings, as it is one of the fastest and most cost-effective imaging modalities, although precautions are taken to limit a patient's cumulative exposure to ionizing radiation. A routine x-ray imaging series to assess knee degeneration includes frontal and lateral views. Radiographs are qualitatively examined for degeneration, mainly tibio-femoral joint space narrowing, presence of osteophytes, subchondral sclerosis, and bone deformity. For epidemiological studies, the Kellgren-Lawrence system[74] is used to grade these changes and define the presence of radiological OA. In some research studies, measurements of joint space width are recorded[75; 76]. However, the utility of joint space measurements for knee x-rays has come into question as the apparent joint space width is extremely sensitive to changes in projection geometry due to patient positioning[77].

Frontal and lateral view radiographs are acquired for specific regions of the spine: sacral, lumbar, thoracic, or cervical. Radiologists inspect images for signs of disc narrowing, endplate calcification, vertebral dislocation (listhesis), fractures, and osteophytes. Observations are rarely detailed at the disc level and only remarkable findings are reported. In the lateral sacral / lumbar view, measurements for pelvic parameters (pelvic tilt, sacral slope, pelvic incidence) and listhesis are sometimes recorded. Research studies favor the use of magnetic resonance imaging over x-ray imaging for the assessment of local degenerative changes in the spine due to improved soft tissue contrast[78]. For global degenerative changes, however, weight-bearing full spine frontal and lateral radiographs are the gold standard as they capture changes in global posture[79]. Radiologists qualitatively report on the location and severity of abnormal spine curvature and well as global balance. Spine curvature measurements, such as Cobb angles[80], and classification for curvature types (Lenke[81], SRS-Schwab[82]) are not routinely used in the clinic but are heavily relied on by orthopedic surgeons during treatment planning[83].

3.3 Physics of magnetic resonance imaging

Magnetic resonance imaging (MRI) is the study of the magnetic properties of atomic nuclei. Several biologically relevant nuclei with odd atomic number possess magnetic moments including ^{1}H, ^{23}Na, ^{31}P. The nucleus of the hydrogen atom (^{1}H, a single proton) has the highest intrinsic magnetic moment (2.79) and is the most abundant in the human body, making it the ideal candidate for medical imaging. While magnetic moments from individual nuclei are undetectable, signals from collections of nuclei (~10^{15}) can be treated as a system by performing the vector sum of all magnetic moments. Under normal conditions, individual magnetic moments are randomly oriented, resulting in zero net magnetization. In the presence of a strong, external magnetic field (B_0), magnetic moments continue to be randomly oriented but show a slight tendency to point along the magnetic field line. This slight tendency results in a net non-zero magnetization M_z.

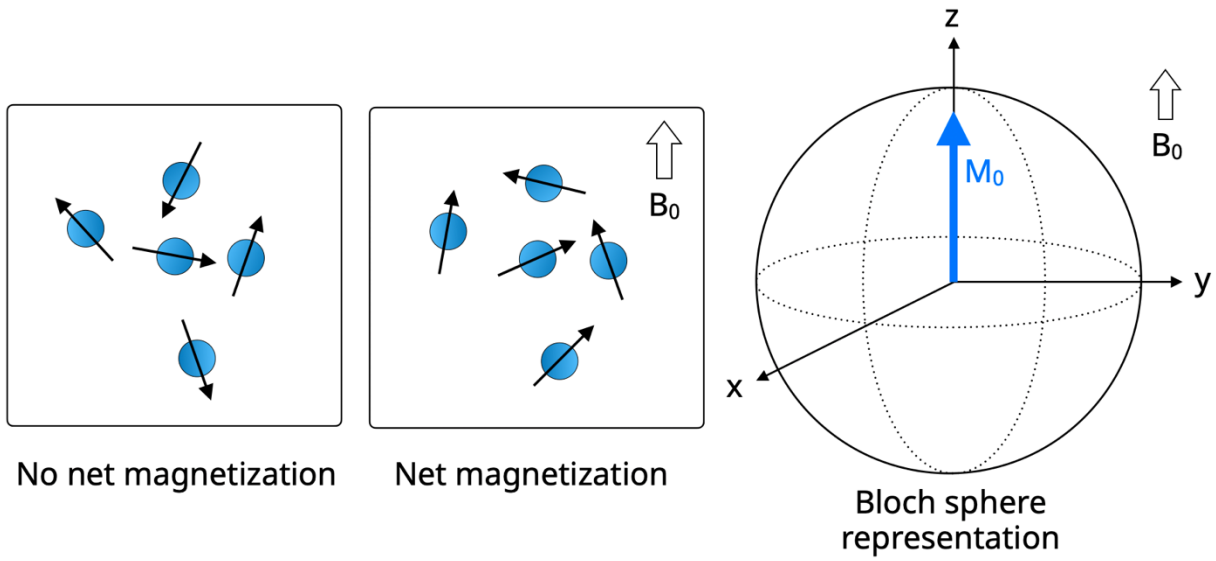


Figure 3.1 Net magnetization under a static magnetic field. Without the presence of an external magnetic field, magnetic moments are randomly distributed, resulting in zero net magnetization. Under a B_0 field, spins show a slight preference towards the magnetic field, resulting in a net magnetization in the direction of B_0. The collection of spins can be treated as a unit, whose net magnetization is a vector in a Bloch sphere.

Moreover, under a B_0 field, individual magnetic moments experience a torque, causing precession at the Larmor frequency. The Larmor frequency ω_0 is defined as the product of the nuclei's intrinsic gyromagnetic ratio and the external magnetic field strength. Precessing nuclei can be treated as a collection of oscillators that can absorb and emit energy at this resonant frequency. A short radiofrequency (RF) pulse tuned to ω_0 and perpendicular to B_0 is sent to excite the nuclei. The magnetic component of the RF pulse exerts a torque on the collection of nuclei, rotating the net magnetization M_z towards the transverse plane by an angle determined by the duration and amplitude of the pulse. M_z returns to equilibrium through relaxation processes. Signals in the transverse plane are detected by an external coil also tuned to ω_0. The design and timing of RF pulses, gradients, and signal acquisition contribute to the final contrast of the MR image. RF radiation is non-ionizing, as it is lower energy compared to x-ray radiation. Tissues absorb RF energy in the form of heat, therefore prolonged or repeated exposure to MRI poses comparatively minimal risk.

3.4 MR imaging contrast mechanisms

The structure and biochemical composition of a tissue determines its magnetic properties, and thus, its appearance on a specific MR sequence. A simple model can help conceptualize MR contrast mechanisms from a non-phenomenological perspective: the collection of nuclei are the system and the environment is a source of noise[84]. Noise can exist at various frequencies, and different imaging sequences are sensitive to different ranges of noise. This is best visualized through the power spectral density function which describes the power of noise at each frequency. The spectral density function $J(\omega)$ for a tissue depends on its local composition and is a measure of how long a spin system takes to lose memory of its interactions. Macromolecular motion contributes to the noise spectrum in MR imaging acquisitions. These motions cause random field fluctuations which are correlated in time. The autocorrelation function compares the field at time t and time $t+\tau$, where the fields are most highly correlated for short τ, and least correlated for long τ, measuring the system's 'memory' of these interactions. Rapid fluctuations have a small correlation time (τc), while slow fluctuations have a long τc. There is a non-linear relationship between τc and a tissue's relaxation time and there is a certain correlation time at which relaxation is most efficient. The power spectral density function $J(\omega)$, is defined as twice the Fourier Transform of this autocorrelation function $G(\tau)$ [84]. Several other mechanisms contribute to the noise spectrum in MR imaging including dipolar interactions, chemical exchange, and J-coupling.

The noise spectrum for biological tissues is dominated by water. Water can exist in a bound state or a free state. Bound water exists as a hydration shell surrounding macromolecules (loosely or tightly bound), has restricted motion, and thus has a longer τ_c. Free water can freely diffuse out of the tissue and has a shorter τ_c. Water in intervertebral discs and cartilage is mostly free water (AF 89% and NP 97%, cartilage 96%[85]) with a smaller portion bound to collagen and proteoglycans matrix macromolecules.

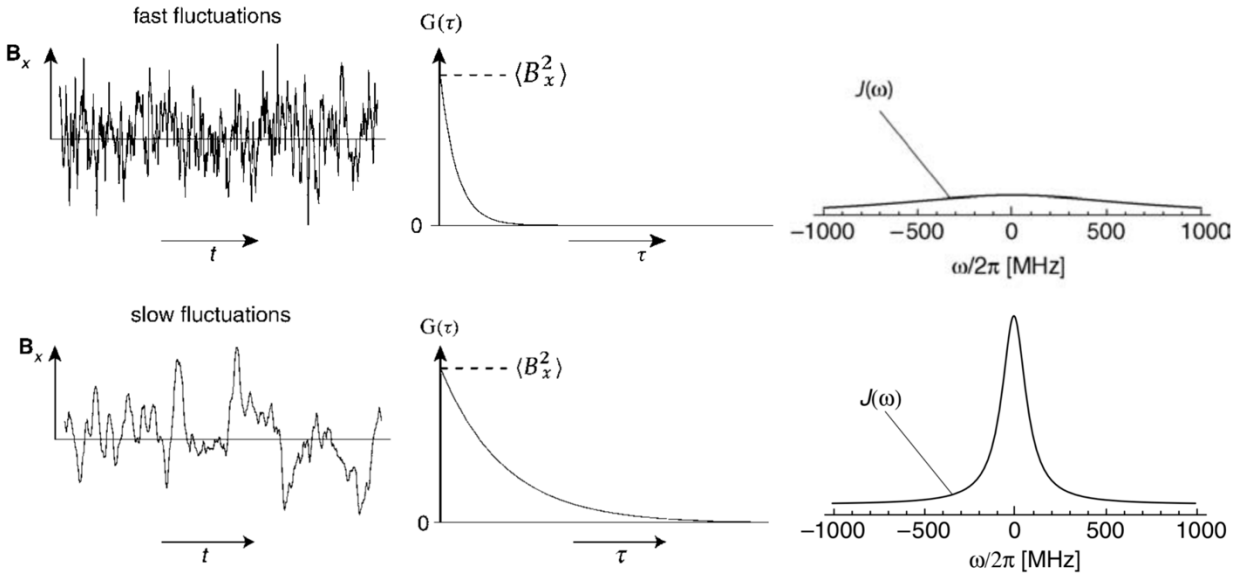


Figure 3.2 Spectral density functions for fast and slow fluctuations in local magnetic field. From L to R, field fluctuations, autocorrelation function, and spectral density for fast fluctuations (τc=0.2ns) and slow fluctuations (τc=2.0ns). Figure reproduced with permission of Wiley, from Levitt Spin Dynamics: Basics of Nuclear Magnetic Resonance[84].

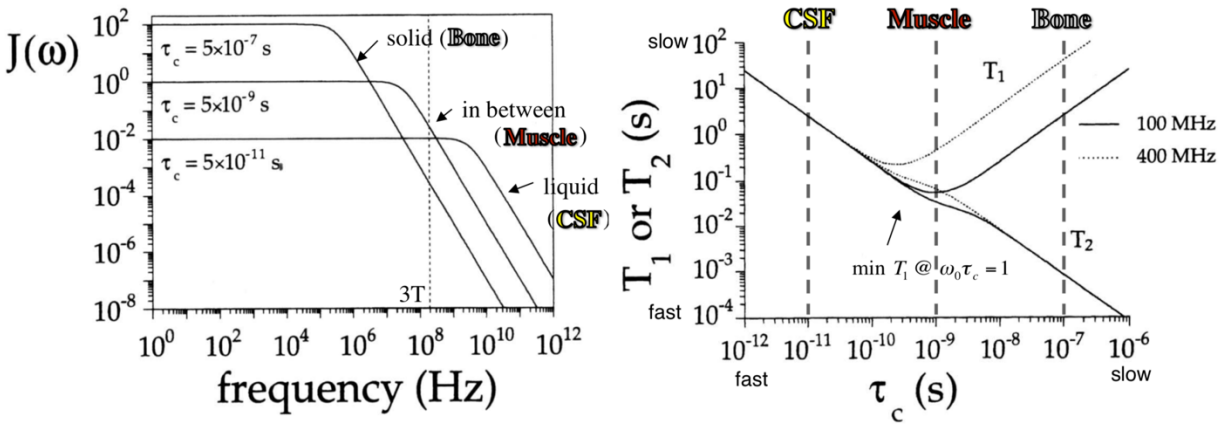


Figure 3.3 Log-log plots for power spectral density function and T_1/T_2 relaxation times vs correlation times of liquids, solids, and tissues in between. Liquids such as cerebrospinal fluid (CSF) have a shorter correlation time than solids like bone, values are approximate and depend on chemical composition of nearby tissue. Source: Prof Daniel Spielman, Stanford Rad226b 2016 Lecture 11 (https://web.stanford.edu/class/rad226b/Lectures/Lecture11-2016-T1rho.pdf)

T_1 relaxation probes the spectral density at ω_0. The higher the noise at this frequency, the more opportunities the nuclei will have to exchange their energy with the environment, thus the lower the T_1 relaxation time. Since the Larmor frequency is a function of the external magnetic field strength, it follows that T_1 relaxation measurements at varying field strengths (1.5T vs 3T) are sampling slightly different windows of the spectral density function. T_2 relaxation describes the loss of nuclei coherence due to a combination of energy relaxation (T_1) and pure dephasing (T_Φ). Pure dephasing is the dominant component in the T_2 relaxation processes because biological tissues have long T_1 relaxation times compared to T_Φ. Unlike T_1 relaxation, pure dephasing does not require energy transfer: pure dephasing (T_Φ) occurs when fluctuations in the local magnetic field cause fluctuations in the nuclei's ω_0. In the simplest T_2 setup (spin-echo experiment), a T_2 measurement probes the spectral density function both at the nuclei's Larmor frequency and at very low frequencies, therefore contains a T_1 and a T_Φ contribution. For the pure dephasing processes (T_Φ), the higher the noise near $J(\omega=0)$, the more quickly dephasing will occur, thus the lower the T_2 relaxation time. At a specific field strength, T_2 relaxation time of a tissue depends on both the acquisition sequence parameters and the intrinsic tissue parameters, unlike T_1 relaxation which is intrinsic to the tissue.

3.5 Clinical MR imaging

T_1 and T_2 relaxation times can be derived from specific MR imaging sequences (see Quantitative Imaging section below), however, scan times are long as they require the acquisition of several images. Instead, clinical imaging relies on the acquisition of a single image to examine relative, rather than absolute, signals from tissue. An image is said to be T_1, T_2, or proton-density (combination of T_1,T_2) weighted based on the tissue properties the acquisition parameters emphasize. The exact weighting is determined by the pulse repetition time (TR) and echo time (TE): T_1 weighting has short TR/ short TE, T_2 weighting has long TR/ long TE, and PD weighting has long TR/short TE. Fat appears bright on T_1-weighted

sequences. Yellow bone marrow is roughly 80% fat and has high signal intensity on a T_1 weighted image. Cartilage, menisci, discs, muscles, red bone marrow, and cerebrospinal fluid have low signal intensity on T_1 weighted sequences. Cortical bone, menisci, calcified cartilage, ligaments, and cartilage endplates have the lowest signal intensity. Specific pulse sequences for fat suppression can be applied to null the fat signal for contrast enhancement. Fat and free water appear bright on T_2-weighted sequences, therefore the disc nucleus, inner annulus, cerebrospinal fluid, and superficial cartilage have higher signal intensity on T_2 than on T_1. Cortical bone, menisci, calcified cartilage, and the cartilage endplate have very short T_2 relaxation times and appear dark on both T_2 and T_1 weighted sequences.

A standard degenerative knee MR imaging protocol includes proton-density (PD) weighted, T_1-weighted, T_2-weighted images in the sagittal and coronal planes, with and without fat suppression. Radiologists examine these images for signs of damage in the bone (bone marrow edema, osteophytes), cartilage (thinning, lesions, denudation), menisci (meniscal tears, cysts), and other soft tissues (ligament tears, synovitis). A small set of research studies have radiologists apply semi-quantitative scoring methods such as WORMS/MOAKS[86; 87] to evaluate the knee. An even smaller set of research studies perform a quantitative evaluation of the knee by means of cartilage/menisci thickness or volume measurements, given the time-consuming nature of the annotations.

Compared to knee imaging, spine imaging is slightly more varied as MR protocols are catered to specific radiological specialties and tend to focus on isolated anatomical regions. To give an example, a typical musculoskeletal lumbar spine protocol includes T_1-weighted and T_2-weighted sagittal and axial views, with and without fat suppression. Radiologists will look for signs of degeneration in the nucleus (loss of signal intensity, thinning, inhomogeneity), annulus (bulging, herniation), endplate (defects, calcification), and vertebra (marrow edema, inflammation, sclerosis). Semi-quantitative scoring methods such as Pfirrmann grading [88] for discs and Modic grading [89] for vertebra are not common in clinical practice. Similarly,

quantitative measurements of disc and vertebra volume are useful but are only performed for research studies. The use of MR imaging to assess global spinal deformities is not common, as it requires the acquisition of several localized views at oblique angles and therefore, has long scan times. Moreover, supine positioning during MR image acquisition does not reflect true sagittal spinal curvature.

3.6 Quantitative MR imaging

Standard MR imaging sequences are effective for the characterization of gross morphological changes in the knee and spine. However, as explained in Chapter 2, local biochemistry and structure determine tissue biomechanics. Therefore, it is desirable to detect early degenerative changes in microstructure (collagen organization) and biochemical composition (proteoglycan, collagen, and water content) before irreversible damage occurs. Several quantitative MR imaging sequences including T_2 mapping and $T_{1\rho}$ mapping have shown promising results for in-vivo characterization of musculoskeletal tissues. In contrast to standard MR imaging, quantitative MR imaging creates parametric maps of the tissues, where voxel values are measurements of tissue relaxation properties and can be used for cross-sectional or longitudinal comparisons. T_2 mapping and $T_{1\rho}$ mapping are used in Chapters 6,7 and will be introduced in detail.

3.6.1 T_2 mapping

T_2 and $T_{1\rho}$ relaxation time parameter maps are derived by acquiring several image snapshots that differ only by varying a single parameter in the acquisition sequence, echo time and spin lock time, respectively, then performing pixelwise exponential fitting on the set of images to calculate the exponential decay constant for each pixel. Many T_2 mapping sequences are based on the Carr-Purcell-Meiboom-Gill (CPMG) sequence used for NMR spectroscopy [90]. When measuring the contribution of pure dephasing to the relaxation processes, CPMG probes the spectral density function near a frequency. First, a π/2 pulse along the x-axis rotates the net

magnetization onto the y-axis. Then, a train of equally spaced π pulses are applied along the y-axis, acting to refocus the dephased spins. The number of π pulses (N) and acquisition time determine the window of the spectral density function that will be probed. For a fixed experiment length, a greater number of pulses (N) will create a bandpass filter whose center is shifted to higher frequencies. For fixed pulse to pulse spacing, each additional pulse alters the width of the bandpass filter and shifts it towards 0 frequency. N=0 (no π pulse) is equivalent to a Ramsey acquisition, and N=1 (a single π pulse) is equivalent to a spin-echo acquisition. Experiment length is bounded by T_1 relaxation time, as signal intensity decreases over time and can reach a noise floor.

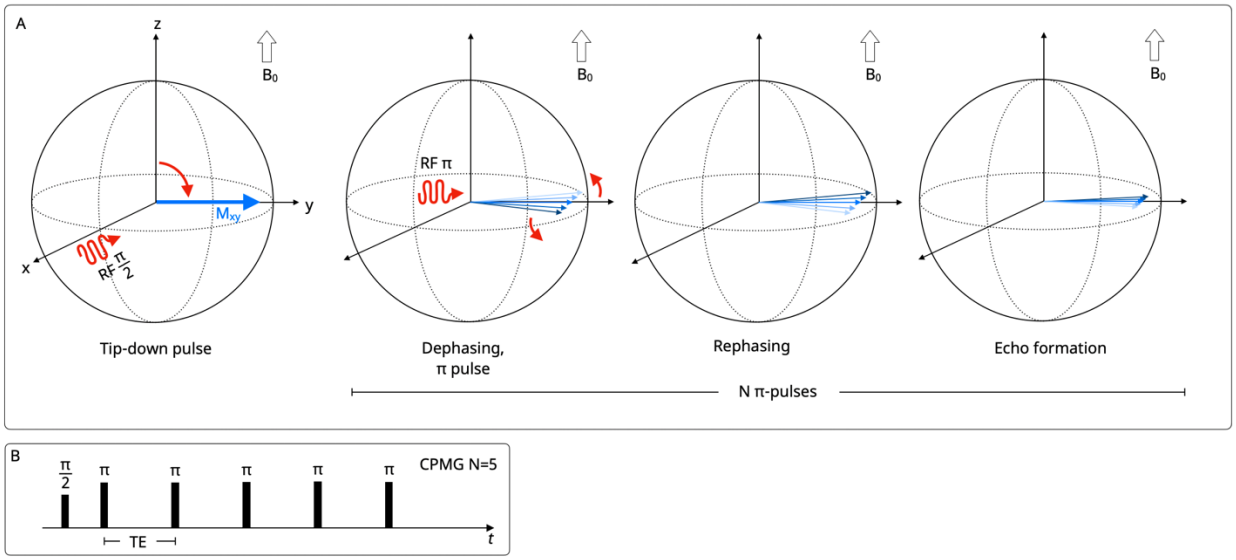


Figure 3.4 The Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence. A) Bloch sphere schematic representation of a single π pulse. M_0 magnetization is tipped down into the xy plane, and spins start to dephase. Then, a short π pulse flips the spins about the y-axis, spins rephase and form an echo. This process can be repeated for several π pulses, with M_{xy} net magnetization approximately exponentially decreasing with each additional echo. B) Simplified pulse sequence diagram for CPMG with 5 π pulses, each equally spaced by echo time (TE).

The maximum number of π pulses that can be delivered in a specific time frame is determined by the total RF energy deposited into the tissue (specific absorption rate limits avoid tissue overheating), as well as the RF pulse characteristics. Due to these constraints, CPMG based T_2 mapping sequences typically probe a spectral window close to 0. The higher the noise

at the spectral window probed, the more quickly nuclei will dephase and the lower the T_2 relaxation time. T_2 relaxation time depends on the specific acquisition sequence, mainly, the number and spacing of pulses. An important consideration for T_2-weighted images and T_2 mapping is the magic angle effect. The magic angle effect is a localized increase in T_2 relaxation time when regions of highly organized collagen are aligned 55˚ relative to the main magnetic field B_0. At this alignment, dipole-dipole interactions due to B_0 field are minimized and T_2 becomes longer. Cartilage has a highly organized, layered structure and is located on a surface with high curvature (distal femur), which makes it susceptible to magic angle effects.

When normal values for T_2 relaxation times of biologic tissues are reported in literature, they are derived from a variety of CPMG-based sequences (including Fast Spin Echo) which probe slightly different ranges of the spectral density function depending on acquisition parameters. It follows that correlations between T_2 relaxation values and proteoglycan/collagen content will vary with tissue orientation, pulse sequence, and biochemical assay.

T_2 relaxation times in disc and cartilage are widely regarded as proxies to quantify proteoglycan, collagen, and/or water content, however literature on the relationship between T_2 values and tissue biochemistry is inconclusive. In cartilage, T_2 relaxation times have shown moderate[91-93] and no[94; 95] correlation with proteoglycan content. Studies have shown moderate [93] and no [92; 94] correlation between cartilage T_2 relaxation times and collagen content measured with biochemical assays. In-vivo, higher cartilage T_2 relaxation times have been observed in patients with OA[96; 97] as water that was restricted by the highly organized collagen structure and bound to proteoglycans becomes mobile. There is evidence to suggest T_2 values sensitive to changes in cartilage collagen content, as T_2 values increased after collagenase treatment[98]. In the meniscus, correlation with proteoglycan content is moderate and correlation with collagen content is weak[95].

In intervertebral disc nucleus, T_2 relaxation times showed strong[99; 100], moderate[101], and no[102; 103] correlation with proteoglycan content. T_2 relaxation measurements in the annulus were

moderately[100] and not[102] correlated with proteoglycan content. Whole disc measurements showed strong correlation between T_2 and proteoglycan content[99]. Nucleus and whole disc measurements in one study showed moderate negative correlations between collagen content and T_2 relaxation time (r=-0.554 and r=-0.735, respectively)[99]. A different study also found a moderate positive correlation between $1/T_2$ and nucleus and whole disc collagen content (r=0.532, r=0.672) [102]. In-vivo, degenerated discs have lower nucleus T_2 relaxation values than healthy discs.

3.6.2 $T_{1\rho}$ mapping

T_2 relaxation sequences are sensitive to the orientation-dependent magic angle effect and are not designed to probe processes in the higher spectral range (0.1-3kHz). Continuous wave $T_{1\rho}$ measurements present a feasible alternative to sample different windows of the spectral density function and minimize the contribution from dipolar interactions. $T_{1\rho}$ is often called “spin-lattice relaxation in the rotating frame” but it is more accurately described as transverse relaxation in the presence of a continuous RF pulse. First, a π/2 pulse along the x-axis rotates the net magnetization onto the y-axis. Then, a continuous RF pulse is sent along the y-axis, preventing the spins from precessing away and ‘locking’ magnetization to rotate about the y-axis for a specified spin lock time (TSL). The carrier frequency of the RF pulse is the Larmor frequency (ω_0) and the spin-lock frequency (FSL) is determined by the amplitude of this pulse. Similarly to the CPMG sequence, the spin-locking pulse refocuses the dephased magnetization but does so in a continuous manner.

By creating a local magnetic field B_1, the spin-lock pulse determines the window of the spectral density function nuclei relaxation will be sensitive to, or the “filter function”. In a continuous wave $T_{1\rho}$ experiment, the FSL is fixed and several spin lock times TSL are acquired. $T_{1\rho}$ relaxation has an energetic exchange component near the Larmor frequency and a dephasing component at the user specified frequency (FSL). The greater the noise at the

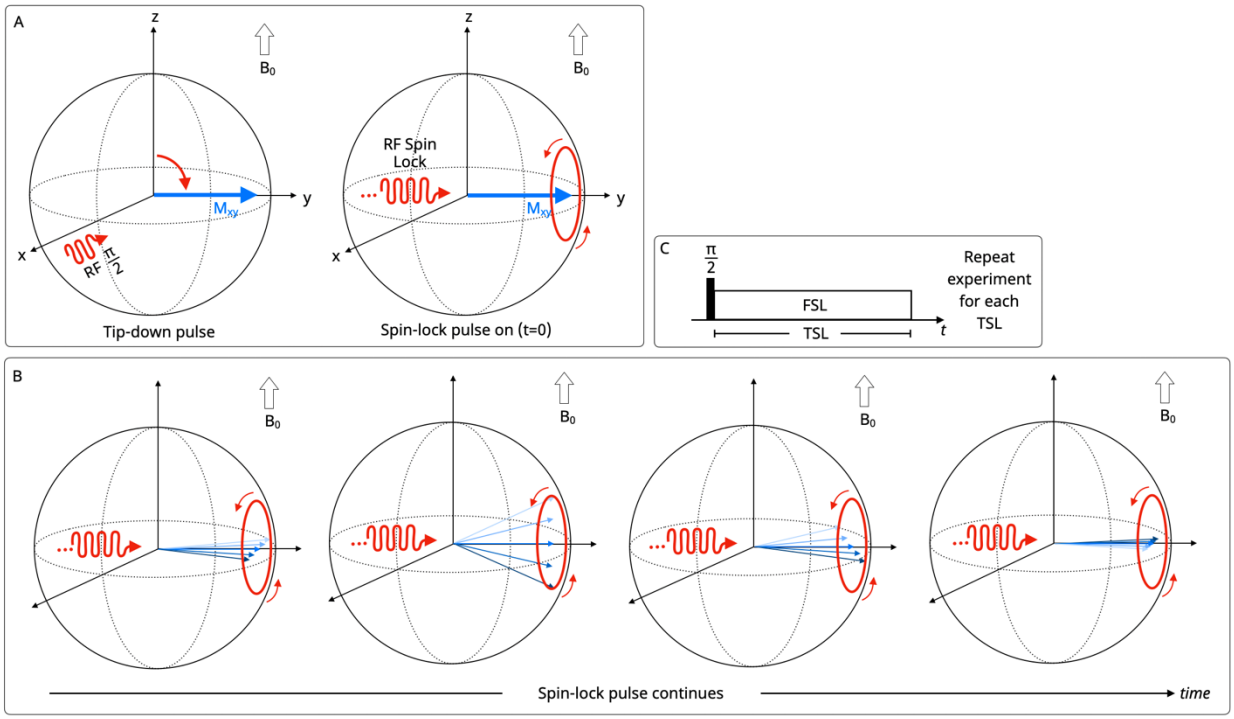


Figure 3.5 Spin-relaxation under a continuous RF pulse ($T_{1\rho}$). A,B) Bloch sphere representation of the spin-locking pulse. First, magnetization is tipped down into the xy plane by a short RF pulse, then a continuous spin-lock pulse is applied along the y axis, causing the spins to dephase and rephase locally. Mxy decreases over the duration of the spin lock pulse, approximately following an exponential decay. C) A simplified pulse sequence diagram for $T_{1\rho}$: after the tip-down pulse, a pulse at a desired spin lock frequency (FSL) is applied for a set time (TSL). The experiment is repeated for several TSL and fit using an exponential decay model.

specified spectral window, the faster the dephasing and the lower the $T_{1\rho}$ relaxation time. Several modifications to the basic $T_{1\rho}$ sequence have been devised to increase robustness to imperfections in RF pulses and B_0 field inhomogeneities, including the addition of refocusing echoes and composite phase pulses (reviewed in Chen 2015[104]). At higher spin lock frequencies (1kHz, 2kHz) the spectral range probed is less affected by dipolar interactions, which in turn minimizes the sensitivity of the imaging sequence to magic angle effects. This is nicely illustrated in a study by Hanninen et al[105] examining the orientation dependence for T_2 and $T_{1\rho}$ at different spin lock frequencies. Increased specific absorption rates at higher spin lock frequencies often limit the spin-lock frequencies feasible to apply in-vivo. An alternative to

continuous wave $T_{1\rho}$ is Adiabatic $T_{1\rho}$, where the TSL is fixed and the FSL is modulated. The clinical utility of adiabatic $T_{1\rho}$ over continuous wave $T_{1\rho}$ has not been established.

Empirically, literature has found mixed associations between continuous wave $T_{1\rho}$ and matrix content. In cartilage, $T_{1\rho}$ relaxation times have shown to be strongly[106], moderately[91; 94], and not correlated[107] with proteoglycan content. However, several studies have reported $T_{1\rho}$ is sensitive to changes proteoglycan content[106; 108; 109]. Studies have not found a correlation between cartilage $T_{1\rho}$ and collagen content[94; 107]. In menisci, $T_{1\rho}$ is moderately correlated to proteoglycan content, and moderately correlated to collagen content[95]. The correlation with proteoglycan content is negative: as cartilage and menisci become degenerated and lose proteoglycans, $T_{1\rho}$ relaxation times tend to increase.

In the intervertebral disc nucleus, $T_{1\rho}$ shows strong positive correlations with proteoglycan content[101; 110; 111]. As the nucleus becomes degenerated and proteoglycan content decreases, $T_{1\rho}$ relaxation times tend to decrease. $T_{1\rho}$'s correlation with collagen content has not been verified with biochemical assays. $T_{1\rho}$ measurements have been correlated to radiographic measures of degeneration such as WORMS grades in knee OA for cartilage and meniscus[112; 113] or Pfirrmann grades for disc degeneration [114].

Overall, mixed results in literature are unsurprising, as different experimental setups for $T_{1\rho}$ measurement fundamentally probe different ranges of the power spectral density function, this includes variations in sample source (human/bovine), scanning in-vivo/in-vitro, biochemical assays, B_0 strength, FSL, tissue orientation, and pulse sequence. Readers are referred to work by Rautiainen et al where cartilage tissue is examined with multiparametric imaging (Figure 3.6).

To summarize, x-ray imaging is well-suited for imaging of gross structural changes in the musculoskeletal system but provides limited soft tissue contrast. MRI, on the other hand, is a powerful imaging modality for the characterization of cartilage and disc structure and composition[115]. Both x-ray and clinical MR imaging are routinely used in practice, yet images are analyzed qualitatively, as quantitative annotations are time-intensive and require clinical

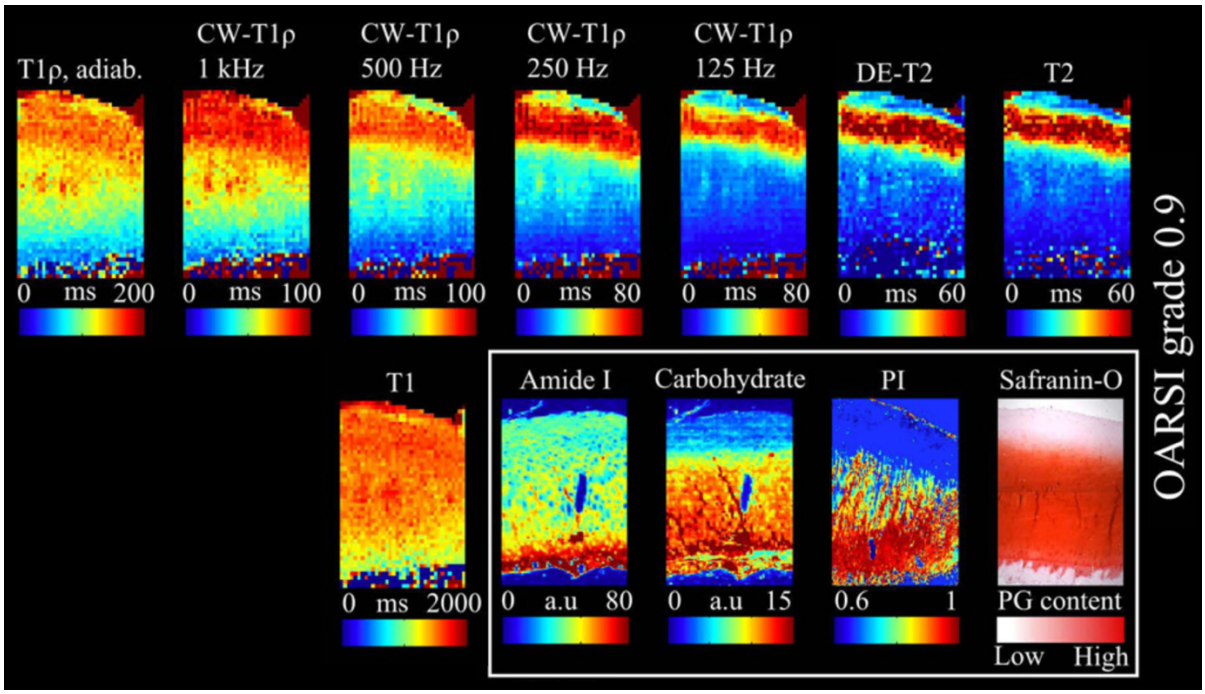


Figure 3.6 Multiparametric imaging of cartilage sample from a human knee with early stages of osteoarthritis, indicated by the loss of proteoglycan in the superficial layer. Top row: $T_{1\rho}$ and T_2 imaging sequences. As the spin lock frequency for CW- $T_{1\rho}$ decreases, $T_{1\rho}$ contrast starts to resemble T_2 contrast, with slightly higher relaxation times. T_2:T_2 echo, DE-T_2: adiabatic double-echo T_2 Bottom row: T_1 relaxation sequence, amide I (collagen content), carbohydrate absorbance (proteoglycan content), parallelism index (collagen anisotropy), and tissue histology (proteoglycan content). Absorbance maps acquired using Fourier transform infrared imaging. Reproduced with permission from Rautiainen et al 2015[93].

expertise. Quantitative MR imaging is not routinely used in clinical practice, as acquisition, post-processing, and analysis methods are complex and time-consuming. In this Chapter, we have introduced two imaging modalities, explained their underlying physics and contrast mechanisms, and explored their role in the diagnosis and characterization of knee osteoarthritis and spinal degeneration in clinical and research studies. Imaging information is underutilized, and there is a need to improve image analysis sensitivity, speed, and precision. In this dissertation, we develop fully automatic tools for the analysis of medical images. The next chapter will introduce the technical advancements that have spurred the development of several fully automatic analysis tools for medical imaging, specifically representation learning.

4 Representation learning in medical imaging

Over the past two decades, hardware advancements have vastly increased the efficiency of computers. This increase in computational power, coupled with wide clinical adoption of digital picture archiving and communication systems (PACS), has enabled researchers to develop automatic tools for medical image analysis. An introduction to representation learning will be presented, followed by an overview of problem formulation using convolutional neural networks and a discussion on longstanding challenges in the field of machine learning for medical imaging. The last part of the chapter will include a walkthrough of DioscoriDESS: our research group's modular, user-friendly codebase for developing deep-learning segmentation algorithms.

4.1 Representation Learning

The performance of machine learning algorithms relies heavily on the representation of the input data. Good representations– or features– capture the variations and structure in the input data that are most useful to the downstream predictor. In feature engineering for images, significant effort is dedicated to the design of filter sets to extract relevant features (textures, edges, averages). One such example are Gabor filters, consisting of Gaussian kernel functions of a set scale modulated by sinusoidal plane waves with specific rotations.

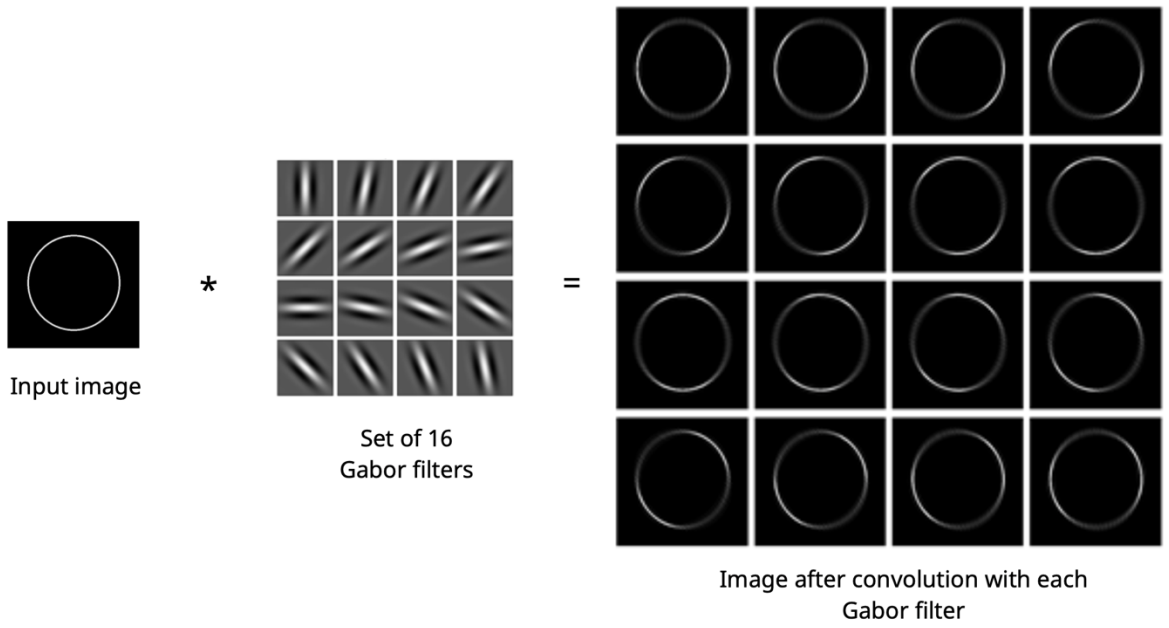


Figure 4.1 Example feature extraction using Gabor filters. Each Gabor filter is convolved with the input image to create the output feature map. Figure adapted from (https://medium.com/@anuj_shah/through-the-eyes-of-gabor-filter-17d1fdb3ac97)

Filters slide along the input image and convolve with the image patch to extract local texture information which is then used as input to a machine learning algorithm. Over the years, similar approaches have fallen out of favor, given the high manual effort and difficulty in determining optimal features a-priori. Rather than engineering features, representations can be learned from data. Representation-learning is broadly used to refer to the supervised or unsupervised process of discovering variations and underlying structure in the input data[116]. Unsupervised representation learning techniques do not have information about the downstream task and only rely on prior assumptions about data smoothness, sparsity, and natural clustering. In addition, for regularly sampled, spatially coherent data like medical images, the intrinsic dimensionality of the input is assumed to be much lower than the input dimensionality, i.e. data is expected to concentrate on a lower dimensional manifold. Some common unsupervised learning techniques include autoencoders[117], k-means clustering, and principal component analysis[118]. On the other hand, in supervised representation learning, the feature extraction and prediction tasks are

connected, where the optimal feature extraction approach is learned for the task at hand. Supervised learning techniques still exploit the prior assumptions discussed above, but benefit from access to labeled examples to guide the learning process. Most convolutional neural networks (CNNs) fall under this category. Instead of convolving hand-engineered filters with images, CNNs have learnable filters, whose weights are randomly initialized and automatically adjusted during network training to improve feature representation for the prediction task[116]. Abstract, non-linear representations are learned through the stacking of several convolutional layers, activation layers, pooling layers, batch normalization layers, and fully connected layers, from which the term 'deep' in deep-learning originates.

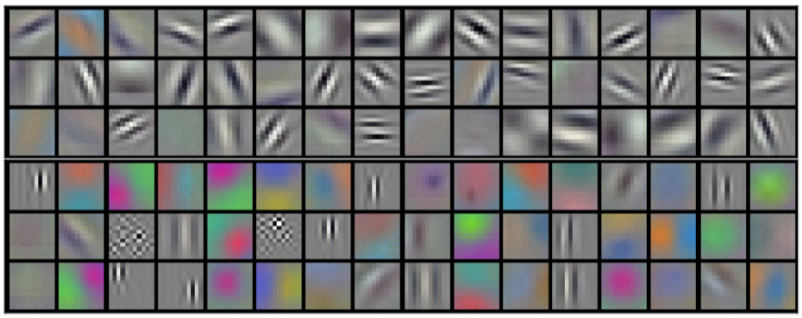


Figure 4.2 Convolutional kernels (filters) learned during training in Krizinshky et al.'s AlexNet. Each filter is size [11x11x3]. In addition to kernels resembling Gabor filters (at a variety of sinusoidal frequencies and rotations), the network learns filters for specific colors from the multichannel (RGB) image input.

Well formulated representation-learning approaches have outperformed methods with carefully engineered features across nearly all computer vision applications. This is evident as the top scoring teams in imaging challenges have used deep learning: a popular example is the Common Objects in Context (COCO[119]) challenge which has object segmentation, object classification, keypoint detection, and captioning tasks, all of which have been successfully tacked using deep learning. More broadly, this highlights two important points for discussion. First, optimal imaging features are more abstract and dataset dependent than previously thought. This suggests effort should be focused on problem formulation and engineering of effective representation-learning systems. Machine learning algorithms should adapt to or easily

be re-trained on changing clinical data, such as a change in MR receive coil or a shift in patient demographics). Second, expanding beyond the initial task scope, success with representation learning raises the question whether the learning task itself would also benefit from being redefined. This can include an upstream or downstream change in the final task and is dependent on the availability of annotated data and clinical utility of the final outcome. A clinical grade, such as Lenke classification[81] for spinal curvature in x-rays, may be an obvious target for a machine learning algorithm; when in reality, there exists a more informative intermediate representation, spinal contours, from which Lenke and other classifications could be derived. In a similar way, optimizing a task downstream of the original task can improve the accuracy of the final outcome. For example, starting with under-sampled K-space images of the knee, the most straightforward step is to learn a mapping between under-sampled and fully-sampled K-space images[120]. However, if the downstream application is to extract accurate knee imaging biomarkers, an end-to-end formulation can yield more effective feature representation as demonstrated by Caliva et al[121].

4.2 Convolutional Neural Networks

Machine learning can be divided into four major levels of abstraction: Application/Data, Model, Optimization Problem, and Optimization Algorithm. In the next section, these levels of abstraction will be used to explain the basic principles behind convolutional neural networks.

4.2.1 Application/Data

First, the target clinical task, available imaging data, and labels are defined. Data labels can be of several degrees of quality: gold standard (validated by biopsy, blood labs), multi-expert annotator, single-expert annotator, weak labels (automatically extracted from a radiology report, approximated by a different algorithm) or even unlabeled but categorically related. Labels are also of different granularity: patient-wide, image-wide, regional, or dense. During development, more than one problem formulation should be considered to maximize algorithm generalizability

and clinical utility. Data should be divided by patient into non-overlapping splits for algorithm development (training, validation) and testing, with particular attention given to balancing demographic characteristics and labels across splits. Differences between available imaging data and real-world clinical imaging data should be noted (population shift, prevalence shift, acquisition shift). Ideally, datasheets for datasets should be released alongside algorithms following guidelines outlined by Gebru et al[122].

4.2.2 Model

The structure of convolutional neural networks, specifically the concept of learnable filters, encodes implicit assumptions about imaging data. The network examines the input image through a set of small, local receptive fields and extracts elementary visual features (edges, corners, textures) into a feature map. These local feature extractors (filters) share learnable weights, enforcing prior assumptions about sparsity and translation *equivariance*. Translation *invariance* is encoded using pooling layers, as the feature representation should not be sensitive to the exact position of the object of interest in the input image. Other task-specific assumptions (priors) can be encoded into the CNN structure. The work by Winkels et al[123] provides an elegant example of 3D roto-reflection equivariant CNNs, achieved by performing geometrical transformations on the learnable filters before convolving with the image. A thorough presentation of geometric priors for deep learning is found in Bronstein et al[124]. Finally, the number of parameters, or network capacity, can determine the model's ability to fit to the data. The value of novel CNN architectures is usually demonstrated on specific benchmark tasks, such as image classification, but the convolutional backbone can be adapted to other imaging tasks.

One of the first CNNs was introduced by Yann LeCun in 1989[125]. LeNet5 was a simple, 7 layer feed forward network used for handwritten digit classification on 32x32 pixel images. CNNs received little attention over the next several decades as they were too memory intensive

for hardware at the time. This was until early 2010's when Krizinshky et al[126] entered AlexNet, an 8 layer CNN, into the 2012 ILSVR Challenge and won with a top-5 error rate of 15.3% compared to the 26.2% of the second best entry. To train on 224x224 sized images, AlexNet introduced several technical innovations, mainly: overcoming memory limitations by reimplementing the 2D convolution operation on a graphical processing unit (GPU), developing a layer-wise distributed training method, and proposing regularization techniques (dropout, data augmentation) to prevent the 60 million parameter network from overfitting.

Hundreds of CNN architectures and training strategies have been proposed in the last decade, three of which will be explained in detail due to their success in various medical imaging tasks: ResNet (Chapter 7), DenseNet (Chapter 8), and UNet (variant used in Chapter 5, Chapter 7).

ResNet[127] was proposed in 2015 by Hu et al to address issues with unstable gradients and high training errors (degradation problem) seen in increasingly deep neural networks. ResNet introduced residual connections, or "shortcuts", where the outputs of previous layers are summed with the outputs of the current layer block (2-3 convolutional layers), in essence, reformulating the network to learn a residual function or a perturbation of the input features. Residual learning resolved problems with gradients and facilitated training of increasingly deep architectures. In fact, a Resnet with 152 layers won the 2015 ILSVR Challenge* scoring a top-5 error rate of 4.49%. CNNs with ResNet backbones have shown promise in a wide range of medical imaging tasks from prediction of incident osteoarthritis from knee bone shape[128] to vertebral compression fracture detection on CT scans[129].

Layer-wise connections were further explored in 2017 by Huang & Liu[130]. DenseNet proposed to use direct connectivity between all layers in a dense block, where each dense block had convolutional layers with the same feature map dimensions. In contrast to Resnet, DenseNet features are concatenated rather than summed, which results in higher parameter efficiency as it encourages filter reuse. DenseNet authors claimed comparable performance to

ResNets with only half the number of trainable parameters. DenseNet backbones have been used extensively in medical imaging tasks, from hip fracture detection on x-rays[131] to prediction of total knee arthroplasty from MR images[132]. In practice, DenseNets are more parameter efficient than ResNets and thus less likely to suffer from overfitting, however, the dense connectivity places high demands on GPU memory during training.

UNet was published in 2015 by Ronnenberger et al[133] for microscopy image segmentation. Previous work by Long and Shelhamer[134] using fully convolutional networks for pixelwise segmentation laid the foundation for the success of UNet. Long and Shelhamer fused feature maps with varying degrees of spatial precision by combining the final prediction layer (spatially coarse, high-level abstract) with lower layers (spatially fine) enabling the network to use global and local information. UNet further formalized the concept of fusing multiple feature levels by introducing a U-shaped encoder-decoder structure with skip connections at each level. Specifically, through the decoder structure, high-level abstract features are progressively up-sampled, concatenated with spatially fine information from the encoder structure, and convolved to extract precise features for segmentation. In many ways, UNet's encoder with sequential pooling at each level is similar to a CNN for classification, an observation which has been leveraged in several studies since. Mehta et al used an additional output from the bottom of the encoding branch to perform classification and segmentation simultaneously (multi-task learning). The kernel weights for the encoder branch in segmentation networks can be initialized with weights from an image classification task. The impact of UNet cannot be understated, since its publication as a conference paper for MICCAI in 2015, the work has been cited over 25000 times. UNet and its variants have achieved state of the art performance in several medical imaging segmentation tasks including cartilage and menisci segmentation[135; 136]. Volumetric U-Net variants (3D-Unet, V-Net) are used widely in medical imaging. V-Net[137] improves upon UNet by learning a residual function at each encoder level and by replacing pooling and upsampling operations with convolutions, reducing the overall memory footprint.

*The ILSVR challenge ended in 2017, as 29/38 competing teams surpassed 95% accuracy and the organizers decided it was time to curate a more challenging dataset.

4.2.3 Optimization problem

Defining the optimization problem by building a loss function is one of the most important steps in problem formulation. Considering the available data and the granularity of the labels, one can define several imaging tasks: classification, regression, bounding box localization, keypoint localization, or segmentation, among others. Supervised representation learning requires a differentiable loss function to assess the similarity between network predictions and the provided labels. Common CNN loss functions include cross-entropy loss, mean-squared error loss, and Dice loss[137]. Loss functions can be composed of one or many loss terms with equal or weighted contributions. The selected optimization algorithm will aim to minimize this loss by finding feature representations that provide information to match the provided labels. Unlabeled but related images can also contribute meaningfully to training through consistency loss functions that exploit prior assumptions about data continuity, sparsity, and semantic similarity[138]. Training with labeled and unlabeled data is called semi-supervised learning, and requires that the input image P(data) contain information about the posterior distribution P(label|data)[139].

4.2.4 Optimization algorithm

The loss landscape for neural networks is smooth but highly non-convex, and neural network training is framed as an unconstrained optimization problem. Stochastic gradient descent (SGD) is one of the most popular optimization algorithms for this problem. In SGD, training samples are split into minibatches, where the upper limit for batch size is set by GPU memory, and is a function of input size and network size. Per batch, a forward and a backward pass are used to calculate partial gradients with respect to model weights. The weights are updated in the direction of steepest descent by an amount specified by the learning rate.

Learning rate schedulers and adaptive learning rates add robustness to training dynamics by preventing divergence of the optimization algorithm or convergence to a local minima. Learning rate schedulers lower or increase the learning rate as training progresses given a fixed schedule (exponential decay for example). Adaptive methods such as Adam[140] are considered variations of SGD as they compute adaptive learning rates for each parameter and include methods to dampen oscillations near local minima and lead to faster convergence.

4.3 Challenges in representation learning for medical imaging

Seminal papers in medical imaging have borrowed concepts from computer vision, yet there are unique challenges to working with medical data. Computer vision research strives to develop low-latency, low-footprint algorithms while medical imaging research prioritizes calibrated algorithms with easily identified failure modes[141].

Unlike natural images, medical images vary widely even within the same modality. As discussed in Chapter 2, the visibility of anatomical structures of interest are closely linked to these acquisition parameters, which in turn, can introduce significant variability in disease presentation. For example, thick slices in knee MR imaging can obscure the presence of a small meniscal tear. Or two $T_{1\rho}$ acquisitions with spin lock frequencies at 300Hz and 1kHz are sensitive to different windows of the spectral density function and will have different contrast. Heterogeneous datasets are common in medical imaging. Overall, prospectively acquired research data (Chapters 5,6,7) tends to be standardized compared to real-world clinical data (Chapter 8). This presents several challenges when trying to generalize algorithms from research data into clinical practice. The value of imaging data standardization is well recognized[142]. Several efforts are underway to standardize the quality of clinically acquired imaging and increase robustness of medical imaging algorithms to acquisition shift[143; 144].

Beyond uncertainty introduced by the imaging parameters, uncertainty in image labels is common. Even with expert graders, labels suffer from intra-rater and inter-rater variability.

Despite being time-inefficient, repeated annotations and majority consensus labelling are the standard workaround. Research from Sudre et al[145] found modelling individual rater labels and consensus labels together improved classification algorithm performance. In a multi-rater segmentation task, Jungo et al[146] found training on all raters' annotations led to improved segmentations (higher Dice scores) and better estimates of pixelwise uncertainty.

Overcoming label uncertainty is still an active area of research, but perhaps the most significant challenge facing medical imaging is data scarcity. The acquisition, storage, and labeling of medical data is a complex and expensive endeavor. Furthermore, collecting consent and guaranteeing data anonymization impose additional barriers to creating large, multi-institute datasets. When working with small datasets, researchers rely on transfer learning techniques and aggressive regularization strategies to prevent overfitting.

Transfer learning extends the concept of representation learning for two related tasks[116]. An unknown subset of the features learned from the first task– for which a sizable training dataset exists – is believed to be useful for the second task– which has a limited dataset size. Network weights from the first task are used to initialize weights for the second task, after which the features are frozen or further adjusted during the learning process. ImageNet pretraining is a classic example of transfer learning, where features from a natural image classification task are used to initialize a medical image classifier. Transfer learning is ubiquitous in the medical imaging domain and it is hypothesized to work through improved weight scaling and feature reuse at the first layers of the network[147]. Similarly, experiments using synthetic images for hip cartilage segmentation found that transfer learning did not improve overall accuracy but led to faster convergence and rescued network performance as the training dataset became smaller. It follows that success of transfer learning is highly dependent on network parametrization and the relatedness of the tasks at hand.

Another common regularization approach is data augmentation, also called label-preserving transformations, where the training images are realistically modified by changing positioning,

noise, intensity, and shape such that the network learns to be insensitive to those image variations. Augmentation creates plausible image-label pairs, providing more information on the joint distribution of P(data, label) and a better understanding of P(data|label).

4.4 Causality in medical imaging

A recent perspective piece by Castro et al[139] called for the reframing of medical image representation learning through the lens of causal reasoning, arguing that outlining key assumptions about the data generating mechanism can identify sources of bias that could prevent generalizing to real-world clinical data. In other words, understanding if the medical imaging task is causal (predict effect from cause) or anti-causal (predict cause from effect) can shed light on the best strategies to combat data scarcity and dataset shifts. This connects nicely with the concepts of augmentation and semi-supervised learning presented above. In causal medical imaging tasks, such as the contouring of cartilage on knee MR images, changes in image acquisition parameters would change the labels since the labels are directly derived from the image. Therefore, in causal tasks, the P(data) contains no additional information about P(label|data) and semi-supervised learning is not likely to improve performance over supervised learning. For anti-causal medical imaging tasks, such as classifying Kellgren-Lawrence grade from MR images, the label is not directly derived from the input such that input image P(data) contains information about P(label|data) meeting the condition for semi-supervised learning. Data augmentation can provide value to both causal and anti-causal tasks[139].

4.5 DioscoriDESS: developing a deep learning segmentation framework

Democratization of convolutional neural networks for medical imaging tasks is likely to improve the quality and utility of published algorithms. Manual segmentation of knee cartilage and menisci on high resolution MRI acquisitions is a time-consuming task: an expert observer

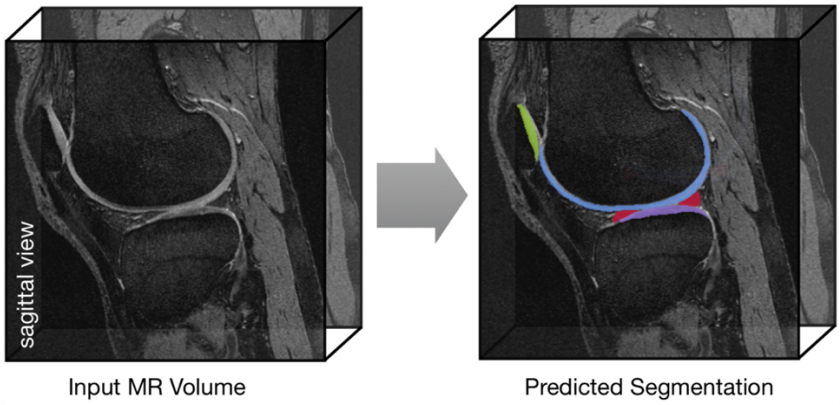


Figure 4.3 Example input MR image volume and predicted segmentation.

may take up to 5 hours per 3D image. If imaging biomarkers– such as cartilage thickness and volume– are to be translated into clinical practice, it follows that the segmentation process must be automated.

4.5.1 Objectives

The road to a deep-learning prototype is long and error prone: for an early graduate student, up to 3 months of engineering effort are dedicated to environment setup, building custom data pipelines, debugging, and developing the network architecture and monitoring dashboards. This iteration cycle is inefficient, produces brittle code, and is prone to the accumulation of technical debt[148]. While excellent deep-learning wrappers exist, they either cater to natural images (Pytorch-Lightning, Fast.ai) or have a steep learning curve (monai.io). DioscoriDESS* was created in 2019 with the help of Francesco Caliva when the Majumdar-Pedoia group participated in the IWOAI cartilage segmentation challenge, challenge results reported in Desai et al[136]. Aiming to maximize organization, reproducibility, and scalability, DiscoriDESS is a bare-bones Tensorflow codebase for training and inference of deep-learning based segmentation algorithms. By reducing time-to-first prototype, researchers have time to experiment with alternate problem formulations and go through additional design cycles. Since 2019, nearly a dozen projects in the Radiology Department have used DioscoriDESS for segmentation, with

several additional projects extending the core template to perform classification, regression, and image reconstruction.

* Pedanius Dioscorides was a Greek Physician who worked in Rome in the 1st Century A.D. and authored 'De Materia Medica', the first seminal pharmacology book in Europe and the Middle East. He recommended the use of ivy as a treatment for musculoskeletal pain. The codebase is an amalgamation of his name and the Dual Echo Steady State (DESS) MR imaging sequence.

A step-by-step walkthrough can be found in the git repository at git.radiology.ucsf.edu/sf048799/dioscoridess a short summary of the key features will be presented here. First, the user follows instructions to activate (or create) a compatible virtual environment on a GPU machine. Then, a use case is selected: training, inference with labels, or inference without labels. Each use case has a main python file (.py) and a template configuration yaml file (.yaml). Only the configuration file requires minor user input and no python programming knowledge is assumed. Once the configuration has been specified via the yaml file, a simple command line call ‘python main.py —cfg config_3D_seg.yaml —desc train_v1’ with the appropriate configuration file and run description, will begin the training and validation loop. Learning is monitored through text logs and Tensorboard dashboards.

4.5.2 Organization/Reproducibility

All training parameters and network architecture configurations are exposed and editable through a yaml configuration file (Figure 4.4), a copy of the exact configuration at runtime is printed at the beginning of the log file, and each run is named using a unique timestamp and user-input description. The timestamp prevents conflicts between runs with identical parameters and enables chronological ordering of experiments. These unique names are used for all logs, checkpoints, predictions. A system-wide random seed is set through the yaml file for experimental reproducibility.

train_template_2D.yaml

```yaml
common:
  seed: 8994
  queue: False
  vis_GPU: '3'
  log_path: 'logs/my_seg_project/'
  save_path: 'ckpts/my_seg_project/'
  print_freq: 15

data_train:
  data_root: 'splits/my_seg_project/train_split1.pickle'
  batch_size: 10
  crop : !!python/tuple [0, 0, 0, 0]
  im_dims: [344,344]
  num_classes: 5
  idx_classes : [0, 1, 2, 3, 4]
  num_channels: 1

data_val:
  data_root : 'splits/my_seg_project/val_split1.pickle'
  batch_size: 10
  crop : !!python/tuple [0, 0, 0, 0]
  im_dims: [344,344]
  num_classes: 5
  idx_classes : [0, 1, 2, 3, 4]
  num_channels: 1

learn:
  num_classes: 5
  comp: ['fem', 'tib','pat', 'men', 'background']
  key_slice: !!python/tuple [0, 0, 0, 0, 0]
  dataloader: 'data_loader'
  optimizer: 'adam'
  lr : 0.0001
  loss: 'wCE_wdice_softmax'
  weights: [0.7,0.75,2,2,0.005]
  metrics: 'spatial_dice'
  max_steps: 50000
  val_freq : 350
  patience: 15
  keep_prob : 0.95

model: 'VNet2D'
model_params:
  num_classes: 5
  num_channels : 16
  num_levels : 4
  num_convolutions : !!python/tuple [1,2,3,3]
  bottom_convolutions : 3

pretrain:
  flag: False
  ckpt: ''
```

infer_template_2D.yaml

```yaml
common:
  seed: 8994
  queue: False
  vis_GPU: '3'
  log_path: 'logs/my_seg_project/'
  pred_path: 'pred/my_seg_project/'

data_infer:
  data_root: 'splits/my_seg_project/test_split1.csv'
  batch_size: 10
  crop : !!python/tuple [0, 0, 0, 0]
  im_dims: [344,344]
  num_classes: 5
  idx_classes : [0, 1, 2, 3, 4]
  num_channels: 1

learn:
  num_classes: 5
  dataloader: 'data_loader'
  loss: 'wCE_dice_softmax'
  metrics: 'spatial_dice'
  save_pred: False

model: 'VNet2D'
model_params:
  num_classes: 5
  num_channels : 16
  num_levels : 4
  num_convolutions : !!python/tuple [1,2,3,3]
  bottom_convolutions : 3

trained_model:
  ckpt: 'ckpts/my_seg_project/model.ckpt-1000'
```

Figure 4.4 Template yaml files for training and inference of a deep learning segmentation algorithm. Detailed documentation is found in the user guide. In the simplest of cases the user would modify paths (log_path, save_path, and data root), number of compartments to be segmented (num_classes), and image dimensions (im_dims, num_channels).

4.5.3 Flexibility

Flexibility: Off the shelf, DioscoriDESS includes 2D and 3D V-Net[149] architectures and accepts single channel or multi-channel inputs. Customizable V-Net parameters are exposed in the yaml and include number of initial filters, network depth, and convolutions per level. Moreover, several softmax and sigmoid based loss functions are available including Weighted Cross Entropy, Dice, Dice with Cross Entropy, and Focal Loss, with optional per class

weighting. A variety of filetypes are supported for images and labels (.mat, .raw, .nii, .h5) as well as pathlists (.txt, .csv, .xls, .pkl). Finally, several regularization options such as early stopping[150] and dropout[126] are included. The breadth and depth of the options provided are sufficient for most segmentation tasks and require no additional programming effort. Notwithstanding, the underlying functions are modular and can easily be extended to support custom dataloaders, losses, or network architectures.

4.5.4 Scalability

Jupyter Notebooks are a ubiquitous format for deep learning tutorials. Interactive runs allow for plotting of intermediate outputs; however, experiment organization and parallelization are almost impossible using Jupyter Notebooks, not to mention code versioning or testing. By monitoring experiments through Tensorboard dashboard and log files, DioscoriDESS is

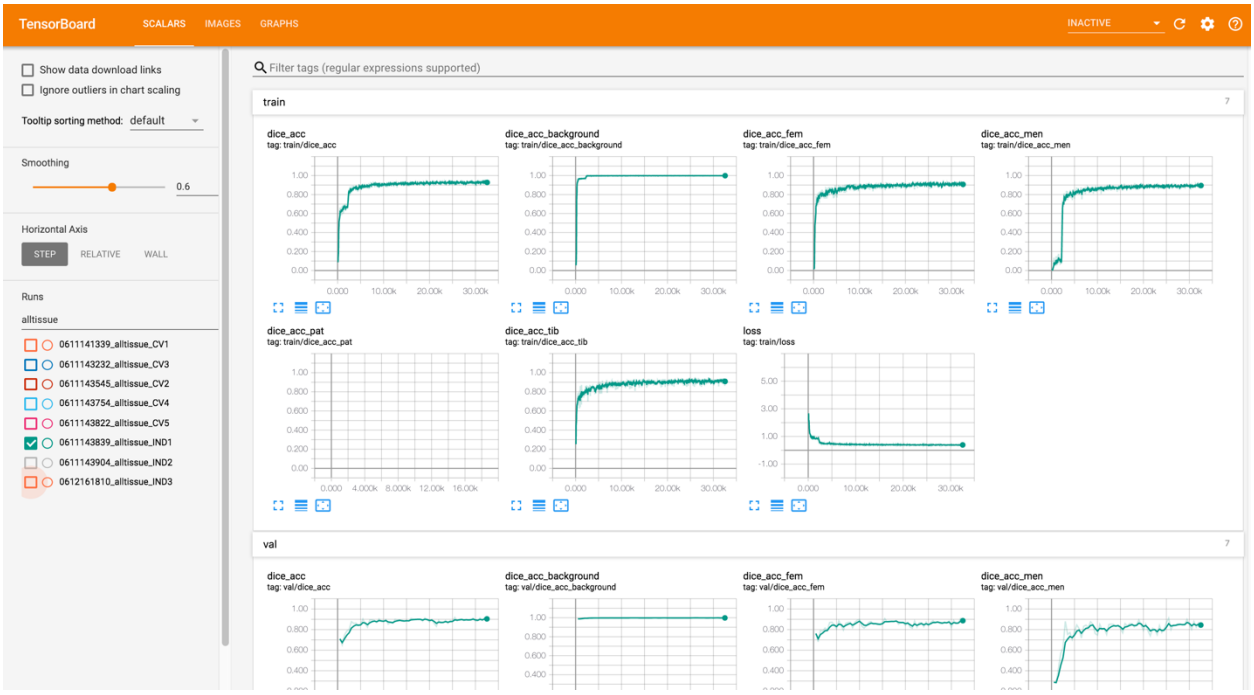


Figure 4.5 Real time monitoring of training through Tensorboard dashboard. Scores for each tissue compartment are plotted separately. Each experimental run is named with a unique timestamp and description, runs names are searchable and can be filtered ("alltissue" is used to filter the runs in the example). Validation curves are sparser than training curves as training data is logged to Tensorboard with every gradient update, while validation occurs once an epoch at most.

compatible with standard queuing systems and can accommodate large scale, computationally expensive tasks such as hyperparameter tuning. In addition, it is space efficient, as data is read in through pathlists rather than relying on pre-specified folder structures or data stored in a common directory. This is ideal for cross-validation setups or for multi-sourced datasets that may be spread across several different folders as it avoids duplication and storage of images.

4.5.5 Interpretability

There are two built-in tools for interpretability: Tensorboard image visualization and uncertainty quantification using inference with Monte Carlo (MC) dropout. Single slice or multi-slice images and segmentation outputs are visualized as RGB images through Tensorflow. Per class, ground truth is encoded in the red channel and non-binarized predictions are encoded in the green channel. Both channels are merged with a fully saturated blue channel and an alpha channel which is the union of both non-zero pixels in the red and green groups. Additive color theory results in 4 distinct pixel values: transparent for true negatives (TN), white for true positives (TP), cyan for false negatives (FN), and magenta for false positives (FP). Visualizations and Dice scores are updated real-time in a Tensorboard dashboard, allowing for qualitative and quantitative monitoring of network performance.

Uncertainty maps are another interpretability method available. The MC dropout scheme is implemented in Tensorflow based on the work proposed by Gal et al[151] and translated to medical imaging by Roy et al[152]. During training, dropout layers act as regularization to prevent overfitting[153]. By keeping dropout enabled during inference and repeatedly running stochastic forward passes through the network, variability in the output predictions can provide pixel-wise estimates of network uncertainty. This process is called Monte Carlo Dropout, as it places a probability distribution over the network weights such that the samples are approximating the posterior distribution. Quantifying uncertainty is crucial for running inference when no labels exist, as high uncertainty measurements can flag out of distribution samples.

To summarize, DioscoriDESS was developed to reduce technical debt by establishing a sustainable workflow and a scalable design pattern for deep learning experimentation in the medical imaging domain.

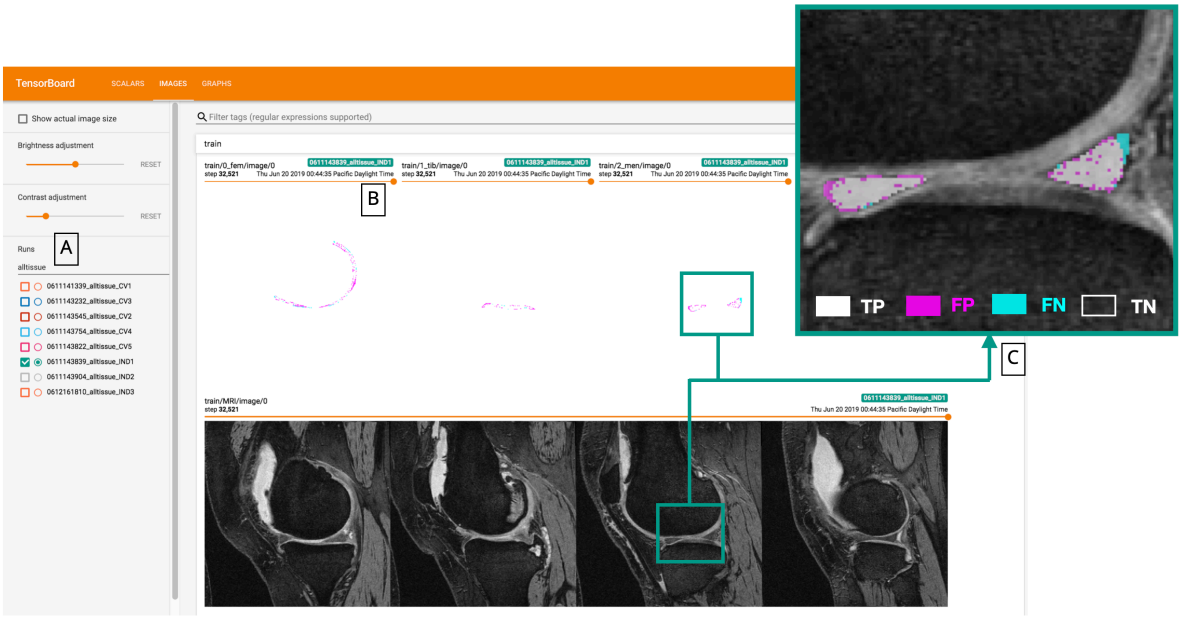


Figure 4.6 Real time qualitative monitoring of segmentation performance through Tensorboard. Segmentation outputs for each compartment are plotted above the corresponding image slice, for volumetric segmentation key slices are specified in the yaml file. A) Runs are uniquely timestamped and sequentially ordered. B) Horizontal scroll bar in Tensorboard allows user to scroll through time. C) True positives (TP), false negatives (FP), false negatives (FN), and true negative (TN) are visualized for each class.

5 Towards Understanding Mechanistic Subgroups of Osteoarthritis: 8 Year Cartilage Thickness Trajectory Analysis

The following manuscript is reformatted and reproduced with full permission from the publisher. It appeared as:

Iriondo, C, Liu, F, Calivà, F, Kamat, S, Majumdar, S, Pedoia, V. Towards understanding mechanistic subgroups of osteoarthritis: 8-year cartilage thickness trajectory analysis. J Orthop Res. 2020; 1– 13. https://doi.org/10.1002/jor.24849

5.1 Abstract

Many studies have validated cartilage thickness as a biomarker for knee osteoarthritis (OA), however, few studies investigate beyond cross-sectional observations or comparisons across two timepoints. By characterizing the trajectory of cartilage thickness changes over 8 years in healthy individuals from the Osteoarthritis Initiative Dataset, this study discovers associations between the dynamics of cartilage changes and OA incidence. A fully automated cartilage segmentation and thickness measurement method was developed and validated against manual measurements: mean absolute error 0.11-0.14mm (n=4129 knees) and

automatic reproducibility 0.04-0.07mm (n=316 knees). Mean thickness for the medial and lateral tibia (MT, LT), central weight-bearing medial and lateral femur (cMF, cLF), and patella (P) cartilage compartments were quantified for 1453 knees at 7 timepoints. Trajectory subgroups were defined per cartilage compartment as: stable, thinning to thickening, accelerated thickening, plateaued thickening, thickening to thinning, accelerated thinning, or plateaued thinning. For tibiofemoral compartments, the stable (22-36%) and plateaued thinning (22-37%) trajectories were the most common, with average initial velocity [μm/month], acceleration [$\mu m/month^2$] for the plateaued thinning trajectories LT -2.66, 0.0326; MT -2.49, 0.0365; cMF -3.51, 0.0509; cLF -2.68, 0.041. In the patella compartment, the plateaued thinning (35%) and thickening to thinning (24%) trajectories were the most common, average initial velocity, acceleration for each -4.17, 0.0424; 1.95, -0.0835. Knees with non-stable trajectories had higher adjusted odds of OA incidence than stable trajectories: accelerated thickening, accelerated thinning, and plateaued thinning trajectories of the MT had adjusted OR of 18.9, 5.48, and 1.47 respectively; in the cMF, adjusted OR of 8.55, 10.1, and 2.61.

5.2 Introduction

Osteoarthritis is a debilitating whole joint disease involving biochemical and structural changes in articular cartilage, bones, ligaments, and muscles. Inpatient US hospital data from 2013 listed osteoarthritis (OA) as the first and second most expensive condition billed to private insurance and Medicare respectively, resulting in over $15 billion USD in direct healthcare costs[154]. The number of individuals suffering from OA and associated costs will increase as the population ages[155]; further studies on the etiology and pathogenesis of the disease are needed to develop effective treatments.

Articular cartilage is a dynamic, responsive tissue that distributes load and decreases friction in joints. Networks of type II collagen, hyaluronan, and aggrecan imbibe water providing cartilage its structure and compressive strength. Chondrocytes in these networks respond to

local mechanical and biochemical cues by actively balancing factors involved in tissue production and breakdown. Animal studies have shown that the disruption of this balance by metabolic, mechanical, or inflammatory tissue stress, and subsequent change of cartilage structure are hallmarks of early OA[47; 156; 157]. Clinical studies using magnetic resonance imaging (MRI) have confirmed the concurrent and predictive validity of cartilage thickness and volume change as biomarkers for knee OA in humans[158; 159], yet quantitative findings vary widely between studies.

Studies have found both cartilage thinning and cartilage thickening at different rates and anatomical locations in knee MRIs of healthy and diseased patients. Buck et al[160] reported annualized rates of cartilage thickness change in 77 healthy knees of 0.28% (std 1.98) in cLF, -0.37% (std 1.31) in LT, -0.17% (std 1.84) in cMF, and -0.32% (std 1.27) in MT; and Wijayaratne et al[161] observed mean volume change of -1.6% (CI 1.2, 1.9) in 148 healthy patellas. In OA knees, rates of volume change observed are as high as -8.6% (std 12.6, n=107) in cMF, -4.6% (std 8.6, n=107) in MT[162], and -4.5% (4.3, n=110) in P[163]. In a cohort of subjects followed for 5 years after ACL injury, Eckstein et al[164] observed annualized thickening of 4μm (CI -1, 9) in cLF, 14μm (CI 10, 19) in cMF, 3μm (CI -1, 7) LT, and 10μm MT (CI 6, 13), with greater magnitude of change occurring in the early interval after injury. Few studies describe the distribution of values observed, and it has been noted that thickening and thinning often occur within the same dataset[165], but averaging across all study subjects might obscure these differences. The ability to detect these changes is closely tied to study design and choice of measurement tool.

OA etiology and pathogenesis are heterogeneous. Studies analyzing cartilage changes in small cohorts (n<200) may not have the power to describe rarer phenotypes of the disease. Large cohorts (n>1000) are preferred for sampling the full spectrum of cartilage health, but manual methods of quantifying cartilage thickness and volume are time-consuming. Furthermore, most studies perform two-timepoint comparisons where the ability to detect significant changes beyond measurement error relies on and improves with longer timescales (2

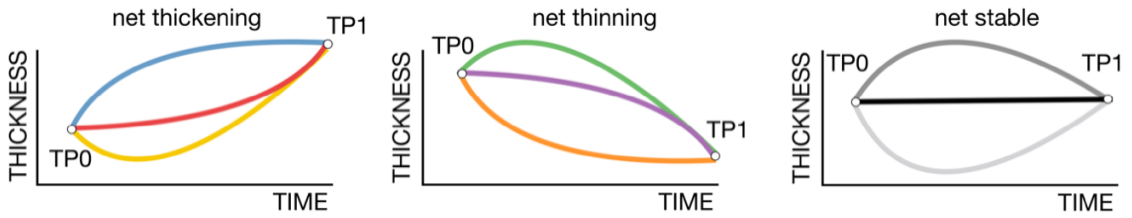


Figure 5.1 Schematic examples of net thickening, net thinning, and net stable trajectories

years, 5 years, etc). However, as timescales become longer, net cartilage thickness changes do not always reflect true cartilage dynamics. For example, net thinning, net stable, and net thickening trajectories illustrated in Figure 5.1 are equivalent under two timepoint comparisons. To identify different trajectories associated with these changes, several longitudinal observations are needed.

Early attempts to increase measurement precision and efficiency used semi-automatic 2D/3D active contours to segment the femoral and tibial cartilage compartments [166-168]. Cartilage volume is estimated directly from segmentations while surface to surface measurements (taken in normal[167] or cylindrical 3D[166] space) are used to estimate cartilage thickness. Recent MRI research has focused on the development of fully automatic algorithms for cartilage segmentation, using voxelwise classification (with 3D statistical shape models[169]), 3D active appearance models[170], 2D/3D convolutional neural networks[171; 172], and combinations thereof [173; 174]. Published methods for segmentation show strong accuracy and reproducibility, but thickness measurements are not validated with external data nor translated into large cohorts to characterize the dynamics of cartilage changes.

In brief, the dynamics of cartilage changes are sensitive to the timescale in which changes are measured, the measurement tool, cohort size, and number of longitudinal observations. In this study, a fully automatic cartilage thickness measurement algorithm and trajectory analysis approach was developed, validated, and applied to 1453 healthy knees deemed at risk of developing OA, aiming to:

(1) Characterize the 8-year cartilage thickness trajectories of five cartilage compartments (lateral and medial tibia, lateral and medial weight-bearing femur, and patella)

(2) Assess if specific subgroups of thickness trajectories have higher odds of incident radiographic OA

(3) Of cartilage thickness trajectories with no net thickness change over 8 years, identify proportion with non-stable thickness trajectories (i.e. differentiate pseudo-stability from true longitudinal stability

5.3 Methods

The Osteoarthritis Initiative is a prospective, longitudinal observational study following 4,796 subjects with radiographic OA or considered at-risk of developing radiographic OA. Among the information collected, X-ray imaging is acquired at most visits to assess radiographic OA status using Kellgren-Lawrence grading. MR imaging is acquired at a subset of visits: structural imaging (DESS) is acquired for both knees while quantitative imaging (T_2 mapping) is acquired for a single knee. An automatic segmentation and thickness measurement algorithm was developed to analyze structural images for subjects that did not have radiographic OA at the baseline visit.

5.3.1 Segmentation Algorithm Development and Validation

A fully automatic method was developed for reliable cartilage segmentation and thickness measurement of knee MRI volumes. A set of 6 convolutional neural networks were trained to segment femoral, tibial, and patellar cartilage and menisci (4 classes) on 176 Double Echo Steady State (DESS) knee MR volumes with manual segmentations of the femur, tibia, patella, and meniscus provided by iMorphics. Twenty-eight volumes were held out from training as a test set from which to report performance. A representative example of image augmentation in Figure 5.2 and details for each of the 6 trained networks and training procedure in Table 5.1.

Image volumes and ground truth segmentation masks were resampled isotropically using cubic interpolation before undergoing geometric transforms. Transforms included 3D rotation, 3D affine deformation, 3D Bspline deformations, and combinations thereof. 3D rotation was performed about the x-axis, z-axis, or both, rotating +-6 to 17. 3D affine transforms rescaled the x,y,and z axes separately with a factor between 1 and 1.25. 3D B-spline deformations were performed by dividing the 2D image matrix into a coarse grid (1/4th of image matrix), creating a field of random deformations. The deformations are then smoothed with a gaussian filter with sigma 45 to 60 and applied to the native images and segmentations. Linear interpolation was performed on all segmentations, images and segmentations (after thresholding at 0.5) were resampled to native, anisotropic dimensions. Intensity based augmentations were used to mimic signal inhomogeneity without modifying the segmentation.

Data was divided participant-wise into 3 independent splits (1/3rd training, 2/3rd validation). Independent training splits allowed for the development of diverse models to better correct for systematic errors[175] and a large validation set provided accurate estimates of model generalizability for monitoring of network training and early stopping. Table 5.2 describes participant demographics for the OAI Dataset and all subsets used in this study. All models were implemented in Tensorflow 1.10 and trained with a weighted Dice loss[149] using Adam on a single Nvidia TitanX (12196MiB) or Nvidia V100 (32480MiB) GPU. Per-class probability predictions from all models were aggregated into the final model ensemble following[26].

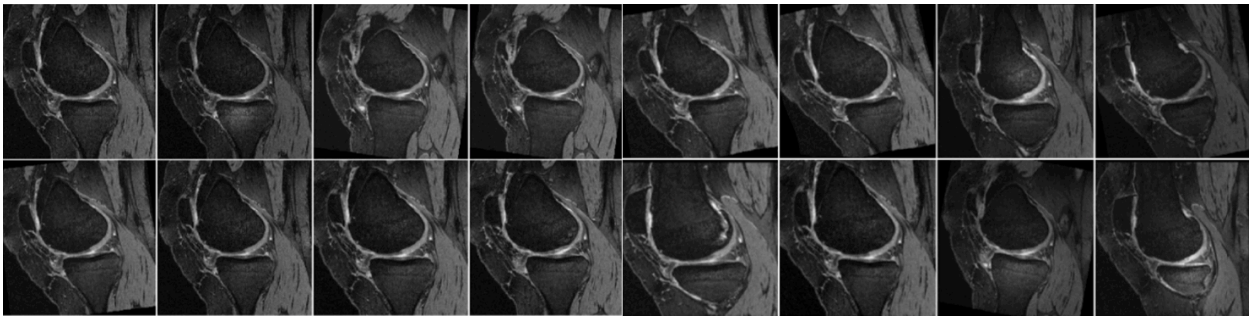


Figure 5.2 Single slice from augmented volume in training dataset.

Table 5.1 Segmentation network details for each of the 6 trained models in the model ensemble.

		2D network A	2D network B	2D network C	3D network A	3D network B	3D network C
data	**batch_size**	16	16	8	1	1	1
	data_split	IND0	IND1	IND2	IND0	IND1	IND2
	im_dims	(344x344)	(344x344)	(344x344)	(344x344x140)	(344x344x140)	(344x344x140)
	num_channels	1	1	1	1	1	1
	num_classes	5	5	5	5	5	5
learn	**drop_rate**	0.05	0.05	0.05	0.05	0.05	0.05
	loss	weighted Dice	weighted Dice	weighted Dice	weighted Dice	weighted Dice	weighted Dice
	lr	5.00E-05	5.00E-05	1.00E-04	0.0001	0.0001	0.0001
	max_steps	100000	100000	100000	50000	50000	50000
	optimizer	Adam	Adam	Adam	Adam	Adam	Adam
	patience	30	30	30	15	15	15
	val_freq	500	500	500	350	350	350
	comp	fem, tib, pat, men, bckgd	fem, tib, pat, men, bckgd	fem, tib, pat, men, bckgd	fem, tib, pat, men, bckgd	fem, tib, pat, men, bckgd	fem, tib, pat, men, bckgd
	weights	1,1,2,2,0.05	1,1,2,2,0.05	1.5,1.5,3.5,3,0.05	0.7,0.75,2,2,0.01	0.7,0.75,2,2,0.01	0.7,0.75,2,2,0.01
model params	**model**	VNet2D	VNet2D	VNet2D	VNet3D	VNet3D	VNet3D
	bottom_conv	4	1	1	3	3	3
	levels	2	6	6	2	2	2
	conv_per_level	(4,4)	(1,1,1,1,1,1)	(1,1,1,1,1,1)	(1,2)	(1,2)	(1,2)
	filters	16	24	24	26	26	26
pretraining	**flag**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table 5.2 Dataset description for each method step. All datasets used in this research are subsets of the OAI dataset. Continuous and categorical variable description of datasets used throughout the manuscript. Continuous variables listed as mean (95% CI), categorical variables listed at raw count (% total). Descriptors calculated using all available samples (baseline and all timepoints thereafter).

			Continuous Descriptors					
	# patients	*#samples*	*age*	*height_m*	*weight*	*BMI*	*KOOS*	*WOMAC*
OAI dataset	4795	46735	63.4 (63.3, 63.4)	1.68 (1.68, 1.68)	80.8 (80.7, 81.0)	28.5 (28.4, 28.5)	86.6 (86.5, 86.8)	10.2 (10.1, 10.3)
seg_trainval	74	148	59.9 (58.4, 61.5)	1.7 (1.68, 1.71)	89.1 (86.5, 91.8)	30.9 (30.1, 31.6)	68.1 (64.7, 71.5)	25.9 (23.0, 28.8)
seg_test	14	28	71.4 (68.4, 74.3)	1.72 (1.69, 1.75)	91.7 (85.6, 97.8)	30.8 (29.3, 32.4)	77.5 (71.2, 83.8)	19.8 (14.9, 24.7)
thick_val	1273	4129	63.5 (63.2, 63.8)	1.68 (1.68, 1.68)	84.5 (84.0, 85.0)	29.8 (29.7, 30.0)	77.7 (77.2, 78.3)	17.2 (16.7, 17.7)
noOA_traj	1200	11970	61.8 (61.6, 61.9)	1.69 (1.68, 1.69)	78.3 (78.0, 78.6)	27.4 (27.3, 27.4)	91.3 (91.0, 91.5)	5.98 (5.8, 6.16)

			Categorical Descriptors			
	# patients	*#samples*	*side=RIGHT*	*sex=F*	*TKR_ever=1.0*	*OA_ever=1.0*
OAI dataset	4795	46735	22136 (50%)	25700 (57%)	2105 (5%)	22011 (49%)
seg_trainval	74	148	80 (54%)	76 (51%)	28 (19%)	144 (97%)
seg_test	14	28	10 (36%)	10 (36%)	8 (29%)	28 (100%)
thick_val	1273	4129	2028 (49%)	2370 (57%)	859 (21%)	3787 (92%)
noOA_traj	1200	11970	6091 (56%)	6132 (56%)	57 (1%)	1758 (16%)

5.3.2 Sub-Segmentation Algorithm Development

The predicted cartilage compartments were then sub-segmented into lateral tibia (LT), medial tibia (MT), patella (P), central weight-bearing lateral femur (cLF), and central weight-bearing medial femur (cMF) compartments Figure 5.3. Medial-lateral femur compartments were divided by the mid-sagittal slice of the femoral cartilage volume. Weight-bearing regions were defined as 60% of the distance between the trochlear notch and the posterior ends (as described in [176]). The weight-bearing region femoral cartilage compartment was using a rule-based method described in Figure 5.3. The weight-bearing femoral cartilage region was selected in order to minimize partial volume effects and such that the region definition matched that of the manual thickness validation data by Chondrometrics.

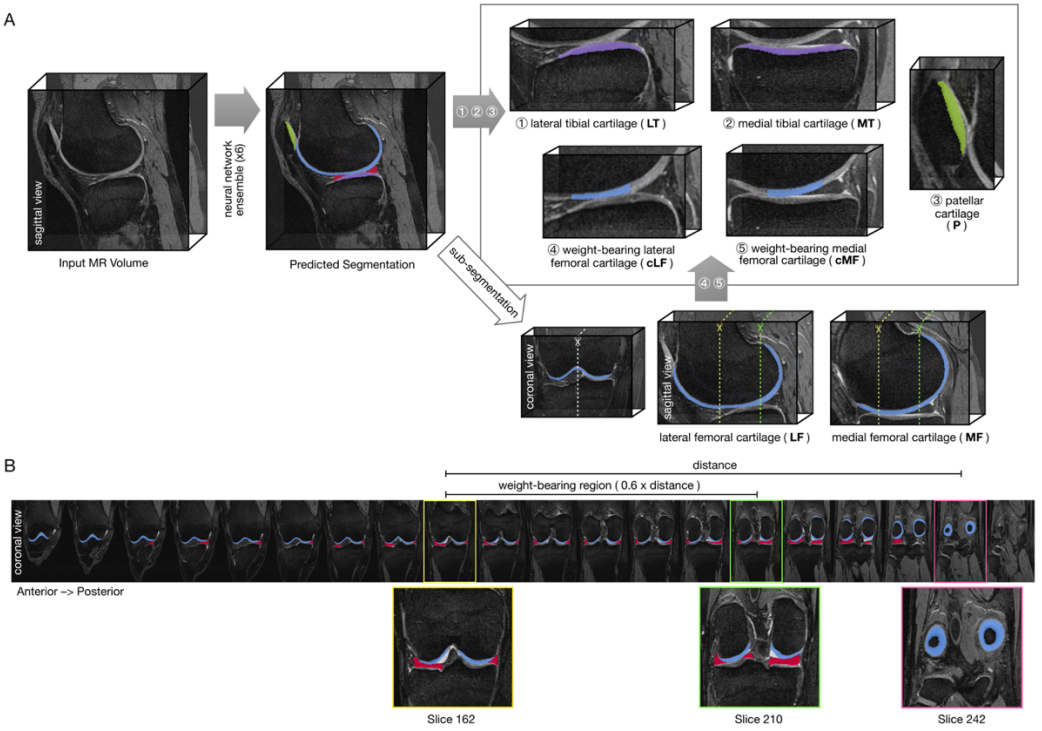


Figure 5.3 (A) Example participant from the Osteoarthritis Initiative dataset. Input MR image is inferred on using the 6 trained neural networks. Probabilities from all the networks are combined to create the predicted segmentation, which is then sub-segmented into 5 compartments for thickness analysis. (B) Identification of weight-bearing region of the femur as defined by the "60% rule" on the example participant. The start of the intercondylar notch is identified as the anterior-most coronal slice with two pieces of femoral cartilage and meniscus present (yellow border) while the "double-bullseye" indicates the last slice of femoral cartilage (pink border). The 60% rule is applied to define the weight-bearing region (yellow border to green border).

5.3.3 Automatic Thickness Measurement and Validation

Per compartment and per sagittal slice, a Euclidean distance transform and skeletonization were performed Figure 5.4. The value of the distance map was sampled at each skeleton point, and all points across all slices were averaged to calculate mean thickness. Lateral and medial femoral compartments underwent Euclidean distance transform and skeletonization before sub-segmentation. Only the weight-bearing region was included in the mean thickness calculation for the lateral and medial femur.

Manual cartilage thickness measurements of the weight-bearing lateral femur, weight-bearing medial femur, lateral tibia, and medial tibia for a subset of the OAI were provided by Chondrometrics. Automatic thickness measurements were tested for accuracy against manual thickness measurements (mean cartilage thickness of the lateral tibia, medial tibia, central lateral femur, and central medial femur) on 4129 knees; manual thickness measurements were not available for the patella. Manual and automatic method repeatability was compared in knees

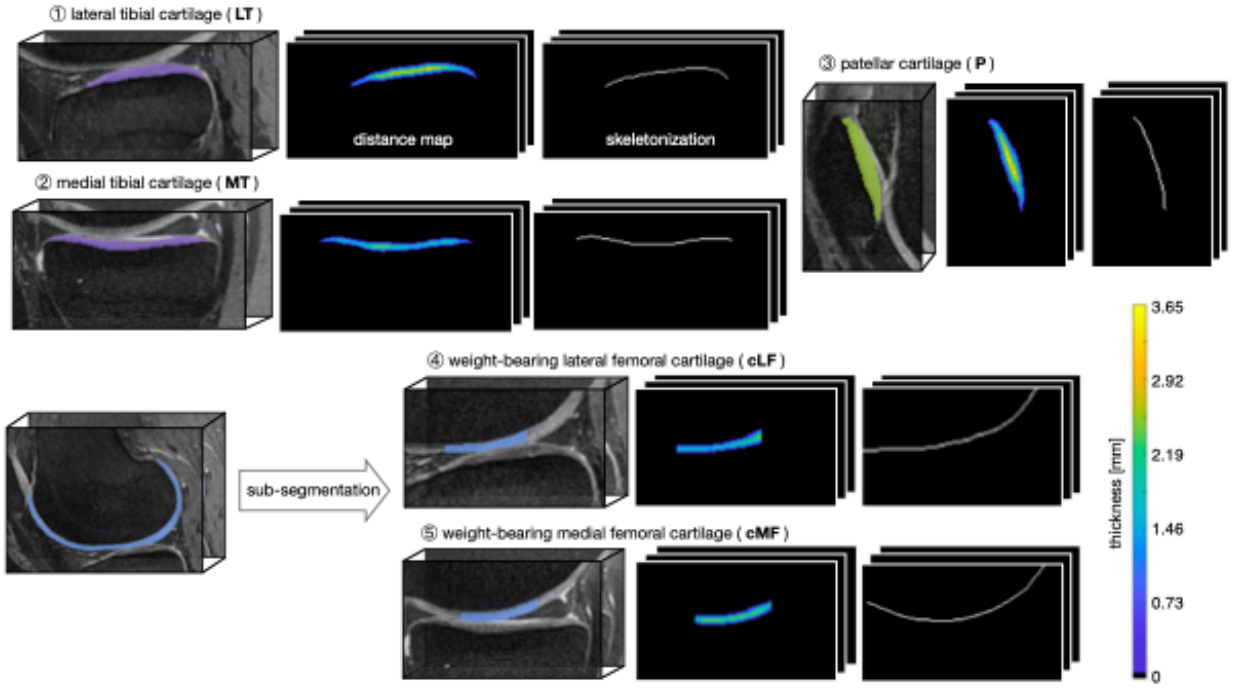
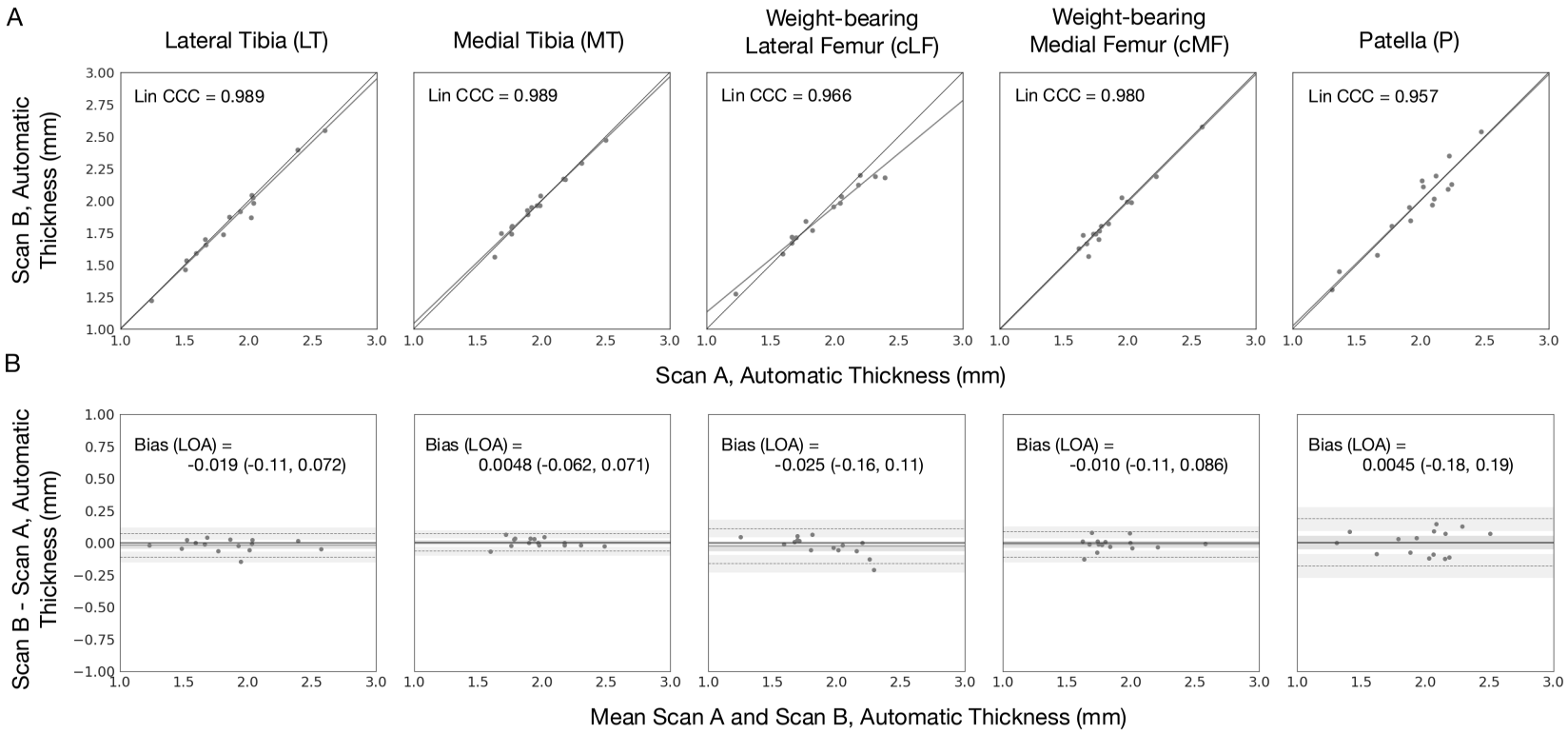


Figure 5.4 Binary masks of each compartment undergo a Euclidean distance transform and skeletonization. Average thickness values are estimated by sampling the distance map at each valid skeleton point and averaging across all points in all slices.

that were without radiographic OA (KLG<2) and had unchanged KLG one year later. Full automatic pipeline scan-rescan results on 15 patients from an OAI pilot study presented in Figure 5.5. A simulated 3D rotation experiment was conducted to investigate the effect of 3D positioning on average cartilage thickness measurements. 19 knees were randomly selected from entirety of the Osteoarthritis Initiative Dataset. These knees were resampled to isotropic (0.3646x0.3646x0.3646mm) resolution, rotated by +- 12˚ about the medio-lateral axis, longitudinal axis, and anterior-posterior axis, then resampled back to their native resolution (0.3646x0.3646x0.7mm). These images were then run through the segmentation, sub-segmentation, and thickness measurement algorithms to investigate the effect of 3D position on average cartilage thickness. Results showing sensitivity of pipeline to 3D patient position in Figure 5.6.

5.3.4 Cartilage Thickness and Trajectory Analysis

Average cartilage thickness per compartment was computed for each timepoint, after which change in cartilage tissue thickness trajectories were analyzed in 1453 without radiographic OA at baseline (KLG<2), and complete MR imaging over 8 years (7 time points, from baseline to 96 months). Incident radiographic OA was defined as KLG >=2 at any follow-up visit. Per compartment, each thickness measurement was treated as an observation in time, and a second-order polynomial fitting was performed to estimate a per-compartment thickness trajectory. The 15th percentile of least accurate polynomial fits were excluded from the analysis. Initial velocity and acceleration values of cartilage thickness change were derived from the first and second order time derivatives of the observed thickness trajectories. These time derivates are independent of initial thickness. Participant subgroups are defined by net cartilage thickness change, velocity, and acceleration per compartment. Each participant-knee is plotted as a single point on an initial velocity vs acceleration graph, to identify different types of cartilage trajectories.



		Mean Absolute Diff (mm)	RMS CV (%)
Scan A vs Scan B	LT	0.036 (0.016, 0.056)	1.88
	MT	0.028 (0.017, 0.039)	1.36
	cLF	0.053 (0.023, 0.082)	2.48
	cMF	0.035 (0.015, 0.056)	2.05
	P	0.085 (0.063, 0.11)	3.27

Figure 5.5 Scan-rescan experiment with pilot OAI dataset (A) Correlation plots between automatic thickness in scan A and scan B with line of best fit (n=15 knees). (B) Bland-Altman plots with bias and limits of agreement per compartment, shaded portions indicate the 95% confidence interval of each. Table at bottom right with summary metrics. CCC = Lin's Concordance Correlation Coefficient, LOA = Limits of Agreement, RMS CV = Root Mean Squared Coefficient of Variation

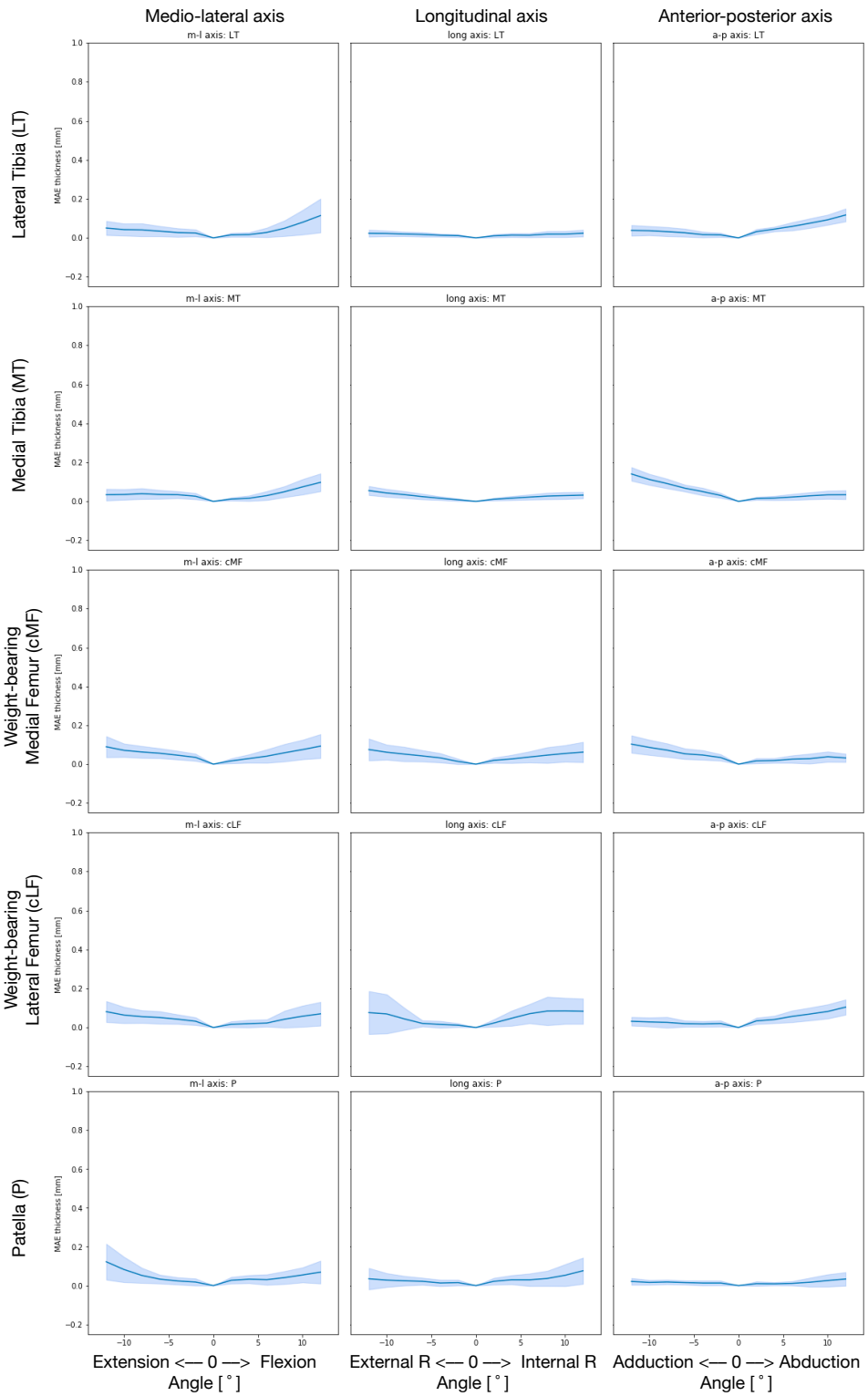


Figure 5.6 The effect of rotation on average cartilage thickness. Mean absolute error (mm) and standard deviation are plotted as a function of rotation angle for each axis, for each tissue. Thickness measurements at angle 0 are considered the ground truth. Errors are the lowest within +-5˚ of the original positioning and increase with rotation angle.

First, the stable subgroup contains knees with cartilage that had no net detectable change in thickness in the 8 year period and a steady thickness trajectory within the limit of reproducibility (±150μm). Then, the remaining knees are grouped according to net thickness change (μm), initial velocity (y-axis, μm/month), and acceleration (x-axis, μm/month^2). Six additional subgroups are defined and color-coded: thinning to thickening (yellow), accelerated thickening (red), plateaued thickening (blue), thickening to thinning (green), accelerated thinning (purple), and plateaued thinning (orange). An example dynamics plot with possible curves is explained in Figure 5.7. Descriptive statistics for knee characteristics at baseline, cartilage dynamics, and odds-ratios for OA incidence during the 8-year period are reported for each participant subgroup. Lastly, knees with no net thickness change between baseline and the 8 year timepoint were divided into longitudinally stable (stable thickness trajectory) and pseudo-stable (non-stable thickness trajectory).

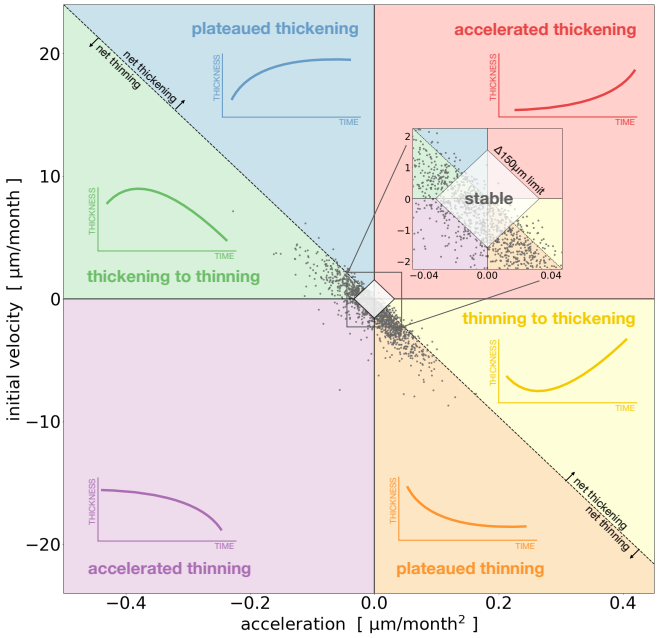


Figure 5.7 Example dynamics plot for a single cartilage compartment. Each marker on the scatterplot represents a single cartilage trajectory over 8 years (n=1453 trajectories). The y-axis is initial velocity of cartilage thickness. The x-axis is acceleration of cartilage thickness. The diagonal-line (dotted) represents the line of zero net change over the 8 years, distance perpendicular to the line is proportional to cartilage thickness change, points further away from the diagonal had greater net thickness changes. Example thickness vs time curves for each subgroup are illustrated.

5.3.5 Statistical Analysis

Segmentation method performance is assessed per compartment using mean and standard deviation Dice overlap scores. Accuracy and repeatability of thickness measurements and scan-rescan are assessed by Lin's concordance correlation coefficient, mean absolute difference, and root mean square coefficient of variation (RMS CV%), reported with 95% confidence intervals. Bland-Altman analysis is performed between manual and automatic thickness measurements per compartment. Mean and standard deviation for cartilage velocity and acceleration are reported per compartment, per subgroup. Two sided t-tests and while $\chi 2$ tests are used to assess baseline differences between the stable subgroup and each other subgroup. Radiographic OA incidence for the examined knee and contralateral knee are compared between the stable subgroup and all others using multivariate logistic regression adjusted for age, sex, and BMI, reported as Odds Ratios with 95% confidence intervals. Race was not included in downstream analyses given small sample size of non-Caucasian participants (91% Caucasian, 7% African-American and 2% other). For each analysis, significance level alpha = 0.05 is adjusted for multiple comparisons using Bonferroni correction resulting in a p-value threshold of 0.001, results that meet Bonferroni threshold are bolded.

5.4 Results

The ensemble model had robust test set segmentation performance with mean (standard deviation) Dice overlap coefficients: femoral 0.890 (0.023), tibial 0.880 (0.036), and patellar cartilage 0.850 (0.068), and meniscus 0.874 (0.024). The ensemble model outperformed individual model predictions, corrected small errors in segmentation without requiring image postprocessing, and accurately segmented denuded bone surfaces Figure 5.8. Ensemble model errors were generally localized near the intercondylar notch and on cartilage edge slices with partial volume effects.

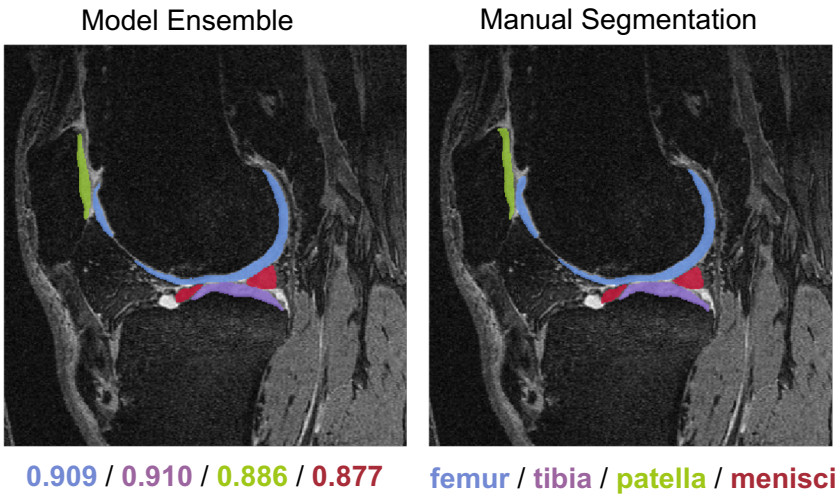


Figure 5.8 Predicted segmentation on a single slice of a test participant for the model ensemble, with the manual segmentation as a reference.

Automatic cartilage thickness measurements showed good agreement with, and similar repeatability to, the independent sample of manual measurements. Manual and automatic measurements had strong concordance correlation coefficients ranging from 0.817 in the weight-bearing lateral femur to 0.929 in the lateral tibia Table 5.3. The upper 95% CI for the largest mean absolute difference was 0.15mm, which is less than half the in-plane pixel resolution. As observed, the automatic method slightly overestimated cartilage thickness in the

Table 5.3 Accuracy and reproducibility results for manual and automatic thickness measurements. Accuracy is assessed by comparing manual and automatic methods (n=4129) using Mean Absolute Difference, and Root Mean Squared Coefficient of Variation reported with 95% confidence intervals in parenthesis. Manual and automatic method repeatability are assessed with the same metrics on consecutively acquired images, approximately 12 months apart, with unchanged KL grading of either 0 or 1 (n=313).

		Mean Absolute Diff (mm)	RMS CV (%)
Manual vs Automatic	LT	0.12 (0.11, 0.12)	6.16
	MT	0.11 (0.11, 0.11)	5.84
	cLF	0.14 (0.14, 0.15)	6.92
	cMF	0.14 (0.13, 0.14)	7.55
Auto Repeatability	LT	0.04 (0.04, 0.04)	1.91
	MT	0.04 (0.04, 0.04)	2.24
	cLF	0.05 (0.04, 0.05)	2.43
	cMF	0.07 (0.06, 0.07)	3.48
Manual Repeatability	LT	0.05 (0.04, 0.05)	2.25
	MT	0.05 (0.04, 0.05)	2.63
	cLF	0.05 (0.04, 0.05)	2.34
	cMF	0.06 (0.05, 0.06)	2.92

tibial compartments (bias 0.06 to 0.08mm) and the weight-bearing medial femur (bias 0.03mm), and underestimated thickness in the weight-bearing lateral femur (-0.07mm) Figure 5.9. Stratified performance by radiographic OA grade (not shown), narrow limits of agreement reveal strong agreement between automatic and manual methods for KLG 0-3 knees (all LOA between -0.41 to 0.39mm) and more moderate agreement in KLG 4 knees (LOA between -0.52 to 0.47mm) particularly in the weight-bearing femoral compartments. Repeatability results on a sample of radiographically unchanged KL0/1 knees reveals comparable performance between manual and automatic thickness measurements, as both methods' RMS CV % was below 3.5% in all tibio-femoral cartilage compartments; results of the automatic method in a 15 patient scan-rescan experiment had RMS CV% below 2.5% in these compartments, and CCC between 0.957 and 0.989 for all compartments. The full pipeline was robust to small changes in 3D positioning.

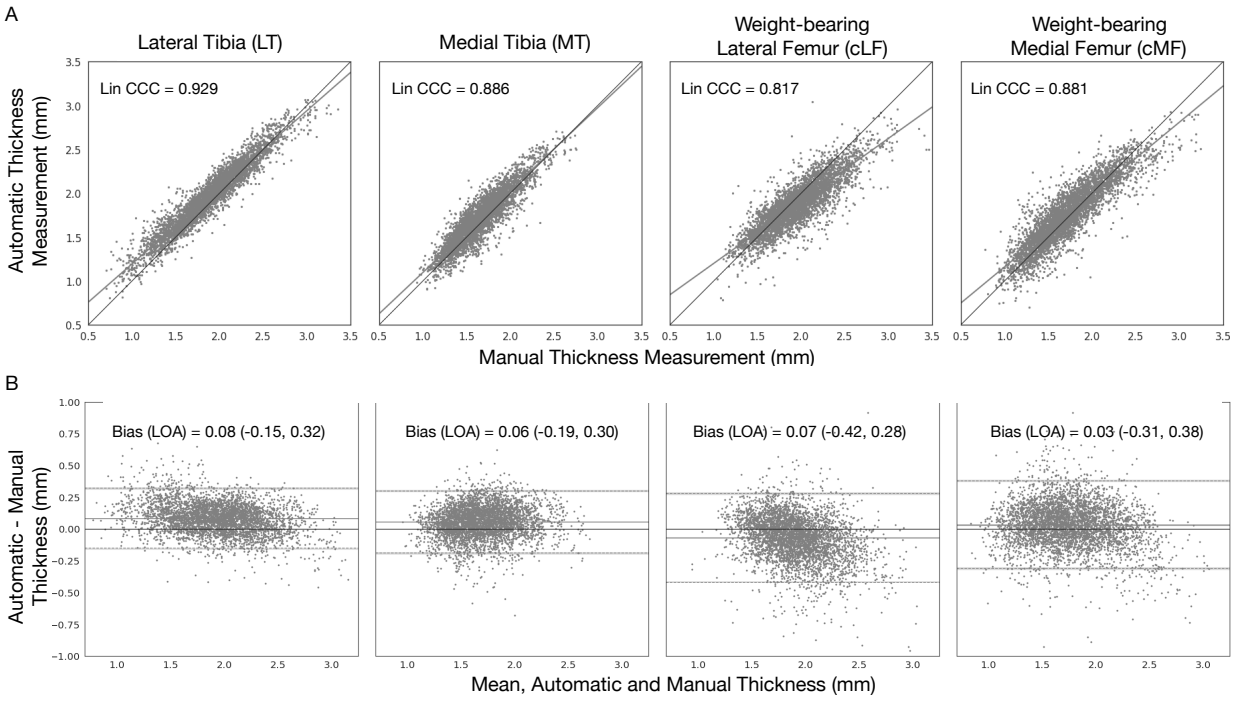


Figure 5.9 (A) Correlation plots between manual and automatic thickness measurements per compartment with line of best fit (n=4129 knees). (B) Bland-Altman plots with bias and limits of agreement per compartment, shaded portions indicate the 95% confidence interval of each. Manual thickness measurements were not available for patellar cartilage. CCC = Lin's Concordance Correlation Coefficient, LOA = Limits of Agreement

Dynamics scatterplots for cartilage trajectories are visualized in Figure 5.10, average values for net cartilage thickness change over 8 years, initial velocity, and acceleration per subgroup are also reported. Thickness changes over time were not constant for any of the non-stable subgroups. Observing the points on the dynamics scatterplot, tibial cartilage trajectories had less variability (higher point density) than femoral and patellar trajectories. Points representing MT trajectories were centered on the diagonal (line of zero net change over 8 years) while a majority 66% of LT trajectories fell below the diagonal (net thinning). Femoral trajectories spanned a wider range of initial velocity and acceleration values, while patellar trajectories had the most variability overall (sparse point density).

62-73% of trajectories that had no detectable net thickness change at the 8 year visit compared to baseline were pseudo-stable (non-stable thickness trajectories). The plateaued thinning and stable subgroups were the two largest subgroups for tibiofemoral compartments (LT, MT, cMF, cLF), together accounting for 49-67% of all knees. Thickening to thinning and thinning to thickening subgroups were the next most frequent, together occurring in 23-38% of knees. Accelerated thinning was less common 1.5-8.9%, and accelerated thickening was the rarest 0-1.7%, full statistics reported are Table 5.4. These relative proportions were not consistent in the patellar compartment (P) where plateaued thinning, thickening to thinning, and accelerated thinning subgroups accounted for 35%, 24%, and 24% of the knees respectively. Only 13% of knees fell within the stable subgroup, 2.5% plateaued thickening, and 0% accelerated thickening. Significant differences in baseline characteristics between the stable tibiofemoral compartment subgroups and other tibiofemoral subgroups were observed. The plateaued thinning subgroup had significantly higher baseline thickness (LT:2.1mm vs 2.02mm, MT:1.79mm vs 1.69mm, cMF:1.95mm vs 1.83mm, $p<0.001$) and a lower proportion of female participants than the stable subgroup (LT:52% vs 59.5%, MT:52% vs 66%, cMF:47.6% vs 59.2%, cLF:45.7% vs 56.3%, $p<0.05$). A similar trend was observed in the accelerated thinning

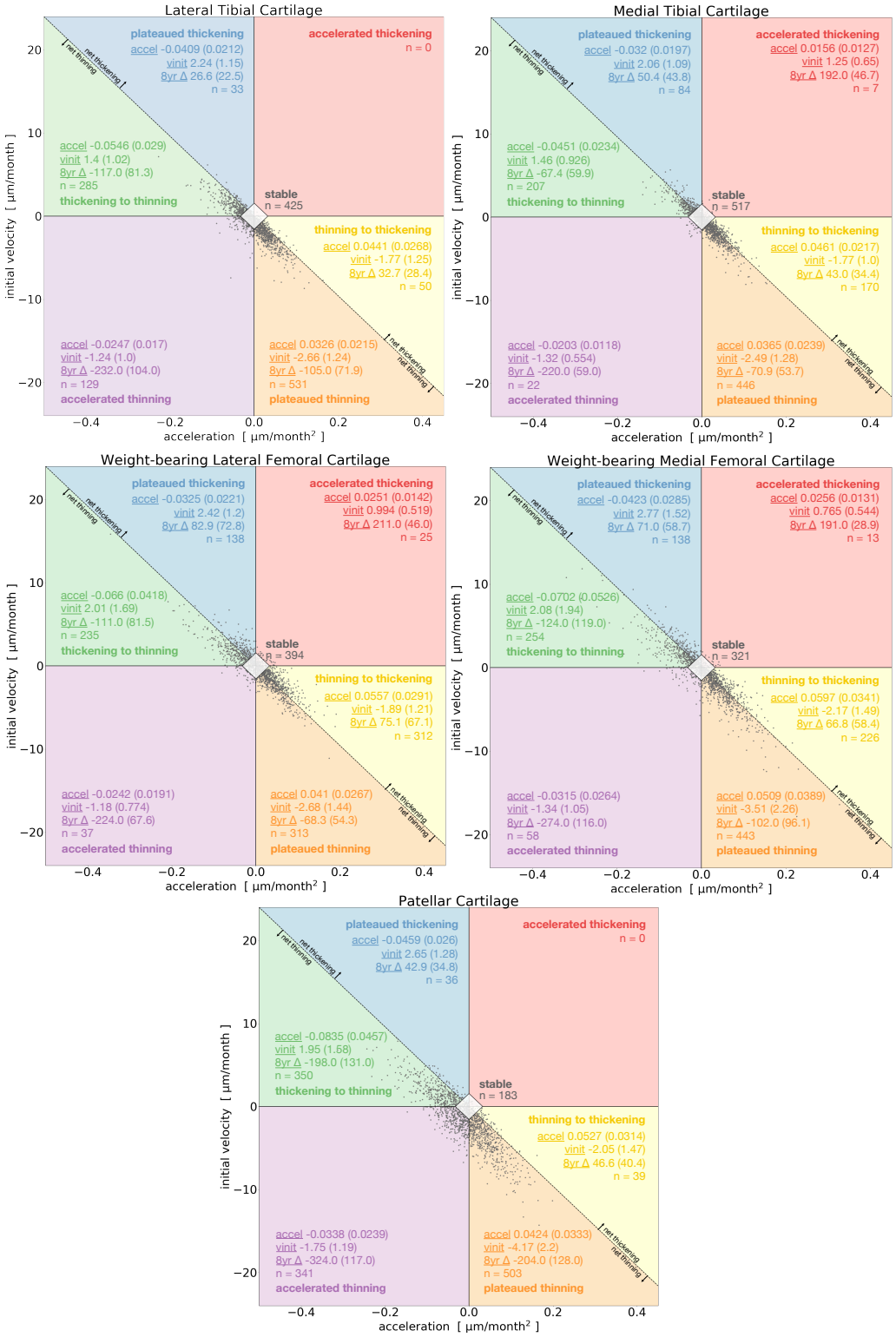


Figure 5.10 8-year cartilage dynamics plot for LT, MT, cMF, cLF, and P compartments. Each marker represents a cartilage trajectory over 8 years. Mean (standard deviation) values are reported per subgroup for net 8-year change (8yr Δ [µm]), initial velocity (vinit [µm/month]), and acceleration (accel [µm/month^2]).

Table 5.4 Characteristics of cartilage dynamics subgroups, where * p<0.05 and ** p<0.001 (**bold**) for comparisons between the stable subgroup and all other subgroups (accelerated thickening, plateaued thickening, thickening to thinning, accelerated thinning, plateaued thinning, and thinning to thickening). All knees included in analysis had no radiographic OA at baseline. Count describes the total number of knees belonging to the subgroup (% of total). Per cartilage compartment, stable subgroup contained thickness measurements within the limits of reproducibility (±150μm) at all timepoints over an 8 year period, other subgroups defined by net thickness change, initial velocity, and acceleration. Continuous variables at baseline– thickness, age, BMI, and WOMAC –shown as mean (standard deviation). Categorical variables at baseline– sex, contralateral OA– shown as raw count (% of total with available information).

			net thickening			net thinning		
		stable	thinning to thickening	accelerated thickening	plateaued thickening	thickening to thinning	accelerated thinning	plateaued thinning
LT	count	425 (29%)	50 (3.4%)	0	33 (2.3%)	285 (20%)	129 (8.9%)	531 (37%)
	baseline thickness [mm]	2.02 (0.285)	2.06 (0.257)		2.04 (0.283)	2.04 (0.323)	2.1 (0.294)*	**2.1 (0.309)****
	age [years]	59.2 (8.34)	55.8 (7.68)*		**53.7 (5.96)****	58.6 (8.43)	60.9 (9.02)	58.8 (8.68)
	BMI [kg/m^2]	27.0 (4.47)	27.6 (4.48)		27.0 (4.27)	27.4 (4.25)	27.5 (4.18)	27.2 (4.18)
	baseline WOMAC	6.16 (9.55)	7.75 (7.65)		5.77 (7.78)	6.63 (10.0)	5.33 (7.74)	6.15 (9.19)
	sex = Female	253 (59.5%)	32 (64.0%)		29 (87.9%)*	178 (62.5%)	60 (46.5%)*	276 (52.0%)*
	contralateral OA = Yes	62 (19.2%)	10 (29.4%)		5 (17.9%)	44 (21.0%)	22 (22.4%)	65 (17.4%)
MT	count	517 (36%)	170 (12%)	7 (0.5%)	84 (5.8%)	207 (14%)	22 (1.5%)	446 (31%)
	baseline thickness [mm]	1.69 (0.237)	1.72 (0.232)	1.62 (0.176)	1.66 (0.237)	**1.76 (0.241)****	**1.86 (0.295)****	**1.79 (0.247)****
	age [years]	58.1 (8.12)	59.3 (8.54)	57.1 (8.65)	58.2 (8.97)	58.7 (8.72)	61.6 (9.54)*	59.5 (8.55)*
	BMI [kg/m^2]	27.2 (4.4)	26.9 (4.27)	26.4 (3.61)	27.7 (4.3)	27.9 (4.52)	29.8 (5.03)*	26.8 (4.12)
	baseline WOMAC	6.25 (9.36)	7.0 (10.2)	10.6 (10.8)	7.92 (11.1)	5.2 (7.97)	6.96 (8.69)	6.2 (9.03)
	sex = Female	341 (66.0%)	106 (62.4%)	6 (85.7%)	61 (72.6%)	111 (53.6%)*	9 (40.9%)*	**232 (52.0%)****
	contralateral OA = Yes	63 (16.3%)	19 (15.1%)	2 (33.3%)	16 (25.8%)	30 (20.4%)	5 (41.7%)	61 (18.5%)
cMF	count	321 (22%)	226 (16%)	13 (0.9%)	138 (9.5%)	254 (18%)	58 (4.0%)	443 (31%)
	baseline thickness [mm]	1.83 (0.276)	1.81 (0.281)	1.7 (0.219)	1.84 (0.263)	1.88 (0.288)	1.91 (0.321)*	**1.95 (0.289)****
	age [years]	58.0 (8.19)	58.0 (8.11)	54.6 (7.14)	57.1 (8.73)	**60.4 (8.67)****	60.8 (9.2)*	58.7 (8.4)
	BMI [kg/m^2]	27.1 (4.3)	26.6 (4.17)	26.7 (3.95)	27.7 (4.72)	27.4 (4.43)	28.0 (4.43)	27.1 (4.32)
	baseline WOMAC	5.57 (9.11)	6.2 (9.21)	7.78 (14.3)	6.73 (10.1)	6.77 (9.55)	7.37 (11.8)	6.17 (8.88)
	sex = Female	190 (59.2%)	162 (71.7%)*	9 (69.2%)	91 (65.9%)	155 (61.0%)	31 (53.4%)	211 (47.6%)*
	contralateral OA = Yes	33 (12.7%)	33 (19.6%)	0 (0.0%)	14 (13.7%)	42 (23.9%)*	11 (31.4%)*	56 (17.7%)
cLF	count	394 (27%)	312 (22%)	25 (1.7%)	138 (9.5%)	235 (16%)	37 (2.5%)	313 (22%)
	baseline thickness [mm]	1.82 (0.248)	1.79 (0.232)	1.7 (0.161)*	1.77 (0.237)*	1.85 (0.272)	1.92 (0.285)*	1.88 (0.239)*
	age [years]	58.2 (8.32)	58.4 (8.07)	55.3 (7.55)	58.2 (8.39)	**60.9 (9.05)****	60.3 (8.77)	58.3 (8.45)
	BMI [kg/m^2]	27.3 (4.26)	26.8 (4.05)	28.5 (5.21)	27.8 (4.58)	27.1 (4.63)	28.0 (4.06)	26.9 (4.13)
	baseline WOMAC	6.32 (9.5)	5.36 (8.26)	7.92 (9.13)	7.19 (9.75)	6.31 (9.33)	5.19 (6.26)	5.73 (8.59)
	sex = Female	222 (56.3%)	209 (67.0%)*	**24 (96.0%)****	94 (68.1%)*	117 (49.8%)	11 (29.7%)*	143 (45.7%)*
	contralateral OA = Yes	52 (16.9%)	37 (16.4%)	9 (42.9%)*	10 (9.26%)	40 (22.1%)	4 (18.2%)	47 (21.4%)
P	count	183 (13%)	39 (2.7%)	0	36 (2.5%)	350 (24%)	341 (24%)	503 (35%)
	baseline thickness [mm]	2.33 (0.316)	2.22 (0.362)*		2.38 (0.339)	2.24 (0.35)*	**2.23 (0.34)****	2.27 (0.322)*
	age [years]	58.0 (7.92)	56.1 (8.28)		53.9 (6.65)*	59.1 (8.59)	60.1 (8.43)*	58.4 (8.59)
	baseline BMI [kg/m^2]	26.1 (3.77)	26.1 (4.72)		27.8 (4.62)*	27.3 (4.66)*	**27.6 (4.12)****	27.2 (4.15)*
	baseline WOMAC	5.62 (9.27)	8.6 (11.5)		6.76 (8.76)	5.92 (9.38)	5.81 (8.68)	6.59 (9.3)
	sex = Female	91 (49.7%)	28 (71.8%)*		13 (36.1%)	198 (56.6%)	192 (56.3%)	292 (58.1%)
	contralateral OA = Yes	21 (15.2%)	6 (18.2%)		6 (22.2%)	51 (20.1%)	48 (18.8%)	66 (17.9%)

subgroup, with an even lower proportion of female participants (LT:46.5% vs 59.5%, MT:40.9% vs 66%, cMF:53.4% vs 59.2%, cLF:29.7% vs 56.3%, $p<0.05$). Thickening subgroups had higher proportions of female participants and tended to be younger. In the patella compartment, these observations differed: thinning subgroups had significantly lower baseline thickness than the stable subgroup (thickening to thinning 2.24mm, accelerated thinning 2.23mm, plateaued thinning 2.27mm vs stable: 2.33mm, $p<0.05$) and no sex differences in thinning or thickening subgroups were observed. The stable patella subgroup was also significantly lower in BMI than the thinning subgroups (thickening to thinning 27.3kg/m^2, accelerated thinning 27.6kg/m^2, plateaued thinning 27.2kg/m^2 vs stable 26.1kg/m^2, $p<0.05$), while this difference was not observed in the tibiofemoral compartments. In all compartments, thinning subgroups had a higher proportion of participants with contralateral OA at baseline and slightly older participants. Baseline WOMAC scores were not significantly different across any of the subgroups for all compartments.

Odds-ratios of radiographic OA incidence and contralateral OA status are estimated by comparing the stable subgroup against all other subgroups per compartment, adjusting for age, sex, and BMI, adjusted OR presented in Table 5.5**.** For all cartilage compartments, belonging to the accelerated thinning subgroup was associated with significantly higher odds of OA incidence (LT:2.68, MT: 5.48, cMF: 10.1, $p<0.001$; cLF:3.17, P:2.47, $p<0.01$), with the medial compartments having the highest OR. Both thickening and thinning cMF subgroups were associated with significantly higher odds of OA incidence (2.61 to 10.1 OR depending on group, $p<0.001$), with thinning subgroups also having higher odds of contralateral OA. All patella (P) thinning subgroups were associated with increased OR for OA incidence (2.47 to 2.56 OR, $p<0.01$). While the accelerated thickening subgroups for the medial tibia and central lateral femur showed high and significant OR (MT: 18.9, cLF:4.97 $p<0.001$), these observations were limited by small sample sizes (n = 7 and 25 respectively, see Table 5.4 for counts per subgroup).

Table 5.5 Adjusted Odds-Ratios of OA incidence and contralateral OA, adjusted for age, sex, and BMI, where * p<0.01 and ** p<0.001 (bold) between the stable subgroup and all other subgroups (accelerated thickening, plateaued thickening, thickening to thinning, accelerated thinning, plateaued thinning, and thinning to thickening). Results reported as OR (95% CI). Empty cells represent subgroups with 0 knees, for subgroup sizes, refer to Table 5.4.

		net thickening			net thinning		
		thinning to thickening	accelerated thickening	plateaued thickening	thickening to thinning	accelerated thinning	plateaued thinning
LT	OA incidence	2.12 (0.99, 4.52)		1.53 (0.57, 4.08)	1.46 (0.93, 2.28)	**2.68 (1.59, 4.51)****	1.17 (0.78, 1.76)
	contralateral OA	1.85 (0.85, 4.05)		1.52 (0.61, 3.79)	1.51 (1.02, 2.24)	1.52 (0.92, 2.50)	0.96 (0.68, 1.37)
MT	OA incidence	1.34 (0.80, 2.24)	**18.9 (3.47, 102.6)****	2.28 (1.29, 4.03)*	0.82 (0.48, 1.39)	**5.48 (2.10, 14.3)****	1.47 (1.00, 2.15)
	contralateral OA	0.96 (0.60,1.55)	1.61 (0.28, 9.31)	1.54 (0.86, 2.77)	1.19 (0.77, 1.83)	2.83 (0.86, 9.29)	1.27 (0.90, 1.79)
cMF	OA incidence	**3.17 (1.71, 5.88)****	8.55 (2.30, 31.8)*	**3.26 (1.69, 6.31)****	**3.67 (2.03, 6.63)****	**10.1 (4.84, 21.0)****	**2.61 (1.49, 4.58)****
	contralateral OA	1.50 (0.93, 2.41)	0.39 (0.05, 3.11)	1.23 (0.71, 2.16)	**2.13 (1.36, 3.34)****	2.56 (1.20, 5.50)	1.63 (1.09, 2.45)
cLF	OA incidence	1.03 (0.64, 1.66)	**4.97 (2.03, 12.2) ****	2.18 (1.30, 3.65)*	1.74 (1.08, 2.81)	3.17 (1.39, 7.22)*	1.06 (0.66, 1.71)
	contralateral OA	0.95 (0.63, 1.41)	5.05 (1.89, 13.5)*	0.54 (0.31, 0.95)	1.02 (0.67, 1.55)	1.09 (0.42, 2.84)	1.11 (0.75, 1.64)
P	OA incidence	2.11 (0.72, 6.18)		2.91 (0.86, 9.89)	2.56 (1.33, 4.95)*	2.47 (1.27, 4.81)*	2.53 (1.33, 4.79)*
	contralateral OA	1.70 (0.65, 4.48)		1.95 (0.72, 5.27)	1.41 (0.84, 2.39)	1.81 (1.07, 3.06)	1.49 (0.91, 2.44)

5.5 Discussion

5.5.1 Cartilage Thickness Trajectory Analysis

Initial cartilage thinning and thickening rates in the non-stable subgroups fit within expected values reported in literature. The thinning[162; 163] / thickening[164] rates in disease establish lower / upper bounds for rates of cartilage change in knees that are at-risk of developing OA but are otherwise healthy. Healthy tibiofemoral rates in literature are centered close to zero with the weight-bearing femoral compartments having the greatest variability[160]; this is consistent with subgroup results. Cartilage change acceleration was non-zero in non-stable trajectories, which logically follows as cartilage could not thin or thicken indefinitely (a majority of tibiofemoral cartilage compartments were stable or plateaued thinning). Observed rates of change (initial velocity) in healthy patellar cartilage skewed negative (61.7% of total trajectories belonging to accelerated thinning, plateaued thinning, thinning to thickening) and were higher in magnitude than the tibiofemoral compartments, as similarly reported in [161; 177]. Plotting all the cartilage thickness trajectories and estimating average rate of initial cartilage thickness change by subgroup highlighted the heterogeneity of the studied population, even between cartilage compartments as factors such as loading and varus/valgus limb alignment can differentially

impact specific cartilage compartments. As suggested by [178], it is possible that previous studies have also observed populations with both thickening and thinning cartilage, but averaging procedures cancelled out any significant effects.

Older age, female sex, and high BMI have been linked to increased likelihood of OA incidence[179]. Interestingly, although there were significant baseline differences in these factors between each subgroup and their stable counterpart, adjusted and unadjusted odds ratios were similar. In other words, age, sex, and BMI differences were not driving increased OA incidence risk among the trajectory subgroups in OAI. Race was not evaluated in relation to cartilage trajectories given insufficient numbers of non-Caucasian participants; recruitment of diverse cohorts remains an important effort to reduce disparities in health research. Within each compartment, the strength of the association between cartilage trajectory and OA incidence varied by subgroup, even among subgroups with similar net changes. This may indicate that trajectories which would be equivalent under two timepoint analysis (plateaued thickening/accelerated thickening/thinning to thickening or plateaued thinning/accelerated thinning/thickening to thinning), have different associations with clinical outcomes. Non-stable trajectories were associated with increased likelihood of OA incidence particularly in the medial compartments of the knee, which is expected given that medial loading tends to be greater than lateral loading. Even thinning trajectories in the patella were associated with higher odds of tibiofemoral OA incidence (~2.5). Cicuttini et al[163] suggests tibiofemoral and patellofemoral cartilage changes have different disease pathways, however it is possible this increased likelihood reflects a common inflammatory mechanism of OA. Notwithstanding, the definition of structural OA incidence by radiographic narrowing (KLG) is limited: studies show joint space narrowing can occur without cartilage thinning (via mensical extrusion)[180] and cartilage thinning can occur without radiographic narrowing[181].

5.5.2 Algorithm Development and Validation

The measurement tool and analysis approach in this study were developed with the purpose of describing cartilage dynamics. Despite the method's robust performance, methodological choices must be justified and limitations acknowledged.

The segmentation model ensemble had sub-voxel accuracy and results were competitive with those in literature (different data splits prevent a direct comparison): Dice overlap coefficients FC 0.897, LT 0.918, MT 0.861, P 0.842, LM 0.895, MM 0.874[171]; at 12m/24m FC 0.894/0.891, LT 0.904/0.900, MT 0.861/0.858[173]. While creating a multi-model ensemble increases computational load during inference, it addresses the shortcomings of individual models without requiring any manually specified post-processing such as removing small connected components (as in [171]) or introducing shape priors (in [173; 174]**)**. In fact, the segmentation training and testing dataset did not reflect the full spectrum of OA pathology and had few non-OA participants, therefore post-processing learned from this dataset could overfit and fail to generalize over the entire OAI dataset. Using aggressive augmentation, independent training splits[175], large validation splits, and a creating a model ensemble[182; 183] likely served to regularize the segmentation algorithms and improve generalizability.

Thickness[158; 159; 164; 166-170; 181; 184-193] and volume[161; 162; 166; 168; 177; 192-197] are the most frequently used cartilage biomarkers in literature; thickness was selected as the outcome variable since it shows similar measurement precision and sensitivity to change as volume[166; 168; 192; 193], and had a large external dataset available for validation, although thickness data was not available for the patella. Scan-rescan precision and accuracy of the thickness analysis method developed in this study were comparable to those of a recent study by Bowes et al[170]. A major limitation of this study is the averaging of cartilage thickness throughout each cartilage compartment. Ideally, analysis would be performed within the patient frame of reference[188; 198] or with anatomical landmarks[189], as location specific thickness changes have been reported in literature[189; 190]. It is possible that simultaneous thinning and thickening within the same cartilage

compartment could be occurring undetected. Further advancements in the thickness algorithm will address this shortcoming to perform local analysis of cartilage changes.

In this study, a trajectory analysis approach used 7 longitudinal observations over 8 years to create a polynomial fit that was robust to individual timepoint noise. Poor polynomial fits were removed from analysis as they likely represent trajectories with rapid changes (ex. thinning to thickening to thinning) where a second degree polynomial is not sufficient to fit the data or significant measurement noise exists at multiple timepoints. Net thickness change, initial velocity, and acceleration values were independent of initial cartilage thickness (this value disappears in the first and second order time derivatives of the polynomial). There are normal variations in cartilage thickness which may be associated with OA risk[199], correlations remained significant even after adjusting for factors associated with baseline cartilage thickness such as sex and BMI. While thickness change metrics were quantified reliably, subgroup definitions were rule-based and the data did not naturally cluster in this form. Representative subgroup curves and names are examples of cartilage trajectories that exist within that subgroup, as cartilage changes exist on a continuum. Net thickness change, initial velocity, and acceleration values are rich descriptors of patient trajectories which can be used with or without the subgrouping scheme in future analyses.

5.5.3 Clinical Relevance of Proposed Methodology

In a recent paper[191] aiming to predict accelerated knee OA incidence from baseline clinical/demographic variables, the addition of cartilage thickness measurements from MRI provided almost no benefit to the classification models. The authors then concluded the high cost of acquisition and analysis of MR images for structural assessment of the joint are not justified. This highlights an important point: this current study is not proposing using longitudinal observations of at-risk populations to predict radiographic OA incidence, as this is neither feasible nor justified. Instead, the methodology and findings presented here hold significant

research potential. Results emphasize the heterogeneity of cartilage dynamics, reaffirming the differences in OA etiology and pathogenesis. Reference values for velocity and acceleration of thickness changes per subgroup per cartilage compartment are established over 8 years, which could serve as valuable benchmarks for developing high fidelity, patient-specific computational models of cartilage degeneration, such as in [198]. Additionally, these velocity and acceleration metrics could be used as objective structural outcomes to investigate OA incidence and progression, understanding which modifiable and non-modifiable risk factors contribute to cartilage dynamics. Research on associations between baseline variables and thickness trajectory subgroups could assist in the a-priori selection of patients in clinical trials of disease modifying OA drugs (DMOAD).

5.5.4 Conclusions

In summary, knees with non-stable cartilage thickness trajectories over 8 years had higher adjusted odds of OA incidence than stable trajectories. Non-stable trajectories did not have a constant rate of thinning or thickening. Up to 73% of knees that would appear stable under a two timepoint comparison between baseline and 8 years, did not have a stable trajectory. Improved phenotyping of these subgroups and further studies on associations between thickness trajectories and clinical endpoints could uncover important insights on the pathophysiology of OA.

6 Deep learning discovery of osteoarthritis biomarkers through dense and hollow point clouds

This chapter contains research that was submitted to the International Society for Magnetic Resonance in Medicine (ISMRM) in 2020 and 2021. These are two proof-of-concept studies exploring the use of point clouds for feature learning from structural and quantitative knee MR imaging sequences. Performing averaging of measured tissue parameters (thickness, T_2 relaxation time) over a region-of-interest is standard practice in osteoarthritis literature. As discussed in the previous chapter, an averaging approach is not ideal as it diminishes spatial precision and lowers sensitivity to local changes. To improve the analysis of structural and quantitative knee MR images, we propose to encode tissues as point clouds and learn OA features directly from these points. Finally, we assess how effectively these learned features perform in discriminating between subjects with and without radiographic OA and examine their utility for predicting incident radiographic OA. Our findings suggest point cloud learned features are promising biomarkers for preclinical OA.

6.1 Introduction

Osteoarthritis (OA) is a painful, whole joint disease responsible for over $15 billion in direct healthcare costs annually in the United States. Greater understanding of modifiable and

non-modifiable OA risk factors is needed to improve preventative care. Several studies have reported distinct cartilage thinning patterns[189 158] and meniscal shapes[200 201] across subjects with increasing disease severity. While these studies have successfully identified morphological features associated with current OA, less research exists on identifying morphological features of preclinical OA. In contrast, quantitative MR sequences, such as $T_{1\rho}$ and T_2 mapping, have been widely investigated for their potential to identify early degenerative changes characteristic of preclinical OA. Although promising, the power of quantitative MR sequences is limited by ineffective feature representation and computationally intensive image processing. As discussed in Chapter 4, handcrafted features, such as region-of-interest averaging, laminar analysis, and texture analysis are limited in their expressive capabilities. In this work, we propose a point cloud deep learning approach for fully automatic extraction of OA features from MR imaging, with the working hypothesis being that features that discriminate between subjects with and without OA can be used to identify preclinical OA. Hollow point clouds are created from structural MR images and used to learn morphological features of OA (cartilage and meniscus shape, thickness, etc). Dense point clouds are created from quantitative MR images, specifically T_2 maps, to learn compositional features of OA (cartilage T_2 texture, local T_2 hotspots, etc). If the models are well calibrated, probability of OA (P(OA)) predicted by the trained network should, in theory, be related to the disease severity: a tissue with a P(OA)=0.1 should be a healthier than a tissue with a P(OA)=0.4, even if both are classified as By representing tissues as point clouds instead of a voxel grid or mesh, we enforce data sparsity while preserving global and local geometric information. Practically, point cloud encoding was chosen as an efficient way limit the number of trainable network parameters and reduce memory demands during training.

6.2 Methods

At the time of writing, the OAI dataset includes MR imaging at seven timepoints over the course of 8 years. Structural imaging was acquired for both knees while quantitative imaging

was acquired for a single knee. Therefore, the datasets for dense and hollow point cloud experiments were split differently and analyzed independently. Specifically, the dense T_2 point cloud experiments use a cross-validation split, while the hollow structural point cloud experiments use a single train/validation/testing split. Each splitting strategy attempts to balance subject demographics and outcomes across the splits.

6.2.1 Dense point cloud creation from T_2 dataset

The OAI dataset includes 25,729 T_2 maps acquired on 3T MR scanners using a 2D multi-slice multi-echo (MSME) sequence[202] (0.313x0.446mm in-plane resolution, 3mm slice thickness, 10/20/30/40/50/60/70ms TE, 2700ms TR). A V-Net[137] neural network developed by Alaleh Razmjoo was trained on 3,921 knee MR images to segment 5 cartilage compartments (lateral femur, medial femur, lateral tibia, medial tibia, and patella). Segmentations were inferred for all the available MSME images in the OAI, after which voxel-wise exponential fitting was performed to calculate T_2 relaxation values[97].

The final T_2 dataset is described in Table 6.1. Cross-validation splits for dense point cloud experiments were created and roughly stratified by demographics and outcomes. Femoral and tibial cartilage compartments are parametrized as dense point clouds. Dense point clouds are a non-ordered collection of 16,384 randomly sampled, non-integer points within the segmentation mask (point-to-voxel ratio roughly 1.6,), where each point encodes its Cartesian coordinates (x,y,z) and T_2 relaxation time (ms), as shown in Figure 6.1.

Each point cloud is zero centered and scaled to its 90th percentile distance, T_2 values in ms are divided by 100 to approximately 0-1 normalized. Creation of single point cloud takes less than 1.5s on IntelCore i7 6700K CPU, making it faster than other parametrization methods.

Table 6.1 Description of subjects and images per balanced split. 20% of data was set aside as a test split, and remaining data used for 5 training/validation splits (64%/16%). Variables are listed as mean with 95% confidence interval or count with percent total. Each split distribution is compared to the population distribution (orange), using a two-sided t-test for continuous variables, and a chi-squared contingency test for categorical variables (**p<0.05**).

				Continuous Descriptors			
		# patients	*# images*	*age [years]*	*BMI [kg/m^2]*	*KOOS*	*WOMAC*
population		4783	22481	63.27 (63.15, 63.39)	28.44 (28.37, 28.50)	86.15 (85.93, 86.36)	10.35 (10.17, 10.53)
test		957	4563	63.36 (63.08, 63.63)	28.47 (28.33, 28.61)	86.23 (85.75, 86.72)	10.16 (9.75, 10.56)
CV0	train	3060	14409	**63.07 (62.92, 63.22)**	28.38 (28.30, 28.46)	86.04 (85.77, 86.31)	10.48 (10.26, 10.71)
	val	766	3509	**63.97 (63.68, 64.27)**	28.60 (28.44, 28.76)	86.47 (85.93, 87.02)	10.06 (9.60, 10.52)
CV1	train	3061	14338	63.27 (63.12, 63.42)	28.47 (28.39, 28.55)	86.17 (85.90, 86.43)	10.45 (10.22, 10.67)
	val	765	3580	63.16 (62.86, 63.47)	**28.24 (28.08, 28.40)**	85.96 (85.42, 86.50)	10.22 (9.78, 10.67)
CV2	train	3061	14308	63.25 (63.10, 63.40)	28.42 (28.34, 28.51)	86.22 (85.96, 86.49)	10.32 (10.10, 10.55)
	val	765	3610	63.23 (62.92, 63.53)	28.43 (28.28, 28.58)	85.73 (85.18, 86.28)	10.72 (10.25, 11.18)
CV3	train	3061	14310	63.32 (63.17, 63.47)	28.38 (28.30, 28.46)	86.14 (85.87, 86.41)	10.37 (10.15, 10.60)
	val	765	3608	62.96 (62.66, 63.26)	28.60 (28.44, 28.76)	86.05 (85.52, 86.58)	10.51 (10.06, 10.97)
CV4	train	3061	14307	63.33 (63.18, 63.48)	28.47 (28.39, 28.55)	86.05 (85.78, 86.32)	10.38 (10.15, 10.61)
	val	765	3611	**62.93 (62.64, 63.22)**	28.26 (28.10, 28.43)	86.42 (85.90, 86.94)	10.48 (10.03, 10.93)

		Categorical Descriptors						
		side = RIGHT	*gender = F*	*KL score > 1*	*TKR*	*OA prog, 1yr*	*OA prog, 2yr*	*OA ever*
population		21901 (97%)	12940 (58%)	9197 (44%)	1161 (5%)	232 (3%)	363 (5%)	11363 (51%)
test		4431 (97%)	**2720 (60%)**	1845 (44%)	204 (4%)	51 (3%)	81 (6%)	2306 (51%)
CV0	train	14037 (97%)	8236 (57%)	5918 (45%)	777 (5%)	152 (3%)	230 (5%)	7290 (51%)
	val	3433 (98%)	1984 (57%)	1434 (44%)	180 (5%)	29 (2%)	52 (5%)	1767 (50%)
CV1	train	13986 (98%)	8188 (57%)	5933 (45%)	770 (5%)	142 (3%)	226 (5%)	7273 (51%)
	val	3484 (97%)	2032 (57%)	1419 (43%)	187 (5%)	39 (3%)	56 (5%)	1784 (50%)
CV2	train	13960 (98%)	8215 (57%)	5869 (45%)	748 (5%)	141 (3%)	224 (5%)	7244 (51%)
	val	3510 (97%)	**2005 (56%)**	1483 (45%)	209 (6%)	40 (3%)	58 (5%)	1813 (50%)
CV3	train	13983 (98%)	8145 (57%)	5821 (44%)	762 (5%)	143 (3%)	225 (5%)	7194 (50%)
	val	**3487 (97%)**	2075 (58%)	1531 (46%)	195 (5%)	38 (3%)	57 (5%)	1863 (52%)
CV4	train	13914 (97%)	8096 (57%)	5867 (45%)	771 (5%)	146 (3%)	223 (5%)	7227 (51%)
	val	**3556 (98%)**	2124 (59%)	1485 (45%)	186 (5%)	35 (3%)	59 (5%)	1830 (51%)

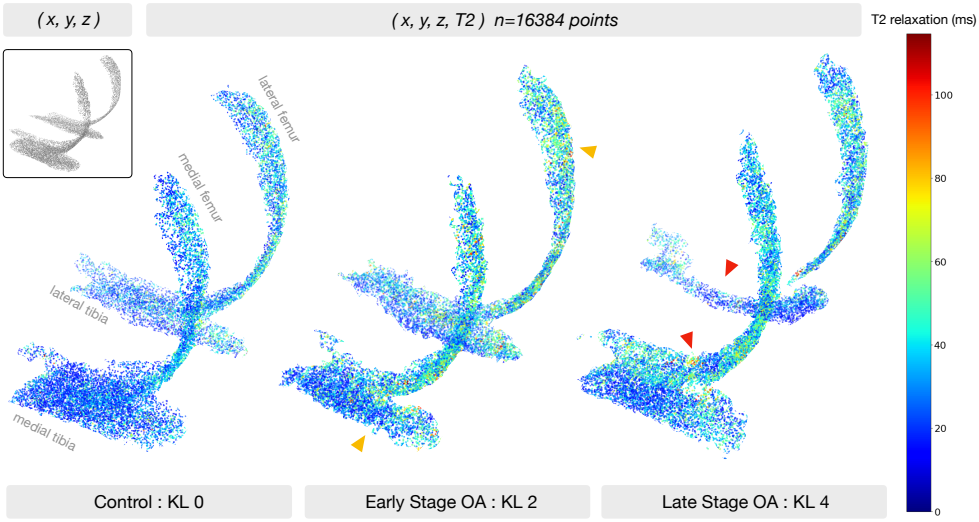


Figure 6.1 Dense femoral and tibial cartilage point clouds with T_2 relaxation values (n x 4) are shown for three study subjects with increasing disease severity. Yellow arrows (L to R) show locally elevated T_2 and loss of coverage in the posterior medial tibia; diffuse elevated T_2 in the superficial layer of the lateral femoral condyle. Red arrows (L to R) show a small T_2 hotspot in the medial femur; complete loss of cartilage in the lateral weight bearing region.

6.2.2 Hollow point cloud creation from DESS dataset

A total of 40,796 DESS images with cartilage and meniscus segmentations were used to create hollow point clouds and split into train/val/test (46%,36%,18%), demographics and outcomes for the dataset are described in Table 6.2. MR imaging parameters and segmentation network details are described in Chapter 5 and [203]. Segmentations were inferred for all the available DESS images in the OAI. Femoral, tibial, and patellar cartilage and menisci point clouds were created using marching cubes and random mesh sampling of 8,192 points per tissue. The experimental setting examined four unique point cloud sets: PAT-FEM-TIB, PAT-FEM, FEM-TIB, and MEN (Figure 6.2) made from patella (PAT), femoral cartilage (FEM), tibial cartilage (TIB) and meniscus (MEN) compartments. Each point cloud set was zero centered and -1 to 1 normalized.

Table 6.2 Description of subjects and images in testing, training, and validation splits. For continuous descriptors, mean and 95% confidence interval is reported, differences against population (all OAI data) are tested using a two-sided t-test. For categorical descriptors, count and percent are reported and differences tested with a chi-squared contingency test. Significant differences ($p<0.05$) are highlighted in blue. OA=Osteoarthritis (r=radiographic incidence, sr=symptomatic and radiographic incidence), KOOS=Knee Injury Osteoarthritis and Outcome score, WOMAC=Western Ontario and McMaster Universities Arthritis Index, KL=Kellgren Lawrence.

			Continuous Descriptors			
	# patients	*# images*	*age [years]*	*BMI [kg/m^2]*	*KOOS*	*WOMAC*
population	4795	44929	63.37 (9.25)	28.46 (4.86)	86.58 (16.70)	10.24 (14.22)
test	819	7348	63.20 (9.18)	28.46 (5.01)	86.77 (16.58)	10.13 (14.11)
train	2069	18900	62.92 (9.07)	28.51 (4.73)	87.26 (16.14)	9.64 (13.64)
val	1620	14548	63.24 (9.36)	28.15 (4.85)	86.99 (16.42)	9.84 (13.99)

Categorical Descriptors					
side = RIGHT	*gender = F*	*KL score > 1*	*rOA within 6yr*	*srOA within 6yr*	*OA ever*
22223 (49%)	25827 (57%)	17282 (42%)	1671 (9%)	367 (10%)	22186 (49%)
3691 (50%)	4089 (56%)	3194 (43%)	259 (8%)	51 (8%)	3496 (48%)
9453 (50%)	11624 (62%)	7864 (42%)	863 (10%)	206 (11%)	8840 (47%)
7328 (50%)	7629 (52%)	6224 (43%)	549 (8%)	110 (9%)	6835 (47%)

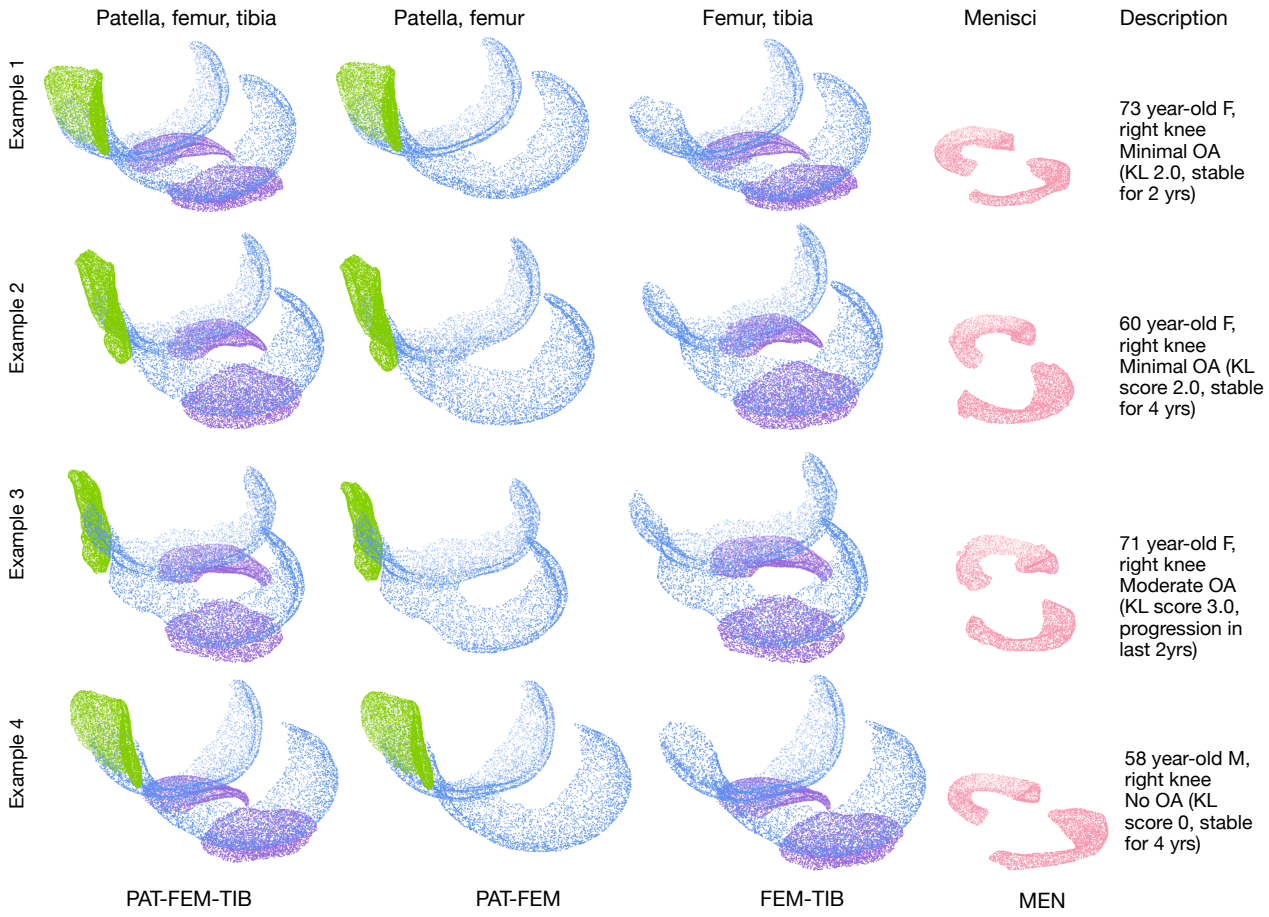


Figure 6.2 Four example subjects with their processed hollow point cloud set. PAT- FEM-TIB, PAT-FEM, FEM-TIB, and MEN compartment combinations were used to train point cloud networks for OA diagnosis. Each compartment is represented by 8192 randomly sampled points. FEM= femur, TIB=tibia, PAT=patella, MEN= menisci

6.2.3 Point cloud network training and statistical analysis

After hollow and dense point cloud datasets were created, a PointNet++like network architecture (LSAnet [204] [205] 2.3M parameters, illustrated in Figure 6.3) was trained for each experiment. Diagnosis of current radiographic OA (binary classification of KL score>=2) was selected as the learning task. Label smoothing was used during model training to encourage learning calibrated outputs and to prevent overconfident predictions: predicted probability of OA should reflect the actual probability of sample OA. The architecture was implemented in Tensorflow 1.8 and trained on an NVIDIA V100 32GB GPU. Dense point cloud networks were

trained for 65 epochs using cross entropy loss with label smoothing=0.3, random point dropout, batch size=20, learning rate=0.0015, and Adam optimizer. At test time, logits averaging was used to ensemble the 5 cross validation model predictions. Hollow point cloud networks were trained for 20 epochs using cross entropy loss with label smoothing=0.2, batch size=20, learning rate=0.001, and Adam optimizer.

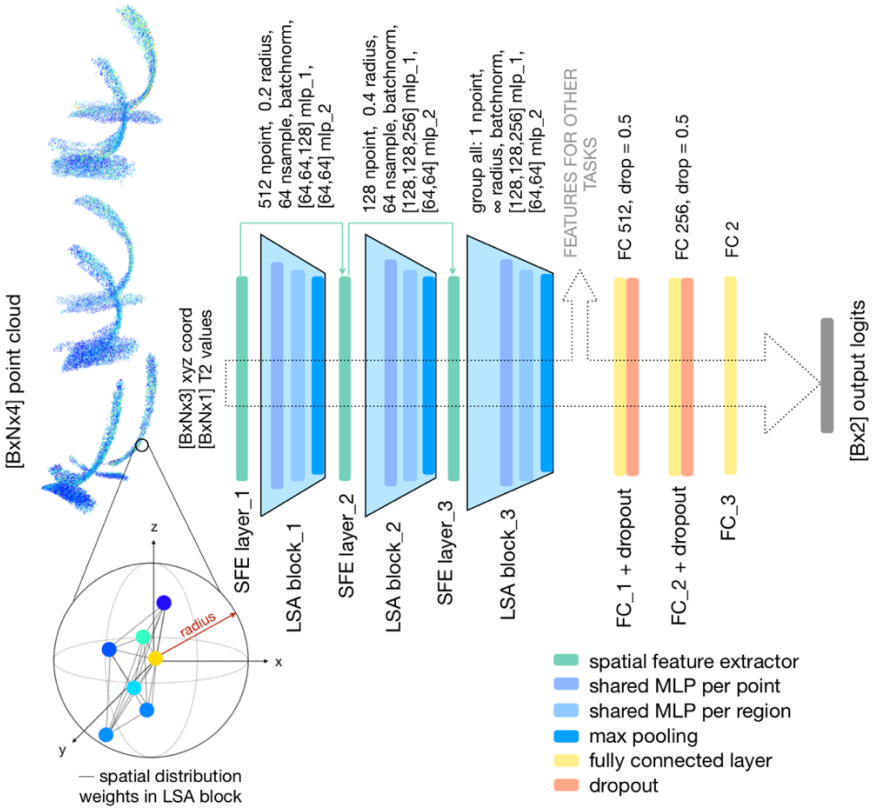


Figure 6.3 Local Spatial Aware (LSA) blocks and Spatial Feature Extractors (SFE) from the LSA-Net architecture proposed in Chen et al, allow neural networks to learn increasingly abstract features from the full point cloud by lifting the dimensions of the spatial xyz features to match the features extracted by the previous LSA block before concatenation. Network parameters are specified above each block in the schematic. Figure input shows example using dense T_2 point clouds, network architecture for hollow structural point clouds is identical with one fewer channel for input [BxNx3].

For the dense point cloud experiments, false positive predictions were examined to understand if the network learned patterns that are characteristic of OA but may precede radiographic OA. Receiver Operating Characteristic, Precision Recall, and calibration curves

were plotted for the test set in each experiment. For the hollow point cloud experiments, downstream statistical analysis was conducted to understand the utility of the learned features for time-to-event prediction of radiographic OA, adjusting for clinical factors. From the test set, patients without radiographic OA at baseline were selected to extract right censored time-to-event data for radiographic OA incidence (860 observations, 103 events). Model concordance index (c-index) was compared between a baseline Cox Proportional Hazards Regression model using only clinical factors and four models with clinical factors and learned OA features, i.e. the P(OA) from a specific point cloud model.

6.3 Results

6.3.1 Dense T_2 point clouds

Test set performance for the diagnosis of radiographic OA using dense point clouds is reported in Table 6.3 and test set ROC, PR, and calibration curves are plotted in Figure 6.4. The networks had higher sensitivity on the test set than on the validation set, while specificity was lower in test compared to validation (except for the CV2 model). The 5-model ensemble had the highest performance with a final sensitivity of 0.8244, specificity of 0.8259, ROC AUC 0.904, and PR AUC 0.899.

The KL grades of the false positive (FP) and false negative (FN) test set predictions from the model ensemble model are examined. 94.7% of FN were KL 2, 5.3% KL 3; FP were 1.5x more likely to be KL 1 than 0. Among the FP group, 9.3% would progress to have radiographic OA within 1 year, and 13.3% within 2 years, which was higher than the average for the entire non-OA group (3.2% within 1yr, 5.6% within 2yr). This suggests the point cloud network is capable of learning T_2 patterns characteristic of early OA, 1-2 years before the onset of radiographic changes.

Table 6.3 Individual network performance on each splits' validation and common test data. Performance of the logits-averaged ensemble on test data is also reported. Detection of T_2 changes close to the onset of radiologically defined OA requires greater sensitivity than specificity, therefore the checkpoints with the highest sensitivity were chosen for inference (epoch 5, 33, 15, 23, 7 for networks CV0-CV4). CV3 had the best standalone performance, while the ensemble was the highest performing overall, competitive with previous literature and current multimodal approaches.

	Validation		*Test*		*Test (ensemble)*	
	Sensitivity	*Specificity*	*Sensitivity*	*Specificity*	*Sensitivity*	*Specificity*
CV0	0.7929	0.7804	0.8190	0.7227		
CV1	0.8132	0.7767	0.8255	0.7386		
CV2	0.8058	0.8099	0.7881	0.8250	**0.8244**	**0.8259**
CV3	0.7427	0.8642	**0.7832**	**0.8340**		
CV4	0.7838	0.7466	0.8363	0.7373		

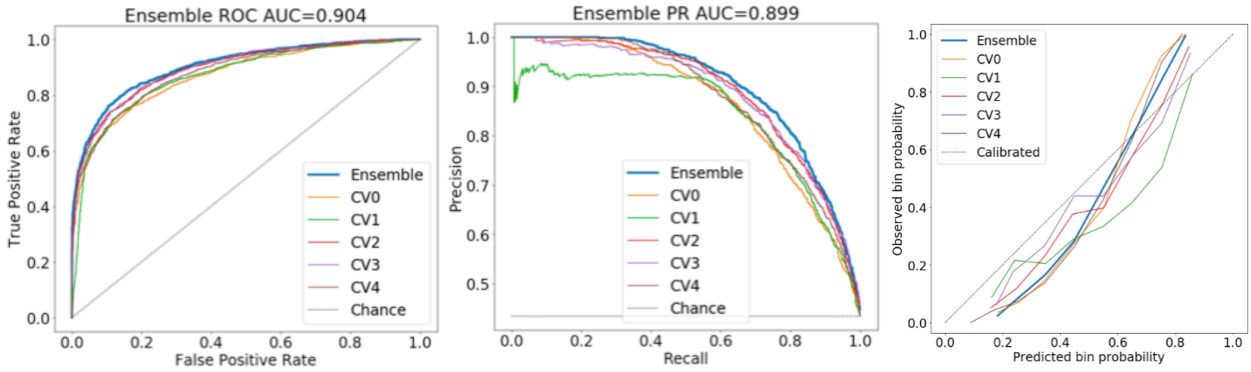


Figure 6.4 ROC, PR, and reliability diagram for OA classification. An ensemble of the 5 cross-validation (CV) folds achieved an ROC-AUC and PR-AUC 0.904,0.899 respectively. Reliability diagram reveals overestimation of probability of OA at low probabilities.

A recent study by Pedoia et al.[206] performed a similar OA diagnosis task on the baseline OAI T_2 dataset and achieved a sensitivity and specificity of 74.53%, 76.13% by using atlas-based registration for cartilage segmentation, then flattening tibial and femoral cartilage T_2 maps, and training a DenseNet convolutional neural on the final maps. Although a direct comparison was not warranted due to different dataset splits, the proposed point cloud method for learning features from T_2 maps showed a promising improvement of 7.91%, 6.46% in

sensitivity, specificity over the previous deep-learning based study, while benefitting from significantly faster processing.

6.3.2 Hollow structural point clouds

Test set performance for the diagnosis of radiographic OA using hollow point clouds is plotted in Figure 6.5. FEM-TIB, PAT-FEM, and PAT-FEM-TIB point cloud models had comparable test performance in the OA diagnosis task (ROC AUC values 0.903, 0.898, and 0.897), while MEN was lower at 0.88. Similarly, the PR AUC values were 0.898, 0.894, and 0.896, while MEN was 0.868. The PAT-FEM-TIB point cloud model was the best calibrated, followed by FEM-TIB, PAT-FEM, and MEN (Brier scores of 0.126, 0.128, 0.138, 0.14).

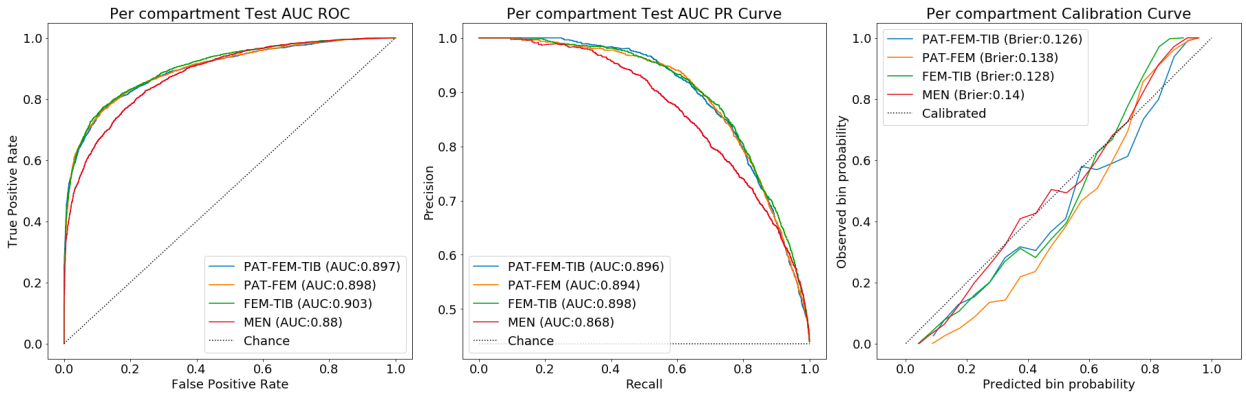


Figure 6.5 Per compartment test results on pretext OA diagnosis task. (L to R) ROC curve, PR curve, and calibration curve with AUC and Brier performance metrics.

Baseline Cox PH Regression with age, gender, BMI category, and KL variables resulted in a c-index of 0.742, while the addition of each shape biomarker increased c-index to 0.752-0.759. Age and gender coefficients were not significant in any of the models. All shape biomarkers except PAT-FEM were significant, with hazard ratios (95%CI) for PAT-FEM-TIB of 2.86 (1.13, 7.27), FEM-TIB 3.95 (1.50, 10.4), and MEN 4.71 (1.82, 12.2). Tabulated Cox PH model results in Table 6.4.

Time-to-event analysis with Cox PH Regression allowed for modeling of radiographic OA incidence using right censored data from the OAI dataset. The results show that structural

features learned by the networks to discriminate OA subjects from non-OA subjects were useful in predicting time-to-event for onset of radiographic OA. We expected PAT-FEM-TIB model features to have the best discriminative and predictive ability, as it was the largest point cloud model (4 tissues * 8192 points = 24576 total) and could encode information on patella-femur and femur-tibia alignment. Surprisingly, FEM-TIB and MEN shape features were more meaningful in the time-to-event analysis than PAT-FEM-TIB shape features. We also observe that MEN network was less effective at the pretext task but had the highest hazard ratio for incident OA, suggesting that high discriminative task performance is not a necessary condition for meaningful feature encoding for downstream predictive tasks.

6.4 Discussion

In this work we demonstrated the potential to use both dense and hollow point clouds to learn OA features from quantitative and structural MR imaging. There are several advantages to the proposed method for OA feature learning through point cloud encoding. (1) Point clouds leverage the representational power of deep learning while overcoming the inefficient representations of such: avoiding the N^3 complexity of working within a voxel grid, the N^2 complexity of mesh methods which are limited to describing the surface of an object, and the artifacts introduced with 2D projection and warping methods. (2) Dense point clouds are a raw representation of T_2 values without requiring ROI placement, averaging, or pooling operations. Point cloud networks can be sensitive to both global changes in T_2, as well as subject specific hotspots. Likewise, hollow point clouds capture global and local differences in tissue thickness, shape, and surface heterogeneity. (3) Point clouds are permutation invariant and able to extract meaningful features without the need to establish point to point correspondence, which could aid the clinical translation of these methods as it lowers the computation burden. (4) This

Table 6.4 Cox Proportional Hazard Regression coefficients for each of the 4 shape biomarker models and the baseline clinical model (bottom row). Only subjects without OA at timepoint 0 are included in the model, events are defined as 1 for incident radiographic OA and 0 for no OA observed (right censored). Time-to-event is defined as the first recorded month where KL>= 2.

PAT-FEM-TIB

	coef	exp(coef)	se(coef)	exp(coef) lower 95%	exp(coef) upper 95%	z	p
PATFEMTIB	1.05	2.86	0.48	1.13	7.27	2.21	0.03
age	0.02	1.02	0.01	0.99	1.04	1.36	0.17
gender	0.21	1.23	0.20	0.83	1.82	1.04	0.30
BMI	0.22	1.25	0.11	1.00	1.56	1.98	0.05
KL	1.16	3.20	0.22	2.09	4.90	5.36	<0.005

FEM-TIB

	coef	exp(coef)	se(coef)	exp(coef) lower 95%	exp(coef) upper 95%	z	p
FEMTIB	1.37	3.95	0.49	1.50	10.38	2.78	0.01
age	0.02	1.02	0.01	0.99	1.04	1.41	0.16
gender	0.30	1.34	0.20	0.90	1.99	1.46	0.14
BMI	0.22	1.25	0.11	1.00	1.55	1.99	0.05
KL	1.11	3.05	0.22	1.98	4.68	5.09	<0.005

PAT-FEM

	coef	exp(coef)	se(coef)	exp(coef) lower 95%	exp(coef) upper 95%	z	p
PATFEM	0.86	2.37	0.56	0.78	7.13	1.53	0.13
age	0.02	1.02	0.01	1.00	1.04	1.62	0.11
gender	0.23	1.26	0.20	0.85	1.86	1.15	0.25
BMI	0.24	1.27	0.11	1.02	1.58	2.12	0.03
KL	1.18	3.26	0.22	2.12	5.00	5.40	<0.005

MEN

	coef	exp(coef)	se(coef)	exp(coef) lower 95%	exp(coef) upper 95%	z	p
MEN	1.55	4.71	0.49	1.82	12.20	3.19	<0.005
age	0.02	1.02	0.01	0.99	1.04	1.51	0.13
gender	0.28	1.32	0.20	0.89	1.97	1.39	0.16
BMI	0.24	1.28	0.11	1.03	1.58	2.22	0.03
KL	1.06	2.89	0.22	1.87	4.47	4.78	<0.005

BASELINE

	coef	exp(coef)	se(coef)	exp(coef) lower 95%	exp(coef) upper 95%	z	p
age	0.02	1.02	0.01	1.00	1.04	1.67	0.10
gender	0.20	1.23	0.20	0.83	1.81	1.02	0.31
BMI	0.27	1.30	0.11	1.05	1.62	2.39	0.02
KL	1.27	3.55	0.21	2.35	5.36	6.00	<0.005

method could easily scale to other tissues such as bones and ligaments, using a combination of dense and hollow point clouds for a multi-structural whole-joint diagnosis. Moreover, dense point clouds represent a lightweight encoding for multiparametric imaging as a single point can encode information from several image contrasts. (5) Finally, hollow point clouds are agnostic to imaging parameters, that is, as long as a segmentation exists for the tissue of interest for each imaging type, point clouds from several sources can by combined for analysis. A potential limitation our method is that a point cloud parameterization for knee tissues simultaneously encodes cartilage shape, T_2, thickness, and patellar-femur/femur-tibia alignment, which makes it difficult to isolate the contribution of specific tissue parameters such as cartilage thickness or T_2.

6.5 Future directions

Encoding MR information in point clouds is a simple, yet effective method for learning features from structural and quantitative MR imaging. Ongoing experiments in our group are using point cloud networks to regress radiographic OA severity (continuous Kellgren-Lawrence grade rather than binary classification) and subject reported Knee Injury and Osteoarthritis Outcome Scores (KOOS) from structural point clouds. Using a similar setup to the hollow point cloud experiments, features are used in a time-to-event analysis to predict the onset of radiographic OA and pain. Future research will focus on implementing techniques for point cloud interpretability, including point occlusion and a recently proposed point-masking technique [207] to identify regions of highest importance for OA and pain prediction.

7 Lumbar intervertebral disc characterization through quantitative MRI analysis: An automatic voxel-based relaxometry approach

The following manuscript is reformatted, shortened, and reproduced with full permission from the publisher. It appeared as:

Iriondo, C, Pedoia, V, Majumdar, S. Lumbar intervertebral disc characterization through quantitative MRI analysis: An automatic voxel-based relaxometry approach. *Magn Reson Med*. 2020; 84: 1376– 1390. https://doi.org/10.1002/mrm.28210

7.1 Abstract

The aim of this work was develop an automated pipeline based on convolutional neural networks to segment lumbar intervertebral discs and characterize their biochemical composition using voxel-based relaxometry, and establish local associations with clinical measures of disability, muscle changes, and other symptoms of lower back pain. This work proposes a novel methodology using magnetic resonance imaging (n=31, across the spectrum of disc degeneration) that combines deep-learning based segmentation, atlas-based registration, and statistical parametric mapping for voxel-based analysis of $T_{1\rho}$ and T_2 relaxation time maps to

characterize disc degeneration and its' associated disability. Across degenerative grades, the segmentation algorithm produced accurate, high confidence segmentations of the lumbar discs in two independent datasets. Manually and automatically extracted mean disc $T_{1\rho}$ and T_2 relaxation times were in high agreement for all discs with minimal bias. On a voxel by voxel basis, imaging based degenerative grades were strongly negatively correlated with $T_{1\rho}$ and T_2, particularly in the nucleus. Stratifying patients by disability grades there were significant differences in the relaxation maps between minimal/moderate vs severe disability; average $T_{1\rho}$ relaxation maps from the minimal/moderate disability group showed clear annulus nucleus distinction with a visible midline while the severe disability group had lower average $T_{1\rho}$ values with a homogeneous distribution. This work presented a scalable pipeline for fast, automated assessment of disc relaxation times, and voxel-based relaxometry that overcomes limitations of current region of interest based analysis methods and may enable greater insights and associations between disc degeneration, disability, and lower back pain.

7.2 Introduction

Low back pain (LBP) is the leading cause of disability globally[1], with a 38%[208] average lifetime prevalence. Treatments, lost wages, and reduced productivity cost the US over $100 billion[209] every year. Although LBP is widespread, its clinical presentation is complex and pathophysiology poorly understood[210]. Identifying patients' pain generating structures and determining the appropriate treatment course remains a challenge[69; 211]: despite a sixfold increase in Medicare expenditures on LBP treatments over 10 years, patient outcomes have not been improved. There is an urgent need for the discovery of non-invasive biomarkers that distinguish LBP phenotypes.

A common mechanism for developing LBP is intervertebral disc degeneration which occurs when disc homeostasis is perturbed by injury or aging[212]. A cascade of biochemical and micro-structural changes take place, including loss of glycosaminoglycans, disorganization of

annular collagen, and dehydration[213]. These early stage changes precede large scale morphological changes which are associated with pain and disability[214]. Conventional MR imaging sequences and grading systems (Pfirrmann[215]/modified Pfirrmann[216]) are used to determine the severity of disc degeneration through qualitative assessment of disc morphology and signal intensity of the nucleus and annulus. These methods are limited by moderate inter-rater reproducibility[215] and broad binning of disc phenotypes[217]. Quantitative MR imaging (qMRI) is powerful tool capable of detecting local variations in disc composition, however its use is limited by coarse, unreliable, and slow manual analyses methods[218-222].

$T_{1\rho}$ mapping, or spin-lock imaging, is a qMRI sequence that probes slow interactions between bulk water and extracellular matrix macromolecules by applying a continuous, low-frequency RF pulse. T_2 mapping, or spin-spin imaging, is a quantitative sequence sensitive to hydrated collagen and its orientation. These sequences create parametric maps that reflect the spatial distribution of biochemical components within an imaged tissue. Both $T_{1\rho}$ and T_2 relaxation times are strongly positively correlated with hydration and glycosaminoglycan content, and negatively correlated with clinical grades of disc degeneration in human intervertebral disc studies.[218; 223-228] For image analysis, the referenced studies calculate average $T_{1\rho}$ and T_2 relaxation times in the whole disc or within user-defined regions of interest (ROIs)– anterior annulus, posterior annulus, and nucleus. The averaging operation performed disregards potentially relevant information about the local distribution of relaxation values, thus decreasing the method's ability to capture subtle changes in biochemistry. In hip and knee cartilage studies[229; 230], local analysis of $T_{1\rho}$ and T_2 relaxation times revealed patterns that could differentiate between osteoarthritic patients and healthy patients, whereas these patterns were not detectable with ROI analysis. Additionally, variability in manual ROI placement introduces selection bias in the quantification of relaxation times and limits method scalability.

While manual ROI methods are common for qMRI analysis, there is longstanding interest in the automation of tools for conventional MR image analysis. For example,

intervertebral disc segmentation on sagittal T_2-weighted images is tackled using computer vision methods including: graph-cuts[231], fuzzy clustering[219], shape modeling[232], and active contours[233]. Classical computer vision approaches have been moderately successful in small datasets of healthy patients, however the handcrafted features they rely on do not generalize well to unseen data or sequences with highly anisotropic voxels. Spinal tissues vary in intensity, volume, shape, and position within the spine, while signal-to-noise ratio (SNR) depends on acquisition parameters. This data diversity presents multiple challenges to classical algorithm development. Recent advancements in convolutional neural network (CNN) architectures and training strategies, have enabled the development of algorithms that can learn the general image features needed to accurately segment one or multiple spinal structures, even on small datasets[234; 235].

We therefore propose a novel analysis pipeline to address current limitations in the sensitivity, reliability, and scalability of quantitative imaging analysis. Unlike ROI based approaches, our method combines deep-learning segmentation and atlas-registration to perform analyses voxel-wise. Our method leverages a recently published CNN to segment the intervertebral disc and guide the registration. We hypothesize that a voxel-based relaxometry approach will reveal localized differences in disc biochemical composition between patients, while still correlating strongly with established measures of disc degeneration.

7.3 Methods

An overview of the voxel-based relaxometry pipeline is shown in Figure 7.1 and consists of three parts: disc segmentation and registration, image fitting, and statistical analysis. Intervertebral discs are segmented automatically, after which the mask for each disc level is used as input into the registration algorithm. The goal of spatial registration is to find a mapping between the input disc mask and a template disc mask, in other words, to find a deformation

field that, when applied to the input disc, will create spatial correspondence between the input

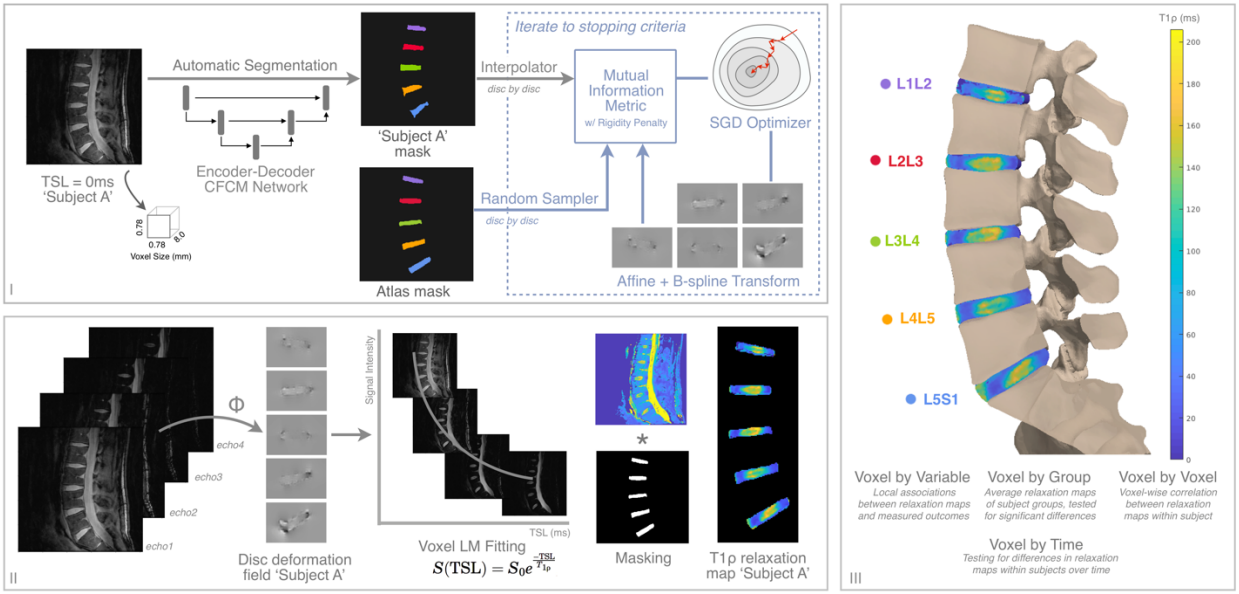


Figure 7.1 Overview of segmentation (I), registration (I &II), fitting (II), and statistical analysis pipeline for lumbar intervertebral disc characterization with qMRI. Once optimized, one or multiple qMRI slices are fed into the 2D segmentation algorithm, after which a single mask or a stack of masks enter the registration procedure. Once deformation fields for each disc are found, they are applied to the various echos of the qMRI sequence before mono-exponential fitting. For visualization purposes, the registered $T_{1\rho}$ relaxation map for Subject A is rendered on a spine mesh in step III. If multiple qMRI sequences exist, such as T_2 relaxation maps, the deformation fields found in step I are applied to additional sequences in step II. Details on network implementation in Supporting Information Figure S1. TSL = Time Spin Lock, CFCM = Coarse-to-Fine-Context-Memory, SGD = Stochastic Gradient Descent, LM = Levenberg-Marquardt

and the template. This deformation is applied to all images before fitting image intensities to calculate relaxation times. Once all subjects are registered to the same space, statistical analyses are performed at each voxel. Mask-guided registration was deemed necessary when intensity-guided registration failed to accurately register the discs and image similarity metrics were unreliable indicators of registration performance. Disc and vertebra vary in intensity, volume, and positioning between subjects, making these tissues ill-suited for intensity-based registration methods.

7.3.1 Datasets

The approach was developed and evaluated on lumbar spine MR $T_{1\rho}$ weighted images from two studies (study A[236] and B[237]) in compliance with the Institutional Review Board. Results from Dataset A and B are presented in the text and color coding of the disc levels is carried throughout all the figures. Dataset A included 16 subjects (10 with documented LBP, 6 controls) scanned at a single time point. The study acquired a single slice $T_{1\rho}$ map (2D Fast Spin Echo) and T_2 weighted images aiming to quantify the biochemical signature of symptomatic degenerative discs. Dataset B consisted of 15 patients with documented LBP scanned at baseline, with 4 returning for a followup scan within a year. The study acquired multi-slice $T_{1\rho}$/T_2 maps (3D Spoiled Gradient Echo), T_2 weighted images, and paraspinal muscle fat-fraction maps with the goal of identifying MR biomarkers related to pain and disability. Demographic variables, clinical variables, and MR sequences for each dataset are detailed in Table 7.1. Categorical variables in each dataset are compared with Fisher's exact test, while continuous variables are compared with two-sided t-tests.

7.3.2 Segmentation Method

Ground truth masks for segmentation network training were generated by annotating lumbar discs L1L2–L5S1 on a single sagittal slice (Dataset A, fully manual) or multiple sagittal slices of the $T_{1\rho}$ sequence (Dataset B, 3D region growing algorithm with manual seeds and manual edits[237]) with an in-house spline-based annotation tool in Matlab 2018a (Mathworks). Throughput for manual annotations was ~90 seconds per disc per slice (7.5 minutes per slice). Data was split per-subject, and a 5-fold cross validation strategy was used to train 5 identical CFCM networks[238] with a 80/20 (62 slices/18 slices) train-test division ensuring Dataset A and B were each represented in the splits. Image preprocessing, network architecture, training, and hyperparameter details in Figure 7.2. Each image was loaded, adaptive histogram equalized to enhance the appearance of local low-contrast tissues, then normalized to zero mean and unit

Table 7.1 Demographic variables, clinical descriptors, and MR acquisition parameters for lumbar spine datasets. Continuous variables described as mean (+/- standard deviation), categorical variables listed as number of subjects (% of total) or number of discs (% of total). IPAQ = International Physical Activity Questionnaire, VAS = Visual Analog Scale for pain, ODI = Oswestry Disability Index, SF36 = Short Form 36 Health Survey.

Demographic Variables

	# Subjects	Age [y]	BMI [kg/m^2]	Height [m]	Sex = F	Is LBP Patient	Follow-up Scan
Dataset A	16	41.7 (±11.9)	27.2 (±4.7)	1.77 (±8.6)	4 (25.0%)	10 (62.5%)	--
Dataset B	15	48.6 (±13.9)	26.1 (±4.5)	1.71 (±9.4)	9 (60.0%)	15 (100%)	4 (21.1%)

Clinical Descriptors

	Pfirrmann	Modified Pfirrmann	Discography +ve Discs	meanVAS / maxVAS	IPAQ>0	ODI	SF36 physical	SF36 mental
Dataset A	2.2 (±0.7)	---	6 (60.0%)	--	--	28.3 (±26.4)	42.9 (±12.3)	49.4 (±13.6)
Dataset B	--	3.6 (±2.1)	--	6.8 (±1.9) / 8.6 (±1.3)	8 (53.3%)	44.9 (±17.5)	---	---

MR Acquisition Parameters

	Scanner	Coil	Sequence Name	View	FOV [mm]	Matrix Size (x,y)	TR/TE [ms]	TSL [ms], FSL [Hz]	Pixel BW [Hz]	Time [m:s]	Purpose
Dataset A	3T GE Excite Signa	8ch CTL spine coil	T1ρ FSE	Sag	200	256,192	2000/80	0/40/80/120, 300	125	6:36	T1ρ relaxometry
			T2 FSE	Sag	200	320,224	5000/70	---	122	2:25	Pfirrmann grading
Dataset B	3T GE Discovery MR750w	8-12ch embedded GEM coil	MAPSS T1ρ/T2	Sag	200	256,128	5.7/51	T1ρ 0/10/40/80, 300 T2 0/12/25/51	488	11:33	T1ρ, T2 relaxometry
			T2 CUBE FSE	Sag	240	160,160	2500/91	---	244	4:13	Modified Pfirrmann grading
			IDEAL IQ	Ax	180	180,180	11/3.9	---	651	6:37	Paraspinal muscle qMR

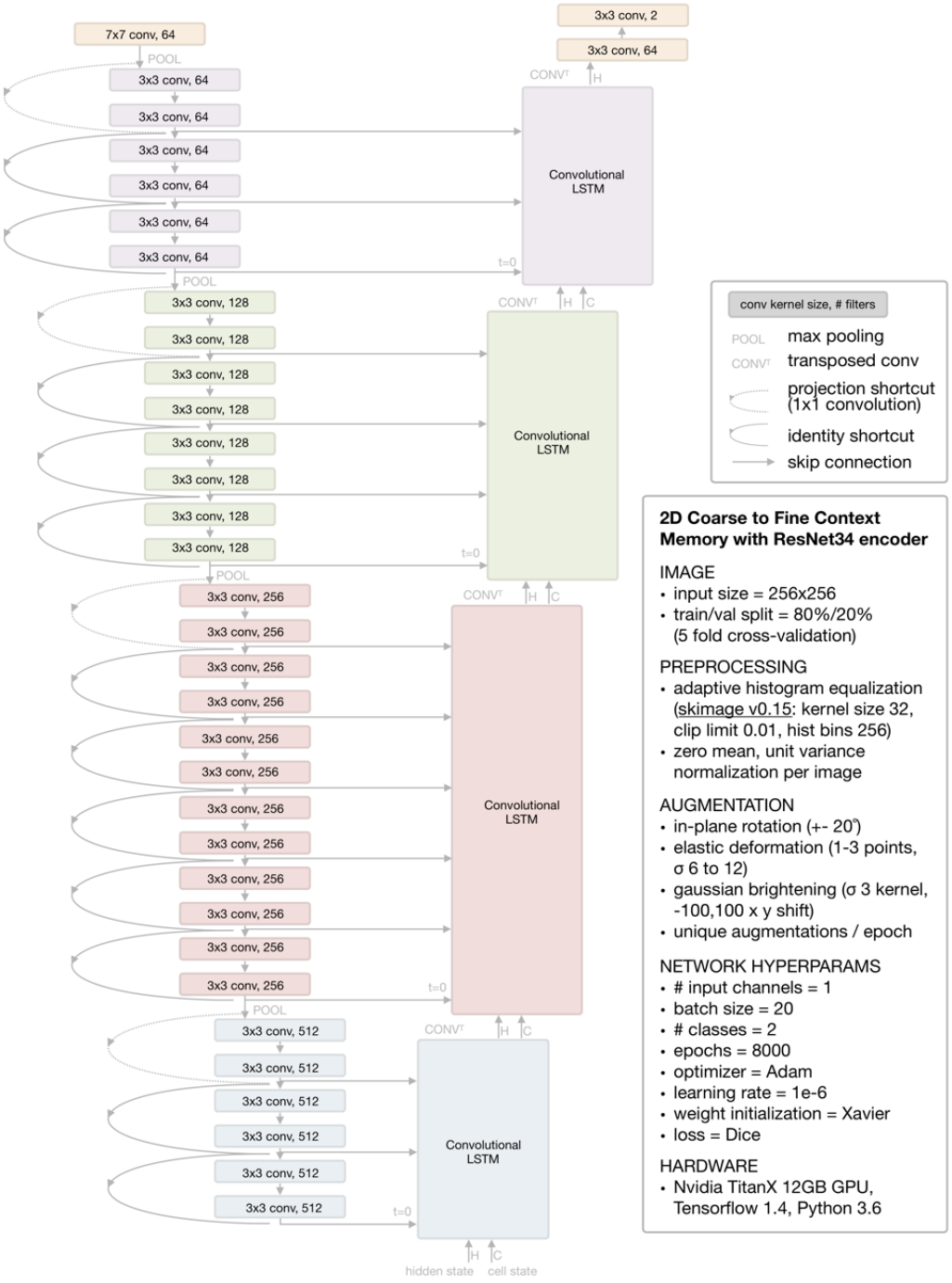


Figure 7.2 Ground truth segmentations were saved as 256x256 binary masks (one 2D mask per slice). The final dataset for network training included 38 scans from 31 unique patients, with a total of 80 segmented slices.

variance. A 2D coarse-to-fine context memory (CFCM) segmentation network (Tensorflow 1.4, Python 3.6) was trained end-to-end using full image slices (256x256) as a single channel input. The CFCM network replaces the decoding path in a classic encoder-decoder with a convolutional long-short-term-memory unit that serves as a memory mechanism to fuse different feature scales and receptive fields, while the encoding path is a ResNet34. Unique augmentations were produced every epoch, introducing enough variability to regularize network training. Aggressive online augmentations were applied to every batch: rotation (in-plane -20 to 20˚), elastic deformations (1 to 3 points, σ 6 to 12), and localized gaussian brightening (image intensity scaled with gaussian kernel σ 3, with -100 to 100 x,y shift). Images underwent adaptive histogram equalization (sklearn v0.15, kernel size 32, clip limit 0.01, histogram bins 256) and zero mean, unit variance normalization. Xavier initialization was used for network weights, trained for 8000 epochs with Dice loss[149], batch size 20, and Adam optimizer (learning rate 1e-6, epsilon 1e-8) on a single Nvidia TitanX GPU, saving the last checkpoint for inference on the test set. After training each of the 5 networks inference was run on an independent test set and segmentations fed into the registration pipeline. To properly evaluate the network's generalization capability, all segmentation results presented herein were inferred using the single network that never trained on those subject's slices. For future applications of this pipeline on new $T_{1\rho}$-weighted data, a 5 network ensemble (logits averaging) would be recommended for segmentation inference.

The performance of the segmentation algorithm is evaluated per disc using semantic segmentation metrics (Dice overlap, mean surface distance, % volume difference, sensitivity, and precision). The 2D version of the metrics is used to analyze single segmented slices from Dataset A, while the 3D version analyzes stacks of segmented slices from Dataset B. To further evaluate segmentation performance, mean disc $T_{1\rho}$ and T_2 relaxation times are extracted using the manual segmentation and the inferred segmentation. Biomarker extraction accuracy on each disc is evaluated by comparing manually and automatically extracted mean $T_{1\rho}$,T_2 values

with Pearson correlation, a paired two-sided t-test for differences, and Bland-Altman analysis for bias. Segmentation performance and biomarker extraction are contextualized with radiological scores for degeneration Pfirrmann/Modified Pfirrmann, scored on a 1 to 5 scale and 1 to 8 scale in terms of increasing degeneration. To examine the effect of changing the segmentation algorithm, three U-Net variants were trained and evaluated as above, results in Table.

7.3.3 Registration

Inferred segmentations were registered to a lumbar disc atlas using Elastix[239], ElastixFromMatlab wrapper (CNRS/Riverside Research), and Matlab2018a. Segmentations were post-processed, identifying connected components in 2D or 3D larger than 125 voxels and labelling them inferior-superior direction (L5/S1 to L1/2). Registration was performed between the inferred disc mask and the atlas disc mask on a disc by disc basis. A healthy spine mask without gross morphological deformities was selected as the atlas to minimize registration artifacts (another healthy spine and a degenerated spine mask were tested as atlases in robustness experiments, extracted patterns were similar). Per disc, the mask is translated to align with the centroid of the atlas disc mask before the two-step registration. First, a 4-resolution recursive pyramidal affine registration rigidly scales, rotates, and shears the disc mask providing initialization for the second step. Then, a b-spline registration elastically deforms the disc segmentation, guided by mutual information with a rigidity penalty term to avoid large local deformations. The two-step registration maximizes the overlap between the inferred disc mask and the template disc mask while preserving the original topology of the inferred disc. The resulting 2D (Dataset A) and 3D (Dataset B) deformation fields are applied to all $T_{1\rho}$, T_2 echos and a two-parameter Levenberg-Marquardt monoexponential fitting is performed voxelwise to create parameter maps of $T_{1\rho}$, T_2 relaxation times in the registered space. B-spline registration parameters including final grid spacing (2), iterations (200), and rigidity penalty weight (0.77) were selected via Bayesian optimization, a method commonly used for hyperparameter tuning

of machine learning models. Bayesian optimization performs registration over many iterations, the choice of the next registration parameters informed by the performance of the previous parameters, which are evaluated for all discs by calling the registration pipeline and treating the result of the objective function (Equation 1) as an observation with loss value L.

$$L = \frac{1}{N} \sum_{i=1}^{N} \left(1 - DSC_i + \sigma(\mathbf{J})_i\right)$$

Equation 7.1 *N* is the total number of discs, *DSC* is the Dice overlap coefficient, and $\sigma(\mathbf{J})$ is the standard deviation of the determinant of the spatial Jacobian.

The determinant of the Jacobian of the deformation field is a pixelwise description of volume changes: expansion ($\mathbf{J}>1$), compression ($0<\mathbf{J}<1$), folding ($\mathbf{J}<0$), or constant volume ($\mathbf{J}=1$). Statistics computed across all pixels in the original disc space quantitatively describe the effect of registration. Per disc, Jacobian determinant values are centered around 1, with the standard deviation describing the severity of local expansion and compression. Evaluations of the objective function guide the Bayesian optimizer, maximizing Dice overlap between the inferred disc mask and the atlas mask, while minimizing the standard deviation of the determinant of the Jacobian across all registered discs for all subjects to find the optimal registration parameters. All resulting deformation fields and relaxation maps were checked to ensure local topology and distribution of relaxation values were preserved after registration.

7.3.4 Statistical Analysis

Four types of voxelwise statistics are performed on the registered $T_{1\rho}$ and T_2 maps from each study. Only voxels meeting threshold criteria (Dataset A $T_{1\rho}$ < 250ms, Dataset B $T_{1\rho}$ < 200ms, T_2 < 150ms) are included. Missing data at the patient level (ex. missing questionnaire) or at the disc level (ex. missing Pfirrmann data) excludes patient maps from analysis concerning those variables. Voxel by variable statistics examine local associations between relaxation maps and measured outcomes with Pearson correlation or partial correlation with adjustments for age, gender, BMI, and group assignment when relevant. Correlation coefficient maps and p-

value maps are visualized. Voxel-by-group statistics compare average relaxation maps of subjects grouped by demographic or clinical variables (for example high disability, low disability) or by group assignment, unpaired t-test checking for significant differences between groups. The average map for each group and p-value map is visualized. Voxel-by-voxel statistics calculate within subject voxel-wise correlation between two relaxation maps (ex. $T_{1\rho}$ T_2). These values are compared with Pearson correlation; correlation coefficient maps and p-value maps are visualized. Voxel by time statistics are primarily for testing longitudinal changes in relaxation maps within subjects. Given the low number of follow-ups, statistical differences between baseline and follow-up cannot be computed. In larger studies, longitudinal difference maps would be tested for association with changing clinical outcomes or for differences between groups where baseline, follow-up, and difference relaxation maps are visualized. Correlation results for Dataset A are visualized as a single slice on a spine mesh while correlation results for 3D volumetric data are visualized as two central slices on a spine mesh. All post-processing and statistical tests were performed using Pingouin(0.2.6), Scipy(1.2.0), StatsModels(0.9.0) using Python 3.6, with $\alpha<0.05$.

7.4 Results

7.4.1 Segmentation Performance

The datasets used for method development are similarly distributed in age, BMI, and height. There exist significant differences in gender ratios, proportion of LBP patients, Oswestry Disability Index (ODI) scores, and degenerative grades (Pfirrmann 1: 12%, 2: 54%, 2.5: 8%, 3: 24%, 3.5: 1%, 4: 1% vs. Modified Pfirrmann 1: 8%, 2: 36%, 3: 23%, 4: 3%, 5: 1%, 6: 14%, 7: 10%, 8: 5%). When Dataset A and B are combined, the final dataset evenly samples the spectrum of morphologic and symptomatic IVDD.

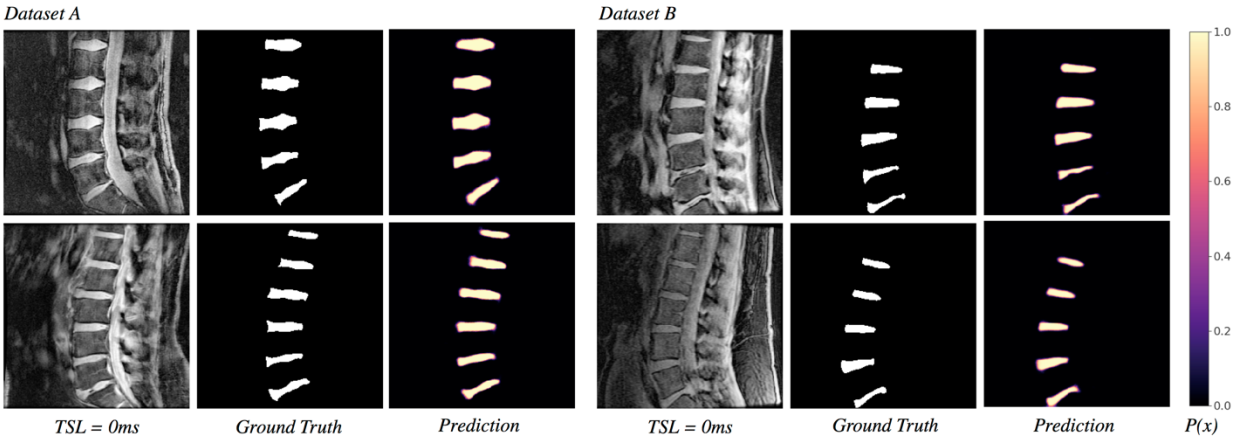


Figure 7.3 Input qMRI slice, ground truth mask, and predicted segmentation probabilities for 4 test subjects, 2 from Dataset A (Left) and 2 from Dataset B (Right). The 2D Coarse-to-Fine-Context-Memory (CFCM) segmentation network demonstrates consistent performance in both $T_{1\rho}$ acquisition sequences, across grades of disc degeneration, and spinal morphology without off-target segmentation predictions. Probabilities are thresholded at 0.5 to create binary masks. TSL = Time Spin Lock

Across datasets and degenerative grades, the CFCM networks produced accurate, high confidence segmentations of the lumbar discs. Representative segmentations before thresholding probabilities at 0.5 are shown in Figure 7.3. Predicted probability maps show the network highest uncertainty along disc boundaries, particularly at the anterior and posterior annulus-ligament interface. Per slice, automatic segmentation of all discs took 0.393 seconds, over 1000 times faster than manual segmentation.

Evaluated using segmentation metrics (Table 7.2), the network produced segmentations with Dice overlap (DSC) consistently above 0.85 and mean absolute surface distance (MSD) less than 1 pixel at all levels, approaching the limit of image resolution. As a metric, DSC is sensitive to the size of the ground truth structure, as a single pixel error will disproportionately lower DSC for a small disc compared to a large disc.

Table 7.2 Dice Similarity score (DSC), Mean Absolute Surface Distance (MSD) at disc boundary, % volume difference, sensitivity, and precision results per disc for each dataset, with 95% confidence intervals in parentheses.

	L5S1	*L4L5*	*L3L4*	*L2L3*	*L1L2*
Dataset A					
DSC	0.871 (0.848, 0.893)	0.916 (0.907, 0.926)	0.932 (0.923, 0.941)	0.915 (0.903, 0.928)	0.883 (0.852, 0.913)
MSD	1.37 (0.976, 1.76)	0.836 (0.748, 0.923)	0.683 (0.596, 0.769)	0.820 (0.687, 0.953)	0.972 (0.716, 0.228)
% VD	2.60 (-6.03, 11.24)	-1.91 (-7.23, 3.40)	0.51 (-3.10, 4.13)	1.19 (-3.18, 5.57)	-2.28 (-11.7, 7.19)
Sens	0.858 (0.824, 0.893)	0.925 (0.902, 0.949)	0.930 (0.910, 0.950)	0.911 (0.883, 0.938)	0.894 (0.843, 0.945)
Prec	0.896 (0.85, 0.942)	0.913 (0.885, 0.941)	0.937 (0.918, 0.956)	0.924 (0.906, 0.943)	0.886 (0.841, 0.931)
Dataset B					
DSC	0.877 (0.858, 0.897)	0.895 (0.877, 0.914)	0.913 (0.898, 0.928)	0.899 (0.875, 0.922)	0.901 (0.890, 0.912)
MSD	0.300 (0.205, 0.395)	0.265 (0.176, 0.354)	0.186 (0.153, 0.219)	0.215 (0.170, 0.260)	0.215 (0.173, 0.258)
% VD	-1.70 (-8.04, 4.64)	-2.95 (-9.87, 3.95)	-5.25 (-10.5, -0.00)	-4.78 (-10.4, 0.86)	-2.71 (-7.81, 2.37)
Sens	0.886 (0.85, 0.921)	0.909 (0.875, 0.943)	0.937 (0.917, 0.957)	0.92 (0.891, 0.949)	0.914 (0.889, 0.939)
Prec	0.876 (0.846, 0.907)	0.89 (0.858, 0.923)	0.896 (0.866, 0.926)	0.884 (0.85, 0.918)	0.894 (0.869, 0.919)

The lowest performing disc segmentation was L5S1, which is the smallest, most likely to be degenerated, and most challenging to manually segment. Highest performing disc was L3L4, which was usually the largest and always centered in the FOV. Volume difference (%VD), sensitivity (Sens), and precision (Prec) between ground truth and network segmentations revealed the networks were biased towards moderate overestimation of disc volume in Dataset B (greater number of False Positive voxels), while Dataset A had slight over and underestimation depending on the disc level. Comparing segmentation metrics against radiological grades of degeneration, the networks showed lower DSC performance in more degenerated discs, while MSD and %VD were invariant to degenerative grade suggesting lowered performance could be a result of the metric itself. Pooled lumbar spine metrics (n=88, n=92) were 0.904, 0.898 DSC; 0.936, 0.236 MSD; +0.07, -3.52 %VD; 0.904, 0.913 Sens; 0.912, 0.888 Prec respectively.

7.4.2 Relaxation Time Extraction

Manually and automatically extracted mean disc $T_{1\rho}$ and T_2 relaxation times show strong, significant correlations at all disc levels (Table 7.3). All disc correlations for Dataset A $T_{1\rho}$ r =

0.995, p=7.4e-84, bias = -0.74ms (-4.35, 2.87), Dataset B $T_{1\rho}$ r = 0.990, p=4.1e-78, bias = -0.01ms (-4.36, 4.33), and T_2 r = 0.984, p=2.5e-70, bias = 0.12ms (-3.55, 3.79), with no trends evident in difference plots (Figure 7.4). Dataset A showed more precise biomarker extraction with a slight bias towards overestimating relaxation times, particularly in L5S1. Dataset B had less precise $T_{1\rho}$ biomarker extraction (as observed with the wider confidence intervals) but produced unbiased estimates of relaxation time. Correlations between manual and automatic $T_{1\rho}$ times in Dataset B were stronger than correlations between T_2 times, in all discs except L5S1. $T_{1\rho}$ and T_2 biomarker extraction accuracy did not change with increasing degenerative grade and remained within 5ms of manually extracted values in all but two discs.

Table 7.3 Comparison of manually and automatically extracted relaxation values with Pearson correlation coefficient r_{coeff}, and bias measurement per disc for each dataset, p-values and 95% confidence intervals in parenthesis respectively.

	L5S1	*L4L5*	*L3L4*	*L2L3*	*L1L2*
Dataset A					
T1ρ r_{coeff}	0.987 (p=3.0e-13)	0.998 (p=3.6e-20)	0.997 (p=1.4e-20)	0.998 (p=4.5e-20)	0.995 (p=1.1e-14)
T1ρ bias (ms)	-2.0 (-8.16, 4.07)	-0.3 (-2.44, 1.74)	-0.5 (-2.33, 1.21)	0.0 (-2.12, 1.99)	-0.6 (-3.72, 2.34)
Dataset B					
T1ρ r_{coeff}	0.969 (p=4.3e-11)	0.991 (p=3.7e-16)	0.990 (p=5.1e-16)	0.993 (p=3.5e-17)	0.988 (p=2.6e-14)
T1ρ bias (ms)	-0.4 (-4.92, 4.01)	0.03 (-4.50, 4.56)	0.0 (-4.41, 4.40)	0.41 (-3.31, 4.15)	0.06 (-4.83, 4.96)
T2 r_{coeff}	0.975 (p=7.1e-12)	0.984 (p=3.8e-14)	0.983 (p=6.5e-14)	0.991 (p=4.0e-16)	0.979 (p=1.6e-12)
T2 bias (ms)	-0.2 (-3.63, 3.23)	-0.1 (3.86, 3.49)	0.18 (-3.55, 3.92)	0.68 (-2.48, 3.84)	0.10 (-4.18, 4.39)

Qualitatively, the two-step registration approach successfully morphed the lumbar discs into the atlas space preserving the spatial distribution of relaxation times in the nucleus and annulus as well as the total distribution of intensity values across the disc, the effect of registration is visualized in Figure 7.5. Performance was consistent across degenerative grades. Histogram plots of disc intensities show good agreement between the values before and after registration, indicating deformations were applied smoothly throughout the disc and disc regions

are represented fairly. Disc boundaries, particularly the anterior and posterior disc-ligament interface, showed the most variability in registration accuracy.

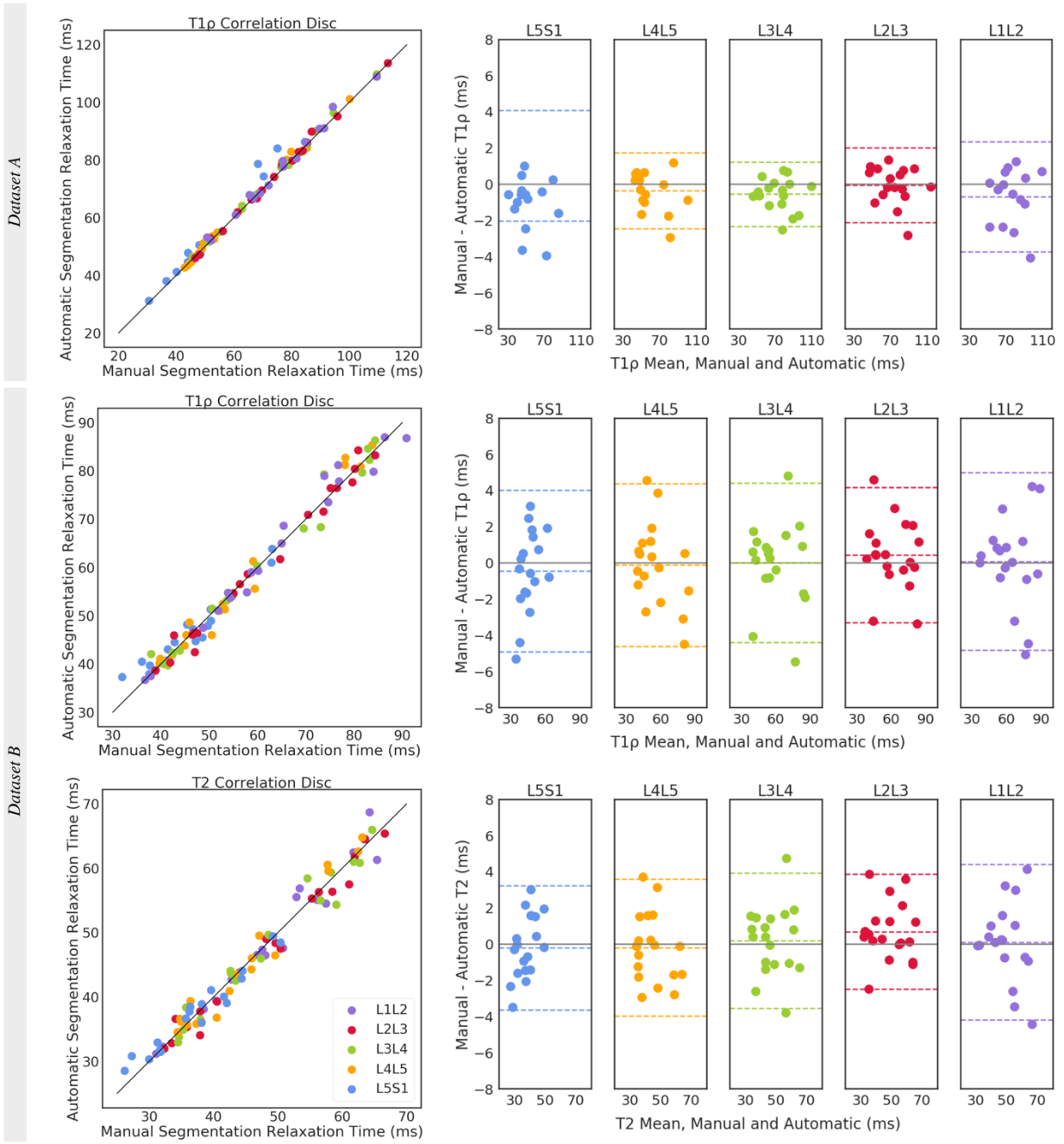


Figure 7.4 Correlation scatterplot for all discs and Bland-Altman plots with the 95% limits of agreement (LOA) for each disc level for comparison of manually and automatically extracted $T_{1\rho}$ and T_2 relaxation times. In Dataset A's L5S1 $T_{1\rho}$ Bland-Altman plot, the lower LOA at -8.16ms was omitted to maintain the same y-axis range between plots.

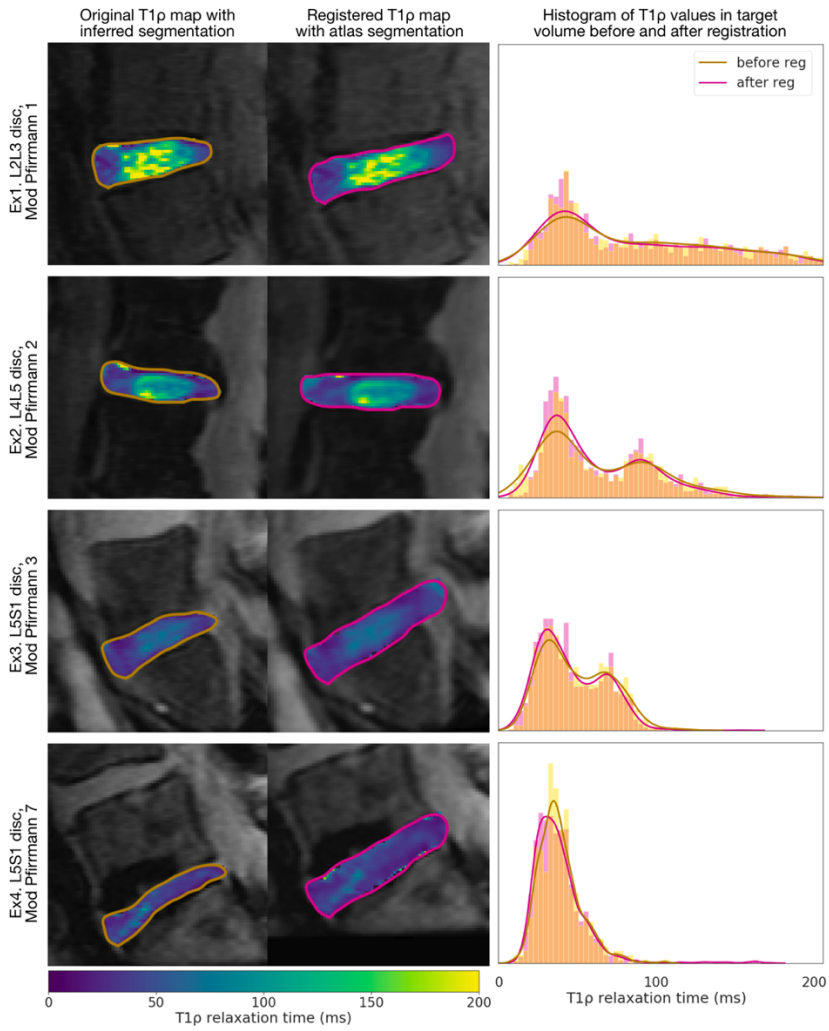


Figure 7.5 Example $T_{1\rho}$ maps before and after registration and distribution of $T_{1\rho}$ relaxation values within the segmented disc before and after registration, from 4 test subjects. The registration process preserves spatial patterns in relaxation maps, with consistent performance across degenerative grades. Histograms showing the density and a gaussian kernel density estimate of $T_{1\rho}$ relaxation times within the segmented disc before and after registration.

7.4.3 Statistical Parametric Maps

Example statistical parametric maps are visualized in Figures 7.6, 7.7, 7.8. Local patterns in relaxation time maps show significant associations with radiological grading, as well as clinical measures of disability. Imaging-based Pfirrmann/Modified Pfirrmann degenerative grades were strongly negatively correlated with $T_{1\rho}$ maps in both datasets and with T_2 maps in Dataset B. Significant correlations are localized to lumbar disc nucleus and inner annulus, with correlation strength and significance increasing in Dataset A's lower disc levels (L3-4 through

L5S1) while associations remain consistent across disc levels in Dataset B. T_2 correlations were similar, yet not identically distributed to $T_{1\rho}$ correlations, with $T_{1\rho}$ showing stronger and more significant correlations around the superior and inferior portion of the nucleus.

Compared correlations with $T_{1\rho}$ values estimated from whole-disc ROI approach, Dataset A (Pearson r L1L2: -0.15 p=0.60, L2L3: -0.458 p=0.07, L3L4: -0.516 p=0.04, L4L5: -0.741 p=0.001, L5S1: -0.772 p=0.0005), Dataset B (L1L2: -0.670 p=0.003, L2L3: -0.660 p=0.003, L3L4: -0.715 p=0.0008, L4L5: -0.732 p=0.0008, L5S1: -0.671 p=0.003), the proposed voxel-based method confirms ROI associations and recovers significant associations in disc subregions not identifiable with ROI methods. For example, in Dataset A's L2L3 disc, Pfirrmann grades are weakly and non-significantly correlated with mean whole-disc $T_{1\rho}$ values, yet the voxel-based method reveals a moderate, positive correlation in the inferior region of the nucleus.

Interestingly, results with respect to disability measures varied between the two datasets (Figure 7.7). Dataset A shows strong, negative correlations between $T_{1\rho}$ and Oswestry Disability Index (ODI) scores while Dataset B shows weak, positive correlations between the two particularly in the nucleus-annulus transition region of L4L5. Dataset B's T_2 correlation maps mostly mirrored those of $T_{1\rho}$ (not shown), however, positive correlations seen in L4L5 were stronger and had significant voxels clustered in the posterior inner annulus. The trends in Dataset A and B appear opposite, however the association between ODI and $T_{1\rho}$ is only consistent in Dataset A, where negative correlations are stronger and present across multiple lumbar disc levels, with the exception of L5S1 where no relationship is evident. Again, these trends support whole-disc $T_{1\rho}$ findings, with the advantage that the voxel-based method can recover the anatomical location of significant associations: Dataset A (Pearson r L1L2: -0.800 p=0.001 , L2L3: -0.650 p=0.016, L3L4: -0.620 p=0.024, L4L5: -0.674 p=0.011, L5S1: -0.231 p=0.45), Dataset B (L1L2: 0.152 p=0.57, L2L3: 0.224 p=0.39, L3L4: 0.415 p=0.098, L4L5: 0.574 p=0.02, L5S1: 0.314 p=0.24). Further stratifying patients in Dataset A by ODI –

minimal/moderate vs severe disability– and performing a group comparison shows significant differences between relaxation maps. Average $T_{1\rho}$ relaxation maps from the minimal/moderate disability group showed clear annulus-nucleus distinction with a visible midline while the severe disability group had lower average $T_{1\rho}$ values with a homogeneous distribution. Relative to other discs, low and high disability groups in both datasets had low mean relaxation values for the L5S1 disc.

Finally, $T_{1\rho}$ maps and T_2 maps were highly and significantly positively correlated, as observed with whole-disc ROI analysis (Pearson r from 0.954 to 0.989 with p<1e-10) (Figure 7.8). However, voxel-by-voxel analysis suggests correlation strength between $T_{1\rho}$ and T_2 relaxation values is localized: the anterior annulus and near endplate regions show weaker correlations than the rest of the disc space, for all lumbar discs. Additionally, correlation values in center of the disc are heterogeneous, suggesting the relationship between $T_{1\rho}$ and T_2 values fluctuates throughout the disc.

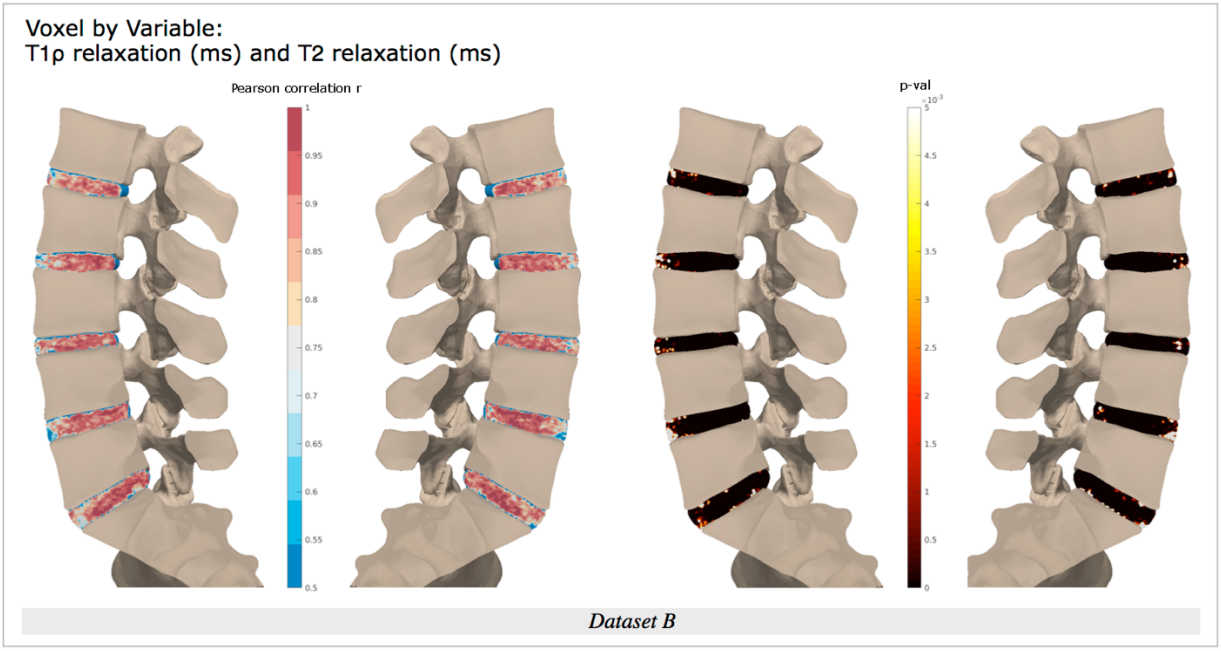


Figure 7.6 Voxelwise associations in Dataset B between $T_{1\rho}$ and T_2 values, with Pearson correlation r map displayed from 0.5 to 1, and p-value map from p=0 to p=0.005.

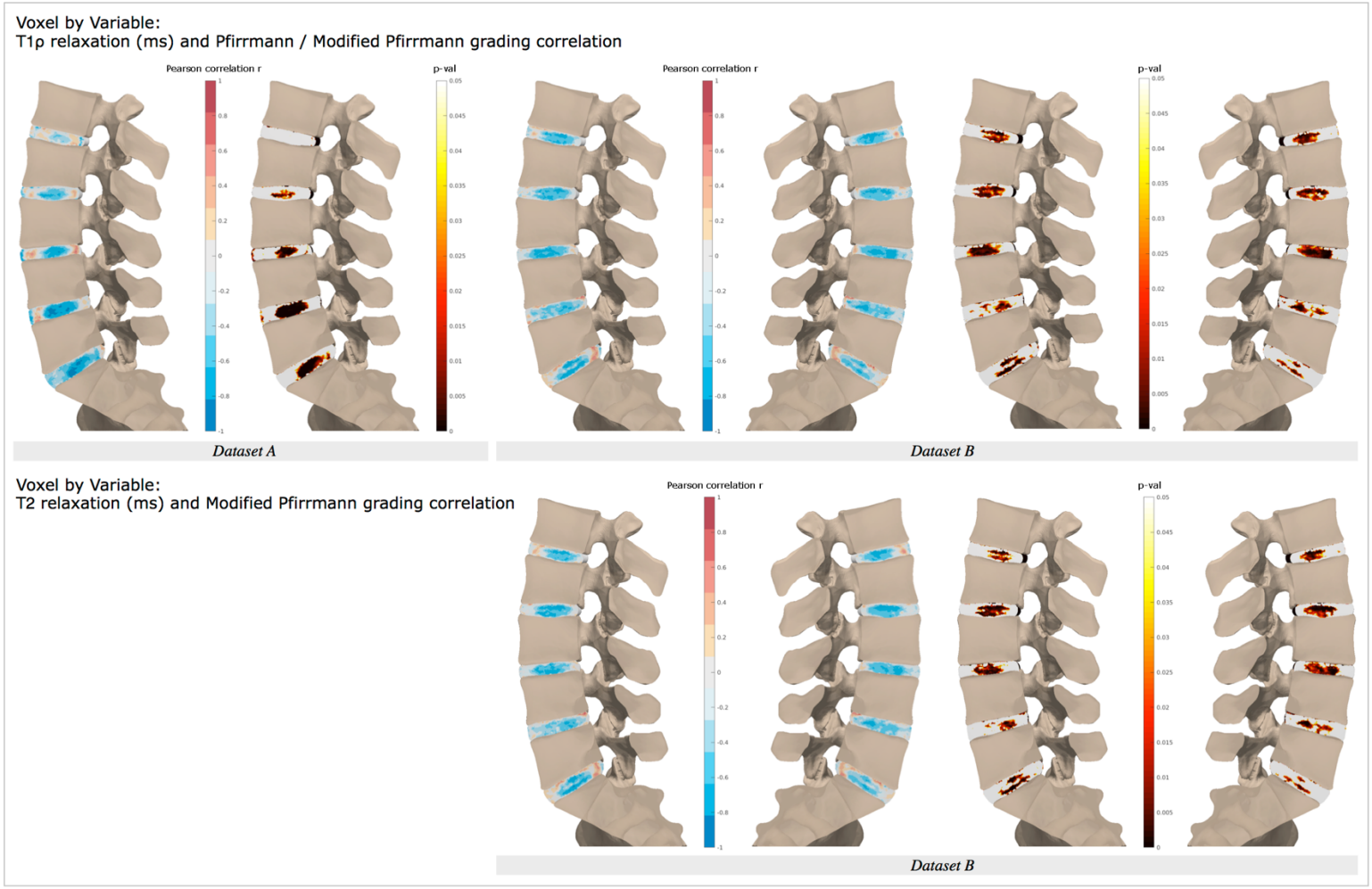


Figure 7.7 Voxelwise associations between $T_{1\rho}$, T_2 maps and Pfirrmann/Modified Pfirrmann grading. Single slice Pearson correlation with 5 point Pfirrmann grade for Dataset A (Left) and two central slice correlation with 8 point Modified Pfirrmann grade Dataset B (Right), each shown with corresponding p-value map.

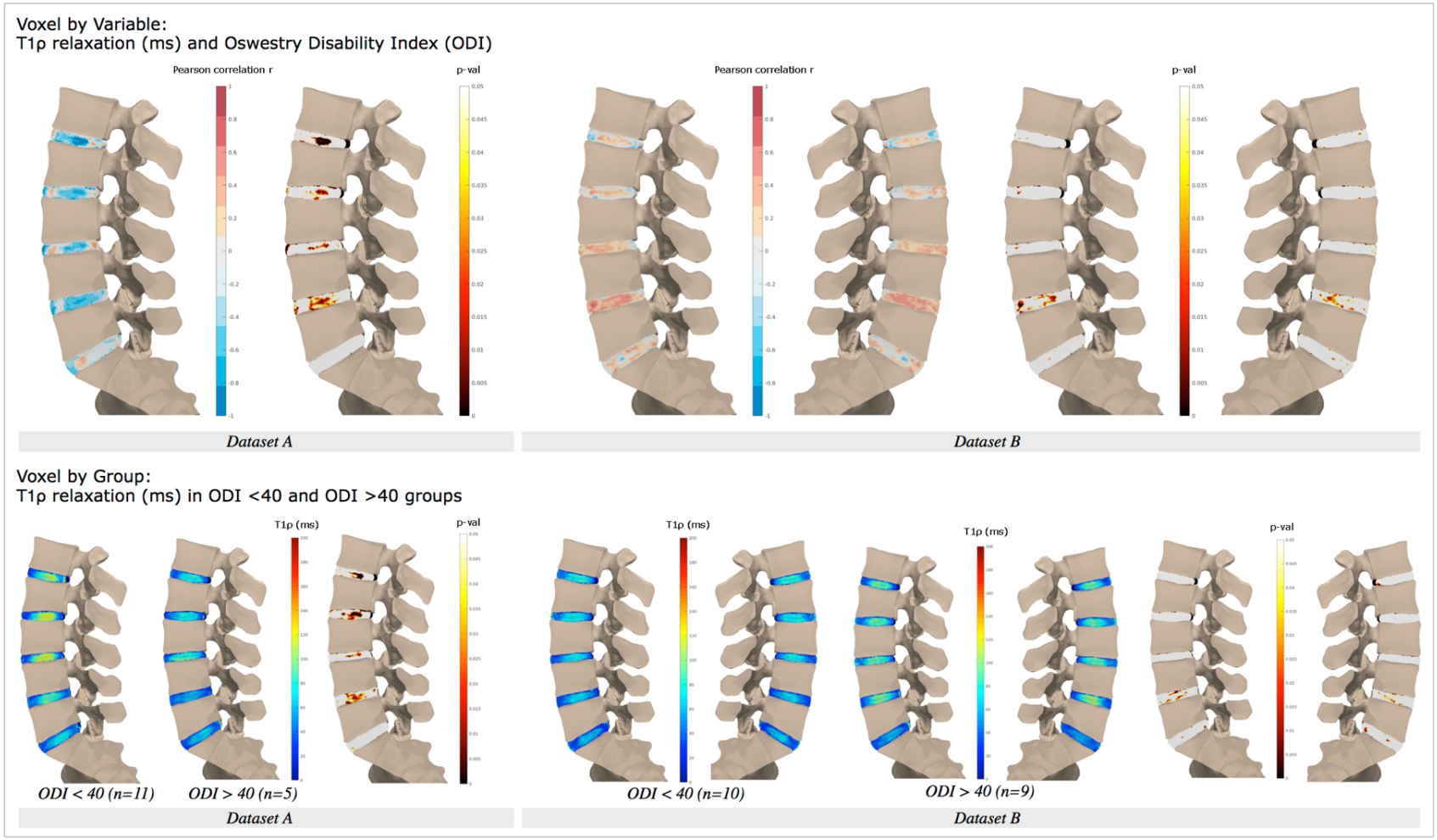


Figure 7.8 Top) Voxelwise associations between $T_{1\rho}$ maps and Oswestry Disability Index scores. (Bottom) Voxel by group analysis, mean relaxation $T_{1\rho}$ maps and p-values for minimal/moderate (0-40 ODI score) and severe disability (>40 ODI score) groups.

7.5 Discussion

We have demonstrated the novel pipeline proposed for qMRI analysis of intervertebral discs is feasible and addresses the limitations of conventional analysis methods. The intervertebral disc is a challenging tissue to analyze due to its deformable structure, lack of anatomical landmarks, and variations in image intensity. By integrating a CNN for segmentation into our atlas-based registration pipeline we developed a fast, robust, and scalable solution to analyze local patterns in intervertebral disc qMRI.

Merging of Dataset A and B was necessary for the development of a robust analysis pipeline. Together, the datasets sample the full morphologic and symptomatic IVDD spectrum, including a range of degenerative grades and patient reported outcomes such as pain and disability. Similarity in $T_{1\rho}$ image contrast and image prescription enabled the merging of these datasets for 2D segmentation method development and validation. However, differences in acquisition parameters (spin-lock pulse duration, spin-lock frequency, and voxel size) prevented joint registration and statistical analysis of relaxation maps. Given the limited sample size of each dataset, the appearance of common trends in $T_{1\rho}$ correlation maps demonstrated face validity of our analysis pipeline. Trends in T_2 correlation maps were similar but not identical to $T_{1\rho}$ maps, demonstrating the feasibility of integrating multiple, potentially complementary qMRI sequences into the analysis pipeline.

Our training strategy helped the CFCM segmentation network learn to reliably segment image slices even with limited training data. The network was prevented from overfitting by aggressively augmenting every training iteration, using large batch sizes with batch normalization, no hyperparameter tuning, and creating data-splits by subject. Both contrast and geometric augmentations were chosen to introduce diversity into the images, as disc shape, position, intensity, and texture vary widely between datasets. The choice of network was key in achieving high performance: the memory mechanism in the CFCM learns how to best fuse features to combine local and global context, and had fewer spurious segmentation predictions

compared to U-Net[238]. State-of-the-art disc segmentation performance in the T_2 2015 challenge and the IVD3M 2018 challenge are Dice scores of 0.918 and 0.907 respectively. We believe the performance of our segmentation network is competitive with and more generalizable than other published disc segmentation methods, although a direct comparison is not possible due to differences in datasets and annotation methods. Datasets provided by disc segmentation challenges[234; 235] have more training data (576 segmented slices from 8 subjects in IVD3M) but are only trained on images from volunteers with healthy discs.

There are several reasons segmentation performance in our dataset decreased in discs with severe IVDD. First, manual segmentations are less reliable; loss of nucleus glycosaminoglycans, annular collagen, and dehydration lead to a decrease in disc volume. In turn, these changes are reflected in shorter tissue relaxation times and lower disc signal intensity on the $T_{1\rho}$ weighted images, obscuring the boundary at the interface of the annulus and spinal ligaments, which compounds with partial volume artifacts on edge slices. Second, there are fewer training examples of severely degenerated discs and many degenerative phenotypes exist. Healthy discs are often surrounded by normal presenting anatomy, while severe IVDD is associated with fattier vertebral bone marrow, narrowed spinal canal, and even signal voids due to the vacuum phenomena[240]. Finally, Dice coefficient and % volume difference are sensitive to segmented tissue size, and given the smaller disc volume in severe IVDD, single pixel errors disproportionately impacted these results.

Our results demonstrate that errors in disc segmentation were not propagated to errors in biomarker extraction. Segmentation is performed on the first echo of the mapping sequence and relaxation times are calculated from the monoexponential fit of all acquired echos. Errors by the segmentation network represent a small fraction of the total disc area thereby not significantly skewing the mean. Error pixels may also contain intensity values that do not have high enough SNR for monoexponential fitting or produce relaxation values outside of a feasible range, neither which are included in the calculation of mean relaxation times. Even in the worst

performing disc –L5S1 in Dataset A– mean disc $T_{1\rho}$ errors ranged from -8 to 4 ms. Calculation of mean disc relaxation times from automatic segmentations are an intermediate output to validate the segmentation portion of the pipeline. However, our automatic segmentation method is a viable alternative to ROI based analysis, as it is significantly faster (0.393s/slice compared to our 7.5min/slice manual, 12s/slice average for submissions to T_2 disc segmentation challenge[235]) and more reliable than manual segmentation.

The automatic segmentation network provides masks necessary to guide registration. Mutual information guided atlas-registration is successful in other tissues, but intensity-based methods fail to register intervertebral discs. We hypothesize this issue arises with cases of severe degeneration where intensity signals from normal presenting anatomy are absent. The disc mask allows for good initialization of the registration algorithm and calculation of overlap metrics for Bayesian optimization of registration parameters. Our proposed objective function is designed to maximize registration accuracy while preventing significant deformations which would perturb the local distribution of relaxation values. This highlights the flexibility of our proposed pipeline, with automatic parameter tuning for application to other datasets or different atlases.

Local distribution of relaxation values was preserved throughout the registration procedure even with alternate segmentation methods and atlas selection, thus demonstrating the success of the full analysis pipeline. Recent studies have recognized the limitations of coarse ROI methods and have attempted to address this problem with smaller ROIs increasing the time, complexity, and bias introduced. In a group of healthy discs, our method's average relaxation maps show distinctive regions corresponding to the annulus, nucleus, and disc midline; patterns recovered in a fully data-driven manner without introducing user bias. Additionally, our voxel-based method showed greater sensitivity to small, significant associations within the disc such as that were washed out with whole disc ROI averaging. Our proposed statistical parametric mapping methods still performs an averaging procedure, on a voxel scale. This will show

common trends within the studied group, but it is not ideal for the identification of focal lesions[241]. High intensity zones (HIZs), for example, would only be identified if they co-localized for a large group of patients. In voxel-by-group analysis, maximum and standard deviation maps could potentially identify these clusters.

Different physiological loading demands explain variations in geometry, biochemical and microstructural composition between disc levels. A level-specific atlas was used for registration, as it is inappropriate to pool relaxometry results from all discs. Relationships between relaxation values at different disc levels may reveal important associations with clinical outcomes. Additionally, the strength of the relationship between $T_{1\rho}$ and T_2 relaxation times was highly spatially dependent, indicating that each of these biomarkers may reveal differences in local biochemistry, which are observed in human disc specimens[25].

The limitations of this work are discussed in two parts: pipeline and dataset. As a pipeline, segmentation network training and registration optimization impose upfront computational and time costs. However, once these sections have been optimized to the target task, processing time is faster than manual ROI analysis methods. From the dataset side, one or two 8mm sagittal qMRI slices do not fully capture biochemical composition of the intervertebral disc. Both Dataset A and Dataset B reported that SNR prevented the acquisition of thinner slices, although recent developments may address this limitation.

Lastly, low sample size prevented meaningful interpretation of associations with patient reported outcomes. Associations between $T_{1\rho}$ and disability were strongly negative and significant in Dataset A yet were not visible in Dataset B, indicating the studies were underpowered. Similarly, associations with muscle data extracted from Dataset B did not reach statistical significance. A greater sample size is necessary to power proper statistical analysis adjusting for multiple comparisons and demographic/clinical confounders, and to enable feature extraction for IVDD characterization.

Image statistics can be defined in voxels clusters, or peaks. Voxel-wise inference examines if the t-statistic (or F-statistic) is within a predefined threshold at each voxel, to reject the null hypothesis at that voxel (high spatial specificity). Cluster-wise inference defines a t-statistic threshold and minimum cluster size, to reject the null hypothesis of the whole cluster, indicating activity is somewhere within the cluster (high sensitivity, low spatial specificity). Peak-wise inference identifies local maxima in t-statistics greater than a predefined threshold (high spatial specificity). To correct inferences for multiplicity, corrections on p-values with Familywise Error Rate (Bonferroni correction, Random Field Theory) and False Discovery Rate controlling procedures.

There are several promising applications of our analysis method. Broadly, the main motivation of this work was to develop an automatic pipeline for lumbar intervertebral disc characterization, creating a fast, reliable, and robust tool to aid mechanistic disease research of IVDD. Applied to a larger clinical imaging dataset, our approach could be used for LBP phenotyping: selecting patient cohorts for clinical trials, matching patients to effective treatments, or tracking treatment effects over time. ReSPINE, a randomized clinical trial for mesenchymal stem cell therapy for IVDD is underway in Europe, and qMRI will be acquired over 4 timepoints. Our proposed pipeline could provide automatic, reliable processing of qMRI to follow subtle changes in spine biochemistry through statistical parametric mapping. Lastly, there is value for researchers validating new quantitative pulse sequences or compressed sensing schemes, to reliably compare the voxel-based patterns extracted by both methods. Application to other registration tasks and datasets is straightforward given the flexibility of our method.

7.6 Conclusion

This work proposes a novel methodology that combines deep-learning based segmentation, atlas-based registration, and statistical parametric mapping for automatic

analysis of quantitative spine imaging, addressing current methods' issues with sensitivity, reliability, and scalability. Evaluation of the segmentation method demonstrates performance is robust and shows excellent agreement with manual methods of biomarker extraction across the spectrum of morphologic and symptomatic IVDD. Despite the limited data available for method development, the voxel-based relaxometry pipeline reveals local trends in disc qMRI values which were significantly associated with clinical measures of degeneration and disability in two independent datasets. Future research directions include the applying the proposed framework on larger spine qMRI datasets to investigate LBP phenotypes for pathophysiological research, clinical cohort selection, and treatment monitoring.

8 Institution-wide shape analysis of 3D spinal curvature and global alignment parameters

The following manuscript is under review in the Journal of Orthopaedic Research as a Research Article.

8.1 Abstract

The spine is an articulated, 3D structure with 6 degrees of translational and rotational freedom. Clinical studies have shown spinal deformities are associated with pain and functional disability in both adult and pediatric populations. Clinical decision making relies on accurate characterization of the spinal deformity and monitoring of its progression over time. However, Cobb angle measurements are time-consuming, are limited by inter-observer variability, and represent a simplified 2D view of a 3D structure. Instead, spine deformities can be described by 3D shape parameters, addressing the limitations of current measurement methods. To this end, we develop and validate a deep learning algorithm to automatically extract the vertebral midline (from the upper endplate of S1 to the lower endplate of C7) for frontal and lateral radiographs. Our results demonstrate robust performance across datasets and patient populations. Approximations of 3D spines are reconstructed from the unit normalized midline curves of 20,118 pairs of full spine radiographs belonging to 15,378 patients acquired at our institution between 2008 and 2020. The resulting 3D dataset is used to build a statistical shape model to describe global spine shape variations in pre-operative deformity patients via 8 interpretable

shape parameters. This approach allows for the characterization of longitudinal changes in 3D spine shape and –if deployed into an existing database– identification of patient subgroups with similar shape and demographic characteristics without relying on an existing shape classification system. Upon publication, interested readers are encouraged to upload anonymized full spine radiographs through grad.io to test the automatic spine midline extraction and shape mode quantification.

8.2 Introduction

Spinal deformities are deviations from the normative, 3D articulated structure of the spine. Discs, vertebrae, facet joints, spinal ligaments, and paraspinal musculature are key structural elements responsible for spine stability. Pathophysiological changes in tissue composition or neuromuscular regulation can threaten the mechanical integrity of the spine and lead to local and global instability[48; 242]. In turn, biomechanical instability recruits compensatory mechanisms[48] such as pelvic retroversion, which can exacerbate deformity progression through the re-distribution of load.

Accordingly, deformities are prevalent in populations undergoing rapid physiological change. In the pediatric population, scoliosis is the most common spinal deformity and is defined as curvature in the coronal plane >10˚. An estimated 0.47% to 5.2% of the pediatric population (<18 years of age) has idiopathic scoliosis[243], with prevalence increasing as patients go through peak growth velocity. Of all pediatric idiopathic scoliosis, infantile scoliosis (0-3 years) represents <5%, juvenile scoliosis (3-10 years) 10-15%, and adolescent (10-18 years) >80%. By contrast, adult spinal deformity (ASD) encompasses a heterogeneous group of conditions affecting the aging spine, including *de novo* and existing scoliosis, in addition to degenerative spinal conditions which can present concurrently. Although the prevalence of ASD as a group is not known, adult scoliosis is estimated to affect 8.3% of adults (>25 years) with prevalence sharply rising after age 50[244], and 68% of elderly patients (>60 years)[245].

Pain and functional disability are common concerns among patients with spinal deformities. Pain prevalence in adolescent idiopathic scoliosis (AIS) is 68%; pain intensity and functional disability are positively associated with curve magnitude[246]. In the ASD population, pain prevalence is nearly 90%; pain and health related quality of life (HRQoL) are strongly negatively correlated with magnitude of sagittal imbalance[79; 247; 248]. Moreover, in both populations, the rate of progression is closely linked to deformity severity[249; 250]. Treatments for AIS and ASD aim to slow or halt the progression of the deformity through conservative methods (ex. bracing, casting) or surgical intervention (ex. tethering, multi-level fusion).

Treatment planning relies on accurate assessments of the spinal deformity and careful monitoring of its progression over time. Lateral and frontal (Anterior-Posterior/Posterior-Anterior) 36 inch radiographs are the clinical standard for deformity evaluation. Cobb angles[80] and sagittal/coronal imbalance measurements are used to quantify deviations from normal spinal curvature, although several others have been proposed[251; 252]. Sagittal imbalance, also called sagittal vertical axis, is the horizontal distance between the posteriormost point of the S1 endplate and the vertebral center of C7 measured on lateral radiographs. Coronal imbalance, or coronal vertical axis, is measured as the horizontal distance between the center of the S1 endplate and the vertebral center of C7 on frontal radiographs. To measure Cobb angles, the user identifies the most tilted vertebra at the top and bottom of the spinal curve and draws a projection line from each using the frontal radiograph. The Cobb angle for the specific curve is the angle formed by the two intersecting lines. However, these assessments lack widespread clinical adoption as manual measurements are time-consuming and sensitive to intra and inter-observer variability[244 253]. These measurements help clinicians group patients by deformity type using 2D or 3D classification systems.

Over the last decade, significant research has been directed towards automating spine measurements. Automatic methods typically start with vertebral localization using a classic

computer vision algorithm[254] or a deep-learning based segmentation[255], object detection[256], or keypoint regression network [257; 258]. The outputs are then used to geometrically estimate Cobb angles. While several studies report radiologist-level accuracy and precision using their automated pipelines, of the surveyed literature, no studies conducted external clinical validation nor made their algorithms publicly testable. Additionally, fixation on replicating current Cobb measurements has prevented the application of these automatic methods for data-driven assessments of spine shape such as the clustering analysis presented by Thong et al[259].

The main goal of this study was to develop a fully automatic method for spine midline extraction on clinically standard full spine radiographs– applicable to pediatric and adult deformity populations –to approximate 3D spine shape and describe shape variations in our institution's patient population.

8.3 Methods

The automatic models were developed and tested on a subset of manually annotated images (hundreds), validated on a larger subset of images with labels extracted from radiology reports (thousands), and inferred on a retrospective institution-wide dataset (tens of thousands).

8.3.1 Keypoint Model Development

This research was approved by the Institutional Review Board (IRB305285). A random sample of 200 male and 200 female patients' full spine radiographs acquired between 2008 and 2018 were pulled from our institution's database. Four users were trained to annotate radiographs by placing keypoints on each vertebral corner from L5 to T1 (68 landmarks) and on the superior endplate of S1 and the inferior endplate of C7 (4 landmarks). Annotations were checked and corrected by two trainees (R1, R2) with 5 and 7 years of experience in radiological image analysis. In regions with poor visibility, such as the upper-thoracic region in lateral views, users were instructed to accurately identify landmarks on the inferior endplate of C7 and

interpolate the points in-between such that an anatomically standard number of thoracic vertebrae are identified. Trained users took more time to annotate lateral views compared to frontal views, suggesting that sagittal landmark detection would be a more challenging learning task than coronal detection, thus sagittal annotations were prioritized. A total of 194 coronal images and 366 sagittal images were annotated. Dicom images were read into Python, windowed, inverted (if necessary), 0-1 normalized, and zero-padded / cropped to a common FOV based on header information. Finally, images were resized to 1024x512, maintaining float32 precision and image aspect ratio throughout all processing steps. Annotated data did not include bending radiographs, images with spinal hardware, or partial spine views.

Data were split by patient into training, validation, and test (77%/8%/15% coronal, 69%/17%/14% sagittal). A 72 point landmark detection algorithm was developed for each view. All algorithms were implemented in PyTorch and used a convolutional neural network backbone with a differentiable layer for landmark predictions[260]. Image preprocessing (adaptive histogram normalization), augmentation severity, network backbone (Densenet-201[261], DilatedResNet-54[262]), batch size, dropout, weight decay, and initial learning rate were selected through a random hyperparameter search with 200 runs. The best performing hyperparameter combinations (Table 8.1) were selected based on the lowest validation mean squared errors.

Table 8.1 (Top) Hyperparameter settings for best performing algorithms on validation data. (Bottom) Hyperparameter settings investigated using random hyperparameter search with 200 runs.

		Sagittal View	Coronal View
data	**image dimensions**	(1024x512)	(1024x512)
	# image channels	3	3
	histogram equalization	FALSE	TRUE
	augmentation type	jitter, shift, contrast, gaussbright, gaussdark, zoom, cutout, rotate	jitter, shift, contrast, gaussbright, gaussdark, zoom, cutout, rotate
	augmentation prob	0.2, 0.15, 0.2, 0.1, 0.1, 0.3, 0.5, 0.4	0.2, 0.15, 0.2, 0.1, 0.1, 0.3, 0.5, 0.4
	batch size	11	12
learn	**loss**	L2 + heatmap regularization	L2 + heatmap regularization
	lr	1.08E-05	1.08E-05
	weight decay	0.00103	0.00715
	optimizer	Adam	Adam
	max epochs	500	500
	patience	75	75
	validation frequency	3	3
	epochs warmup	50	50
model params	**model**	DenseNet201 + DSNT layer	DenseNet201 + DSNT layer
	dropout	0.18	0.114
	# coordinates	72	72
pretraining	**type**	ImageNet	ImageNet

		Hyperparameter Search
data	**histogram equalization**	TRUE; FALSE
	augmentation type	jitter, shift, contrast, gaussbright, gaussdark, circle, zoom, cutout, rotate, noise
	augmentation prob	Low aug [0.2, 0.15, 0.2, 0.1, 0.1, 0, 0.3, 0.5, 0.4, 0]; Med aug [0.4, 0.15, 0.35, 0.1, 0.1, 0.1, 0.3, 0.3, 0.25, 0.15]; High aug [0.7, 0.2, 0.35, 0.3, 0.3, 0.15, 0.5, 0.5, 0.4, 0.35]
	batch size	6 to 13
learn	**lr**	5e-07 to 0.001 logspace
	weight decay	0.00001 to 0.05 linspace
model params	**model**	DenseNet201 + DSNT; DilatedResNet54 + DSNT
	dropout	0 to 0.4

8.3.2 Keypoint Model Testing

Test performance was assessed between ground truth landmarks and predicted landmarks with pointwise mean absolute error (MAE), imbalance mean absolute difference (MAD), and imbalance concordance correlation coefficient (CCC). CCC was selected over Pearson Correlation Coefficient as it measures bias as well as correlation between two variables. Coronal imbalance was calculated as the x-axis difference between the midpoint of S1's superior endplate and the midpoint of C7's inferior endplate. Sagittal imbalance was estimated as the x-axis difference between the posterior point of S1's superior endplate and the 2/3rd point

of C7's inferior endplate. To assess inter-reader variability and algorithm performance in a clinical scenario, R1 and R2 independently measured imbalance in a small, independent set of images using tools available on a PACS workstation. MAD and CCC were used to assess agreement across measurements.

8.3.3 Quality Control, Midline Extraction, and 3D Reconstruction

Per view, points along each side of the spine were fit using an polynomial degree 8, following an approach similar to the one proposed by Bonnani et al.[263]. Automatic quality control (autoQC) consisted of two tests: (1) polynomial fitting errors are below a predefined threshold and (2) predicted landmark order follows anatomical sequence, for example L3 vertebrae landmarks should be positioned above L5 landmarks. The polynomial fitting threshold was set to 0.01, which was selected empirically by examining 100 predictions and identifying a cutoff with specific to low-quality predictions. The autoQC step was included as a safeguard to detect predictions from out of distribution inputs. Finally, per view, the vertebral midline curve was extracted by averaging points from each side and fitting a polynomial through the vertebra midpoints. Midlines and contours were overlaid onto input images for visualization. For 3D reconstruction, the S1 midpoint was defined as the origin and 0-1 normalization of the z-axis was used to scale the S1 to C7 distance between views, resolving slight differences in magnification. Coronal and sagittal midlines were each sampled with 1000 points, combined using a common z-axis, and isotropically normalized. Due to the lack of calibration objects in the field of view, three major assumptions were used to accomplish the reconstruction: patient posture did not change between acquisitions, intrinsic parameters of the x-ray source were identical for both acquisitions, and acquisition planes were orthogonal to one another. Plots with sagittal, coronal, and axial projections were used to visualize the resulting 3D curve (Figure 8.1).

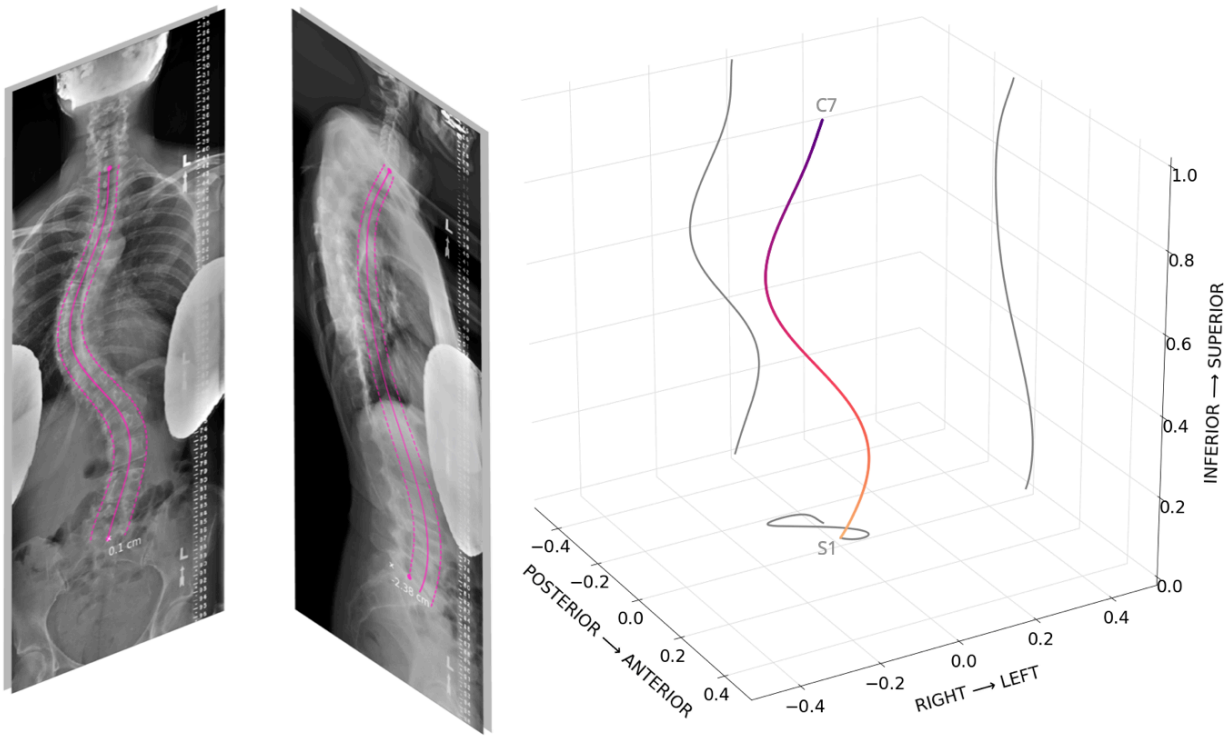


Figure 8.1 [Left] Input images, with predicted spine contours shown in magenta. White markers indicate x-axis coordinates of the C7 plumb line, text annotations show imbalance measurements in cm. [Right] Approximate 3D reconstruction of spine shape using midlines extracted from AP and LL views, from the upper endplate of S1 (light orange) to the lower endplate of C7 (dark purple). Coronal/sagittal/axial shadow projections are shown on each plane, axes are scaled isotropically.

8.3.4 Institution-wide Validation

To further test algorithm validity and generalizability, predicted imbalance measurements were compared to measurements mined from radiology reports. Radiology reports were parsed with a simple regular expression tool to extract imbalance measurements in centimeters. Predicted results were visualized as scatterplots and error histograms, agreement was assessed using MAD and CCC. A final qualitative check of algorithm generalizability was performed by running inference on images from the 2019 Accurate Automated Spinal Curve Estimation (AASCE) challenge test set[264] and examining the vertebral overlays (Figure 8.2).

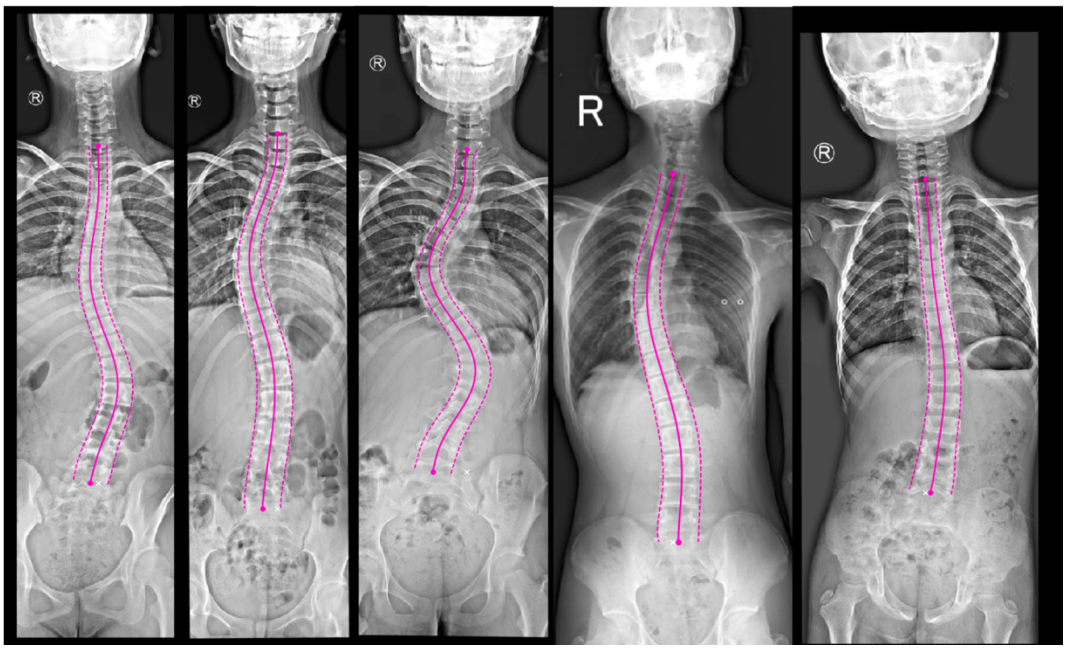


Figure 8.2 Example midline and contour predictions on JPEG data from the AASCE 2019 challenge. One low visibility image (not shown) failed quality control.

8.3.5 Retrospective Institution-wide Inference

Musculoskeletal radiologists and neuroradiologists compiled an exhaustive list of 28 radiology exam codes to identify relevant patient accessions between 2008 to September of 2020. All associated Dicom images and reports were anonymized. Data filtering steps are detailed in a flowchart (Figure 8.3). First, accessions with radiology reports mentioning (‘hardware’, ‘fusion’, ‘rods’, ’screws’) were removed. Then, Dicom headers missing view or pixel information were excluded. The remaining 20788 sagittal and 22893 coronal images were preprocessed identically to the model development images, then run through the landmark detection and midline curve extraction algorithm. Approximately 6.6% coronal images and 4.5% sagittal images failed autoQC; failed images were primarily mislabeled views, patients with spinal hardware not mentioned in the report, and bending radiographs.

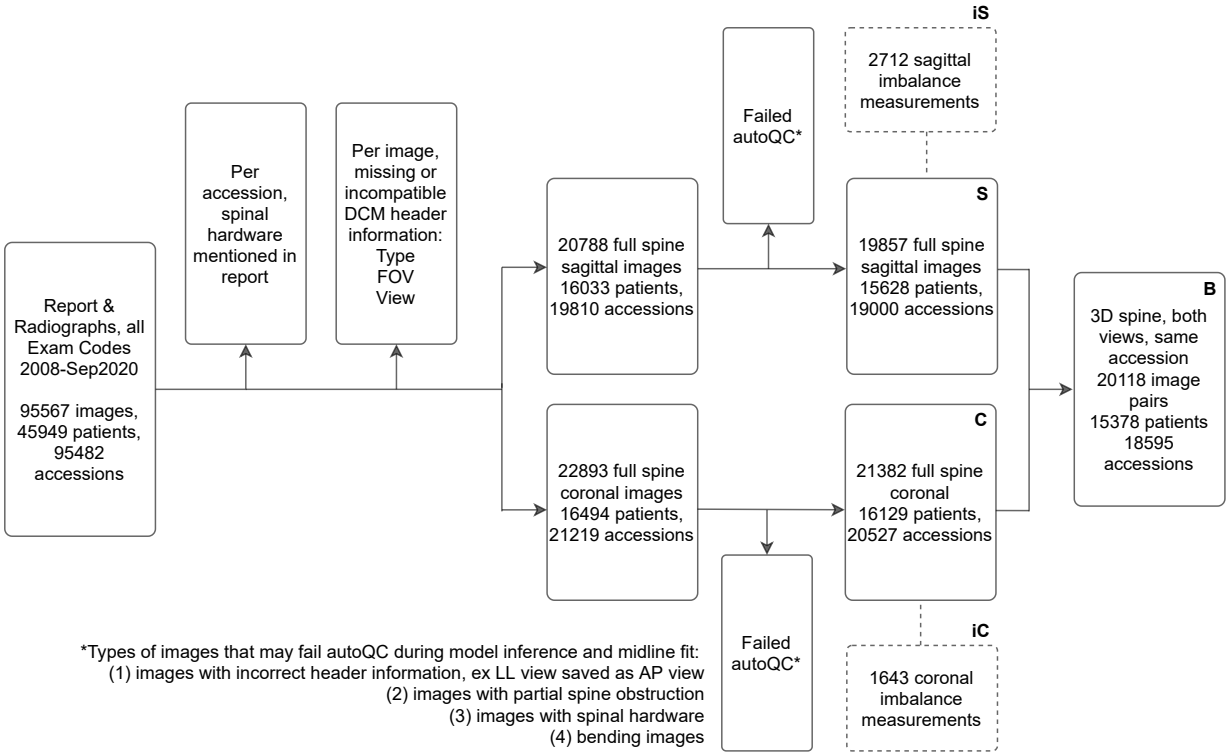


Figure 8.3 Data selection pipeline for institution-wide validation and deployment.

8.3.6 Shape Modeling

20118 3D spines described by 2000 anatomically corresponding points were to construct a statistical shape model. Features were centered before using Singular Value Decomposition based Principal Component Analysis to project the data to a lower dimensional, linear subspace. This resulted in 8 new shape axes (modes) describing shape variability within the patient population. In other words, the curvature of each 3D spine is described by 8 numbers, each describing specific shape characteristics. The average spine shape is visualized alongside -3 to 3 standard deviations of each shape mode to interpret shape characteristics. T-distributed stochastic neighborhood embedding (t-SNE) was used to visualize the distribution of patient spine shapes by creating a nonlinear embedding of the 8 dimensional shape vector into a two dimensional subspace. Twelve patients with scoliosis were randomly selected for Cobb angle evaluation: measurements from two trainees and one radiologist were averaged (R1, R2, radiology report) and plotted alongside the t-SNE datapoints.

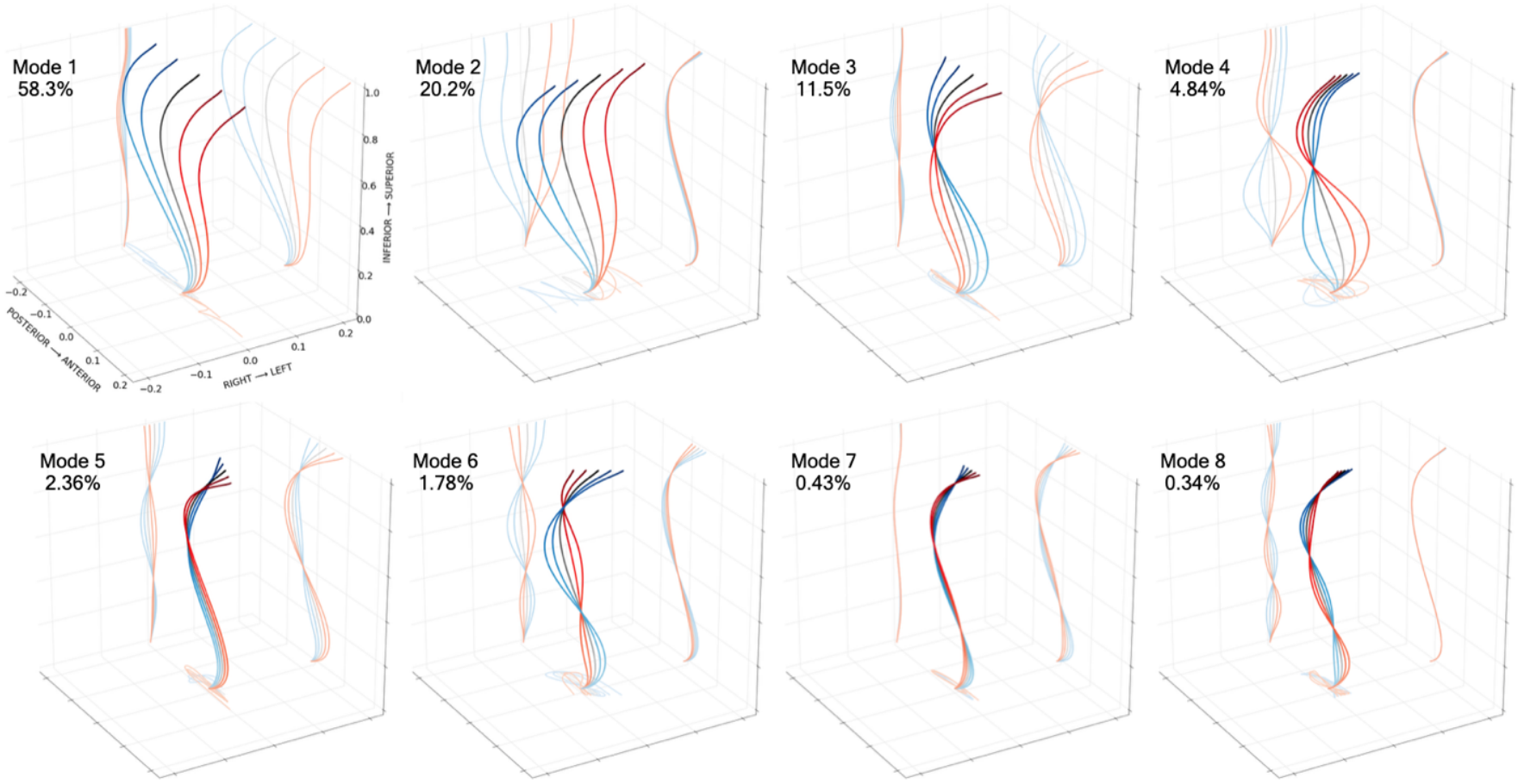


Figure 8.4 Shape mode interpretability plot, with percent variability explained by each shape mode. Average spine shape is plotted in grey. Lowermost curve point (lighter shading) is center of the upper S1 endplate, while the uppermost curve point (darker shading) is the center of the lower C7 endplate. Positive values for 1.5 and 3 standard deviations of each shape mode are plotted in red, negative values for 1.5 and 3 standard deviations of each shape mode are plotted in blue. Sagittal, coronal, and axial projections are plotted on each plane. All plots have the same axes and are scaled anisotropically (x : y : z, 0.5 : 0.5 : 1) to improve visibility.

8.3.7 Hosted model

For interested readers, the spinal landmark algorithm and radiology report parser will be hosted on gradio[265], pending approval by the Data Security Team at UCSF. Anonymized lateral and frontal (AP) Dicom image pairs are required to extract spine contours, a 3D spine plot and shape parameters.

8.4 Results

Dataset information for model development and institution-wide deployment is detailed in Table 8.2. From 2008 to early 2020, acquisition of spine radiographs has been growing at 19% per year. As a result, imaging acquired within the last 6 years constitutes a large proportion of the data used in this study. Ratio of female/male patients was consistent across all datasets (54.9%-66.2%) except S_{rad} and C_{rad} which were 87.5% and 83.3% female. Age distribution was bimodal, with a mean age of 12.6 (3.6) years for pediatric acquisitions and 58 (16) years for adult acquisitions.

Table 8.2 Detailed acquisition and demographic information for each dataset. Acquisition year and age are expressed as mean (range). Calgo, Salgo, Crad, and Srad were used for model development and validation; C,S,iC, iS, and B were used for institution-wide model deployment and validation. When the number of patients was less than the number of accessions for a given dataset, this indicated a subset of patients had more than one visit. When the number of images was greater than the number of accessions, a subset of patients had repeat images within the same accession.

Dataset description and shorthand	Number of patients	Number of accessions	Number of images	Acquisition year	Patient age	Patient sex
Coronal full spine radiographs with curve prediction (**C**)	16129	20527	21382	2015 (2008, 2020)	46.9 (1.16, 99.0)	60.9% F
Sagittal full spine radiographs with curve prediction (**S**)	15628	19000	19857	2015 (2008, 2020)	49.7 (0.59, 105.0)	57.3% F
Radiographs with cm of coronal imbalance in report (**iC**)	1450	1643	1643	2013 (2008, 2020)	50.8 (4.0, 95.0)	66.2% F
Radiographs with cm of sagittal imbalance in report (**iS**)	2415	2712	2712	2013 (2008, 2020)	55.3 (4.0, 95.0)	60.7% F
Bi-planar full spine radiographs with 3D curves (**B**)	15378	18595	20118	2015 (2008, 2020)	49.5 (1.16, 99.0)	58.5% F
Radiographs measured for sagittal imbalance (**Srad**)	8	8	8	2012 (2009, 2014)	53.4 (20.0, 75.0)	87.5% F
Radiographs measured for coronal imbalance (**Crad**)	12	12	12	2011 (2009, 2014)	51.8 (19.0, 75.0)	83.3% F
Coronal radiographs with keypoint annotations (**Calgo**)	188	194	194	2014 (2009, 2018)	51.6 (4.0, 85.0)	56.6% F
Sagittal radiographs with keypoint annotations (**Salgo**)	366	366	366	2014 (2009, 2018)	55.1 (4.0, 88.0)	54.9% F

8.4.1 Model testing

The spinal landmark detection models showed robust performance across the test set (Table 8.3). On average, pointwise MAE for coronal and sagittal images was <10mm, with a majority of pointwise errors <5mm and few errors >20mm. The largest errors were driven by a y-axis shift, where predicted landmarks were placed on vertebra edges rather than corners due to regions with low image quality or disagreement on the location of the landmark endplates. Y-axis shift was preferable over x-axis shift since the former still allowed for accurate fitting of spinal contours. Although a direct comparison is not possible due to different training and testing datasets, transforming our landmark predictions and ground truth points to match Multi-View Correlation Network MVC-Net's[257] scale we saw 79%, 66% lower test error (0.0095 vs 0.0459, 0.0136 vs. 0.0398) in sagittal and coronal views, despite MVC-Net requiring both views as input.

Table 8.3 Model performance on test set, continuous values are expressed as mean with range in parenthesis. Imbalance describes the range of imbalance measurements as defined by the ground truth landmarks. MAE = mean absolute error, MAD = mean absolute difference, CCC= concordance correlation coefficient, mm= millimeters

	Number of images	Imbalance measurements (mm)	Pointwise Mean Absolute Error (mm)	Imbalance Mean Absolute Difference (mm)	Imbalance CCC
Coronal radiographs with keypoint annotations, test (**Calgo**)	28	7.72 (-19.6, 43.4)	8.68 (3.6, 24.8)	2.42 (0.013, 10.3)	0.973
Sagittal radiographs with keypoint annotations, test (**Salgo**)	40	19.4 (-62.3, 127.0)	5.44 (1.97, 20.8)	3.58 (0.039, 14.3)	0.993

Overall, S1 and C7 endplate landmark identification and midline curve extraction was highly reliable, example overlays in Figure 8.5, Figure 8.6. Imbalance measurements derived from ground truth landmarks in the test set spanned from -19.6 to 43.4mm of coronal imbalance and -62.3 to 127mm of sagittal imbalance similar to pathological ranges reported in the literature[79; 266], and included patients with varied spinal deformities. Predicted imbalance measurements were in excellent agreement with landmark measurements (CCC of 0.993 and 0.973 for sagittal and coronal imbalance respectively). In a small clinical dataset (S_{rad}, C_{rad}), algorithm imbalance measurements remained in good agreement with each radiologist (CCC of 0.974, 0.943 for

sagittal and 0.941, 0.948 for coronal, MAD <10mm for all samples, Figure 8.7). For sagittal imbalance rater-algorithm agreement was 3.3% to 6.2% higher than inter-rater agreement. For coronal imbalance, rater-algorithm agreement was 0.3% to 1.1% lower than inter-rater agreement.

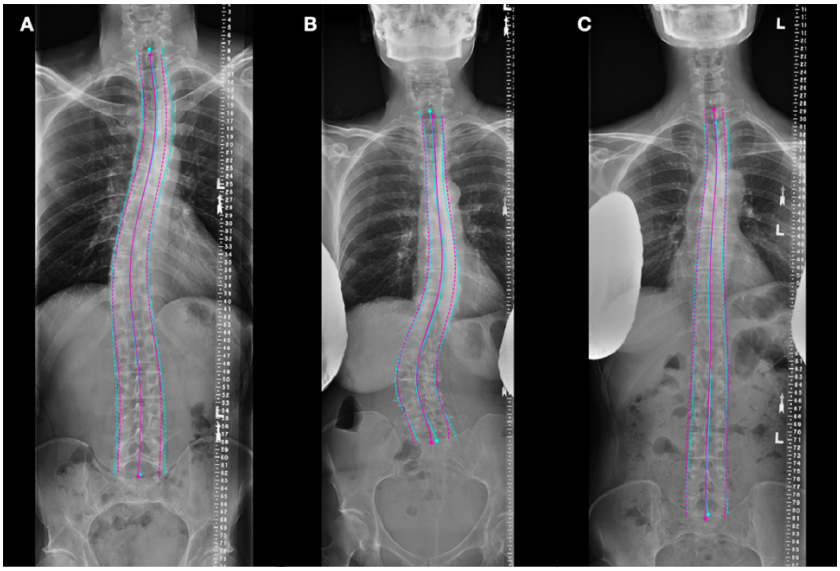


Figure 8.5 Frontal spine radiographs from test set. Ground truth curves in cyan, predicted curves in magenta. Dashed lines are spine contours, solid line is midline. Filled circles are S1 and C7 landmarks for coronal imbalance calculation.

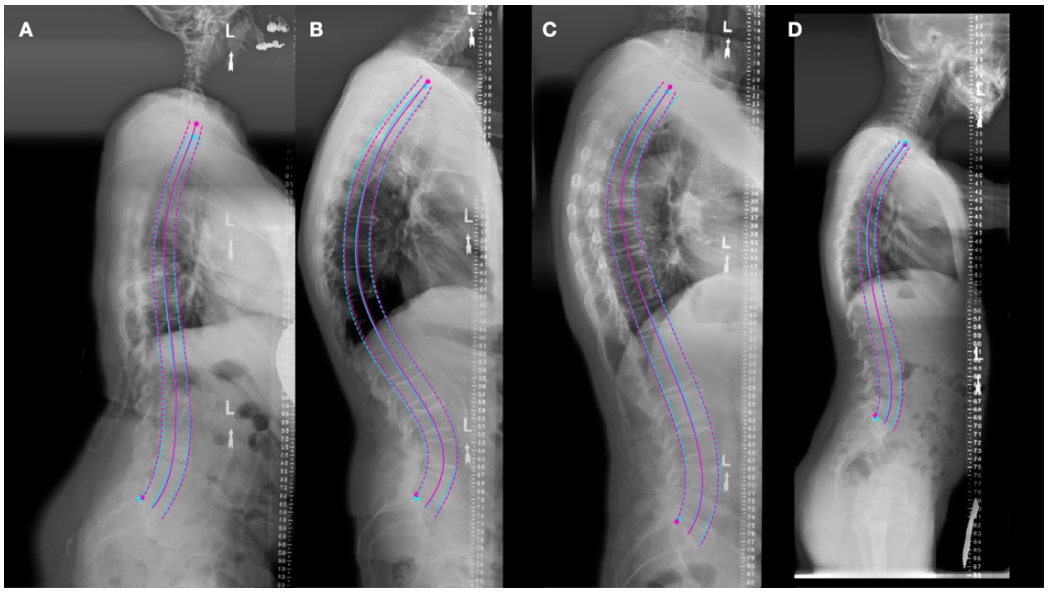


Figure 8.6 Lateral spine radiographs from test set. Ground truth curves in cyan, predicted curves in magenta. Dashed lines are spine contours, solid line is midline. Filled circles are S1 and C7 landmarks for sagittal imbalance calculation.

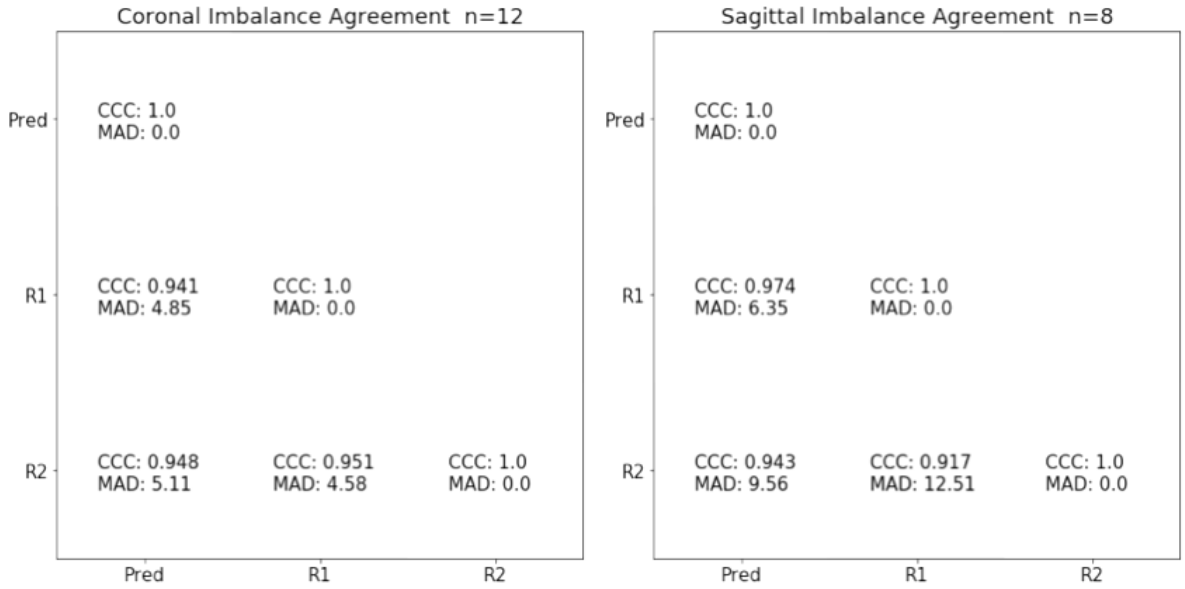


Figure 8.7 Matrices show inter-rater and rater-algorithm agreement on coronal and sagittal imbalance measurements performed on a PACS workstation. Pred = algorithm, R1,R2= Rater 1,2 , CCC=concordance correlation coefficient, MAD=mean absolute difference [mm]

8.4.2 Institution-wide validation

There was moderate agreement between radiology reports and predicted imbalance (CCC 0.916 sagittal, 0.731 coronal) (Figure 8.8). Reduced agreement in this dataset was expected given human errors in radiology reports and errors in automatic text parsing for label extraction. Differences >5cm between reported measurement and predicted measurement were investigated: 60% caused by errors in reporting (for example sagittal and coronal measurements flipped, or use of qualifiers “more than”, ”at least”), 18% from errors in text extraction (for example previous value for imbalance extracted from report), and 12% had report text and accession number potentially mismatched (deep learning prediction overlays are reasonable but significantly off of the reported measurements). This further demonstrated algorithm validity and generalizability to the institution-wide dataset, as this dataset included measurements from several radiologists and x-ray sources. Promising qualitative results on the AASCE challenge images suggested the landmark detection algorithm may be robust to shifts in patient population, data acquisition between institutions, and image compression. However,

understanding how performance is affected by more severe domain shift will require additional investigation.

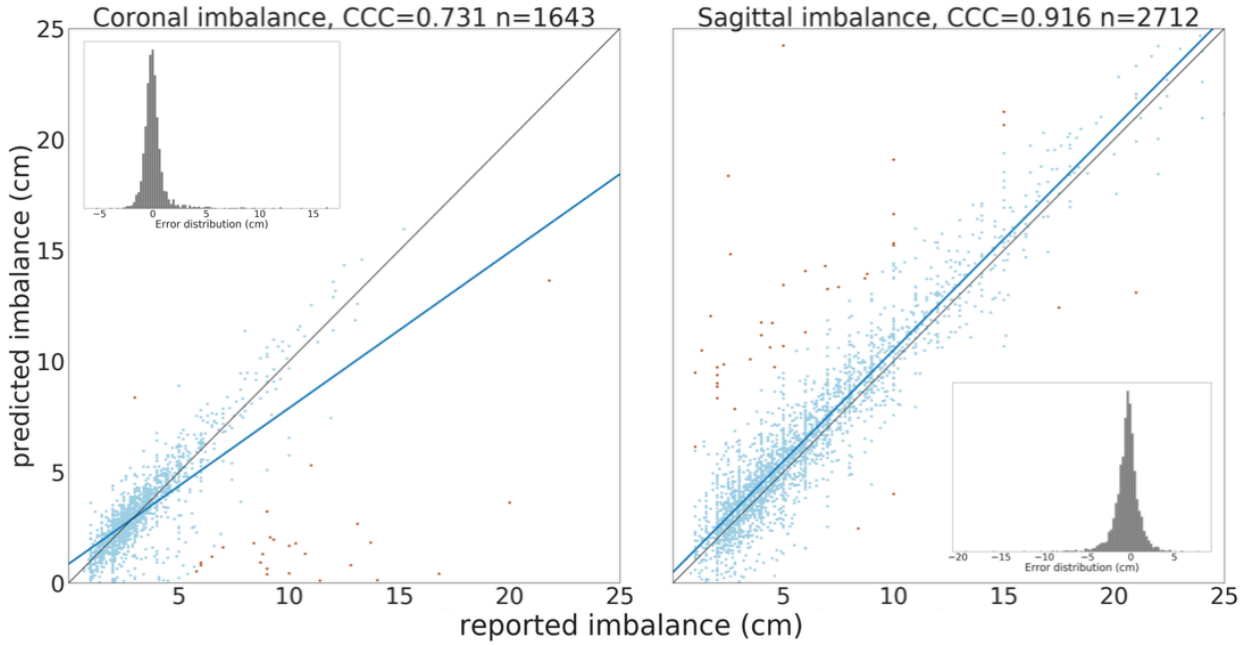


Figure 8.8 Scatterplot showing agreement between absolute imbalance measurements in cm extracted from radiology report and predicted imbalance measurements. Blue points indicate errors within 5cm, red points are errors greater than 5cm. Blue line represents the line of best fit to all data points. A plot of the distribution of errors is shown as a grey histogram, where negative errors correspond to overestimation of imbalance. CCC=concordance correlation coefficient, cm=centimeters

8.4.3 Global Alignment Parameters

Sagittal and coronal imbalance were measured using predicted keypoints and image header information. Figure 8.9 shows the resulting mean value and 95% confidence ellipsoid of sagittal and coronal imbalance parameters by age group. After age 50, global sagittal alignment became increasingly positive. However, these results should be interpreted with caution due to sampling bias: the patient population undergoing full spine x-rays are not likely a representative sample of the population.

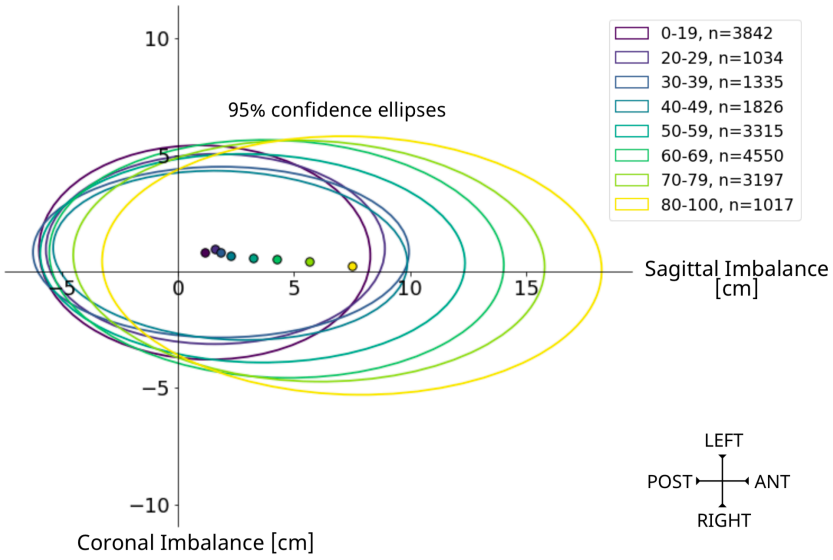


Figure 8.9 Sagittal and coronal imbalance parameters for all patients with bilateral x-ray imaging, separated by age groups. Filled datapoints represent centroid per age group while surrounding ellipse is the 95% confidence interval of the parameter set.

8.4.4 Shape modes

A total of 8 shape modes (Figure 8.4) described 99.68% of shape variation in the patient population. Modes 1 / 3 / 7 were dominated by sagittal plane variations, while variations in modes 2 / 4 / 8 were localized to the coronal plane. Modes 5 / 6 were a combination. Modes 1 / 2 accounted for 58.3%, 20.2% of the total shape variation and reflected changes in sagittal and coronal imbalance respectively. Mode directions agreed with imbalance conventions (positive, negative). Mode 3 plot (11.5% of variation) had a single point of intersection near T6, where increasingly negative values showed exaggerated lumbar lordosis and posterior sagittal balance while positive values a C-shaped lateral spine and anterior sagittal balance. In the coronal plane, negative values were linked to a minor rightward thoracic curve.

Mode 4 plot (4.84% of variation) had a single point of intersection near T11, negative values indicated major rightward lumbar curves, positive values major leftward curves, with curve magnitude scaling with mode values and a compensatory change in imbalance.

Mode 5 (2.36% of variation) and mode 6 (1.78% of variation) plots had two points of intersection, near T4 and L2. Negative values of mode 5 were linked to a ‘flat back’ shape with a

small rightward lumbothoracic curve, while positive values were associated with mid-thoracic kyphosis. Negative values of mode 6 showed a double curve shape with a major rightward lumbothoracic curve and minor leftward lumbar curves and upper thoracic curves, with the opposite true for positive values. Lastly, mode 7 (0.43%) and mode 8 (0.34%) plots had three intersection points and reflected local changes in sagittal and coronal curves. Positive values for mode 7 were linked to upper thoracic kyphosis. While shape modes were interpreted using common clinical terms for spinal deformity, important observations were gleaned by examining shape mode plots. For example, positive and negative values in mode 2 appeared mostly symmetric, however axial projections revealed a twisted loop shape in the negative values. Several axial projection shapes described in Pasha et al[267]– V-shape, S-shape, closed loop– were similarly identified in this study. T-SNE (Figure 8.10) did not separate patients into visually distinct clusters, instead creating a large point cloud where the main directions of variation corresponded to variations in shape modes 1 and 2. Patients with similar cobb angle measurements were only co-localized if sagittal plane curves were also similar. Coloring the point cloud by age, sex, image acquisition year did not uncover any obvious patient clusters.

8.5 Discussion

The developed landmark extraction algorithms demonstrated robust performance across the tested datasets. Once validated, algorithms were run on institution-wide data with pairs of frontal (Anterior-Posterior) and lateral (Left-Lateral) view radiographs to reconstruct approximations of 3D spine shape. A detailed discussion of merits and limitations of this approach is warranted.

As a first point of merit, the proposed method could enhance current clinical assessment of spine radiographs. Landmark annotations are not feasible within the clinical workflow as they require significant user-input and are vulnerable to user error. Several factors can reduce visibility and prevent the reliable manual identification of landmarks including severe spinal deformity, overlapping soft tissues, and visible lead shields. Moreover, anatomical landmarks

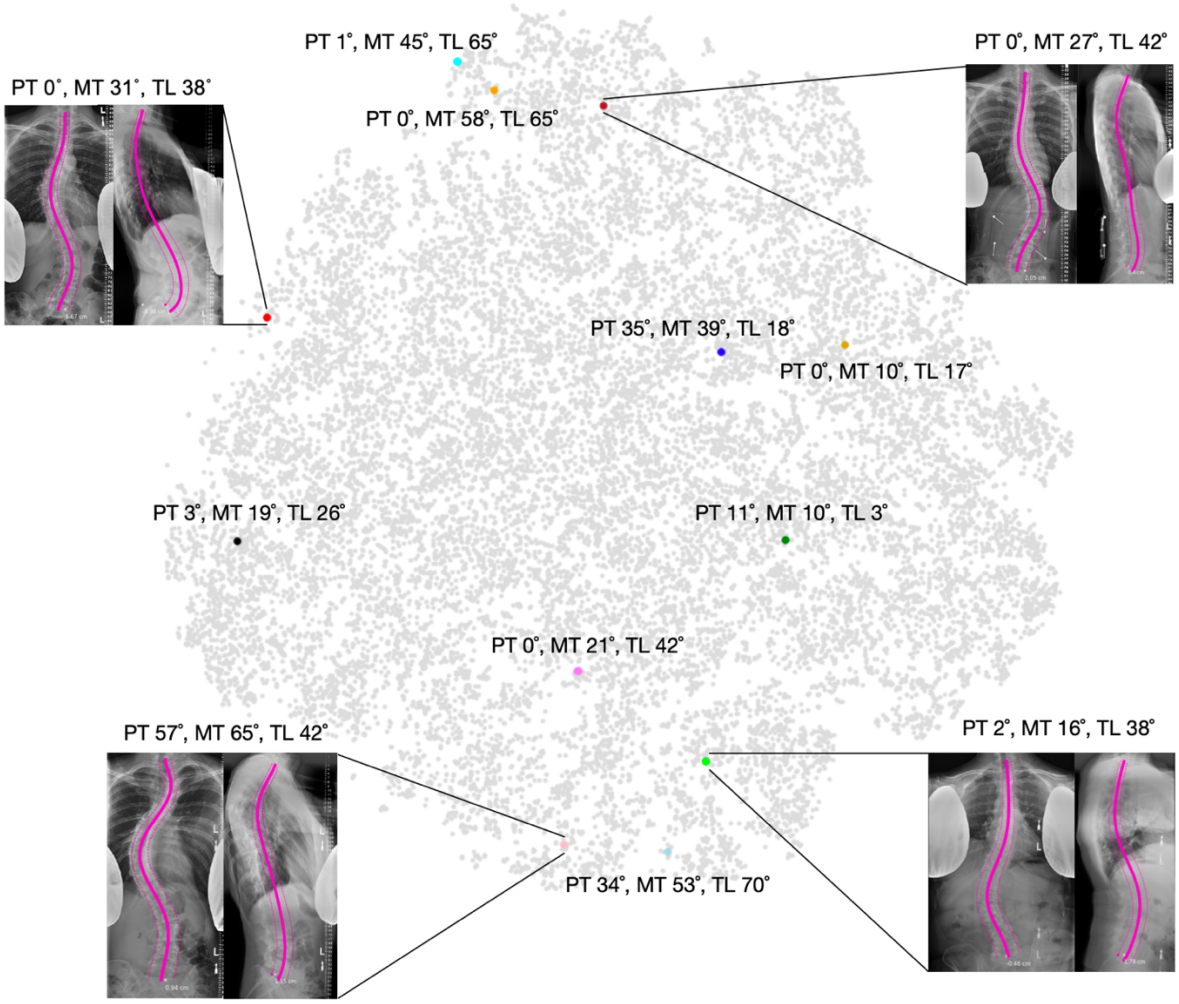


Figure 8.10 T-SNE embedding of 8 3D spine shape modes. Coronal Cobb angle measurements from 12 randomly selected spines displayed near colored point or above inset with frontal and lateral radiographs. Cobb angles were taken as the average measurements from 3 radiologists. Notice the similarities in Cobb between the upper left and upper right coronal images. TSNE parameters: learning rate 200, perplexity 100, iterations 800.

may present differently in patients with obesity, osteoporosis, transitional anatomy, fused bone, or at variable skeletal maturities. The sacral plateau (S1) and lower endplate of C7 were chosen as reliable anchor points, since nearby anatomical landmarks (ribcage, spinous processes, sacrum) allow for reliable identification even in low visibility settings. Automatic imbalance measurements were validated using both manual annotations and a large clinical dataset, providing evidence of the accurate identification of these anchor points. The landmark

prediction, midline curve fit, quality control, and 3D reconstruction for the proposed method are fully automatic and therefore have the potential to scale and integrate into the clinical workflow. Once deployed, prediction algorithms could run online–where clinicians would request results for a specific image pair– or offline– where inference would run on images shortly after acquisition and results would be stored in PACS. For patients with one or more previous visits, clinicians could quickly assess changes in 2D and approximate 3D curvature of the spine (example longitudinal case presented in Figure 8.11).

For patients without historical data, clinicians would be presented with a set of similar cases at their institution based on age, sex, and curvature similarity (defined as Mahalanobis distance in shape space). Retrieval of similar cases, as first proposed for coronal radiographs by Menon et al.[268], would allow for treatment planning based previous experiences with similar patients.

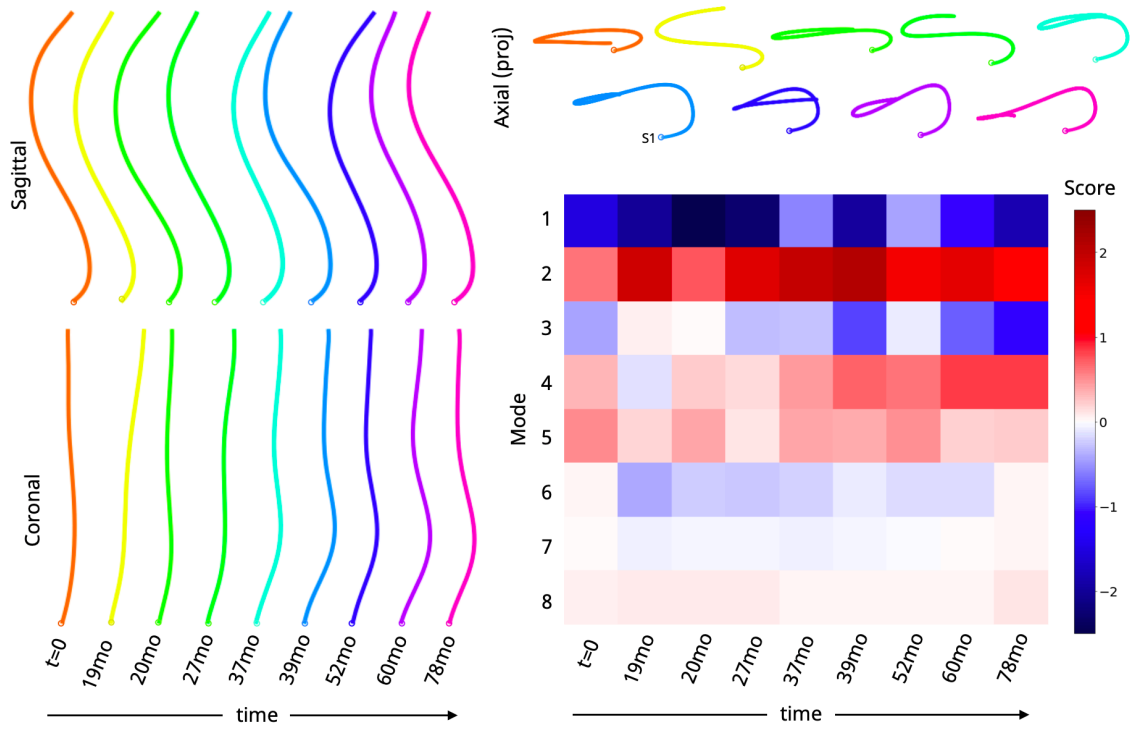


Figure 8.11 (L, Top R) Sagittal, coronal, and axial views of spine midlines for a single patient: a 12M with back pain at initial visit (t=0) monitored over the course of 6.5 years. Bilateral imaging was acquired during each. Slight changes in patient positioning and spinal curve are identifiable between visits. (Bottom Right) Shape mode scores over time. Notice a gradual increase in score for shape mode 4 starting at 37 months, which coincides with the onset of a leftward lumbar curve.

Furthermore, similarity grouping would enable the identification of patient subgroups for retrospective observational research or provide prevalence estimates of specific spine deformities for future cohort designs.

Second, an approximate 3D spine shape provides a rich description of global deformity compared to single plane evaluations, even with noise arising from an imperfect biplanar reconstruction. Shape mode embedding highlighted differences between patients with high curve similarity in the coronal plane but low similarity in the sagittal plane. Global curvature informs the distribution of biomechanical loads on the spine and might be important to understand risk of curve progression and assist in clinical management[269]. For example, recent AIS literature showed that pre-operative 3D shape, early post-operative shape, and information on fusion levels could be used to predict surgical outcomes 2 years post-op[267]. A significant body of literature exists on classification of 2D and 3D spine shapes for AIS and ASD patients[48; 270]. While it would be worthwhile to see if these classes naturally group in 3D shape space, the proposed method does not impose a specific classification scheme on the results, which more appropriately displays the spectrum of spinal deformities.

Third, it is well known that deep learning models are sensitive to shifts in data acquisition and can behave unexpectedly when tested on new populations[271]. Therefore, we have hosted the algorithms online such that they are easily accessed by the research community. To our knowledge, we are the first testable algorithm for spine midline extraction from biplanar radiographs, with an approximate runtime of 20 seconds per image pair.

A major limitation of the proposed method is that the 3D curve reconstruction is only an approximation of 3D shape since it is based on two non-calibrated acquisition planes. Calibration of the two image planes was not feasible for this study as the minimum set of parameters required to establish stereocorrespondence using epipolar geometry (6 anatomical landmarks per vertebra, known calibration parameters, a known distance landmark for scale or focal length information) were not available on all images. Additionally, published methods that

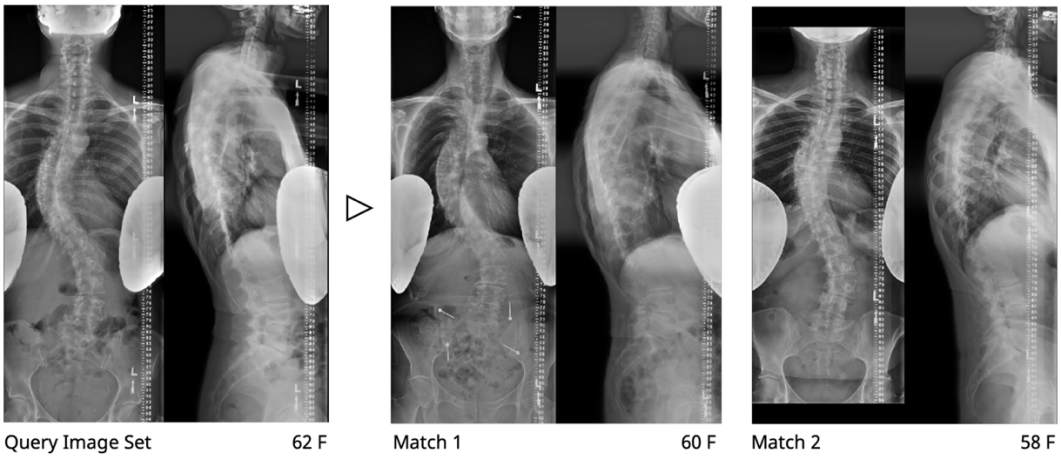


Figure 8.12 Similarity retrieval using Mahalanobis distance. Spine shape is extracted from query images and compared to a patient database of spine shapes. Patients with high demographic and shape similarity are returned.

perform self-calibration on landmarks such as spinal contours or midlines rely on statistical shape models based on separate databases of CT scans to infer or constrain a full 3D spine reconstruction[272-274]. Given our interest in understanding global shape variations across our institution we opted against using pre-existing shape models for reconstruction. It is important to emphasize that limitations associated with non-calibrated acquisitions apply to all spine measurements performed on biplanar radiographs including imbalance, Cobb angles, lordosis, kyphosis, and pelvic parameters. It follows that our proposed method would not be appropriate for applications where high-precision 3D skeletal reconstructions of pedicles and vertebral bodies are required, such as finite element simulations.

Calibration issues can be addressed during acquisition by using dedicated simultaneous low-dose stereoradiographic (EOS) systems achieving equivalent or improved image quality and measurement reliability with less radiation than a conventional radiograph[275; 276]. While this technology holds great promise, upgrading to these systems can be cost-prohibitive and even institutions with EOS systems have a wealth of historical data available that could be analyzed retrospectively to aid clinical decision making.

A minor limitation of the proposed method is the limited scope of the development dataset. While the dataset included a wide range of spinal deformities, radiograph sources, and demographics, it did not include bending or post-operative radiographs. Therefore, an abstention mechanism was built into the pipeline: algorithm predictions for these out-of-distribution images have high variability and are flagged as invalid during the autoQC step. Automatic landmark prediction for instrumented spinal radiographs has been investigated in[258], but was considered out of the scope for this study given that key landmarks are often obscured[277]. Furthermore, fused spinal levels in an instrumented patient population were likely to result in different spine shape modes as compared to the pre- or non-operative population.

The most surprising finding in this work was the range in actual sagittal and coronal imbalance among images linked to radiology reports stating “no imbalance”. Our early experiments attempted to train an image classifier to recognize balance/imbalance on full spine radiographs using labels extracted from the radiology report, but performance was suboptimal. After pivoting to keypoint annotations, we ran inference on images whose radiology reports stated “no imbalance” and found true measurements of approximately +/- 2.5 cm of coronal imbalance and +/- 5 cm sagittal imbalance, beyond the clinical threshold for imbalance of 2 cm. This has important implications: when no exact measurement was provided, qualitative descriptions of spinal alignment (‘no imbalance’, ‘neutral balance’, ‘mild’) were subjective.

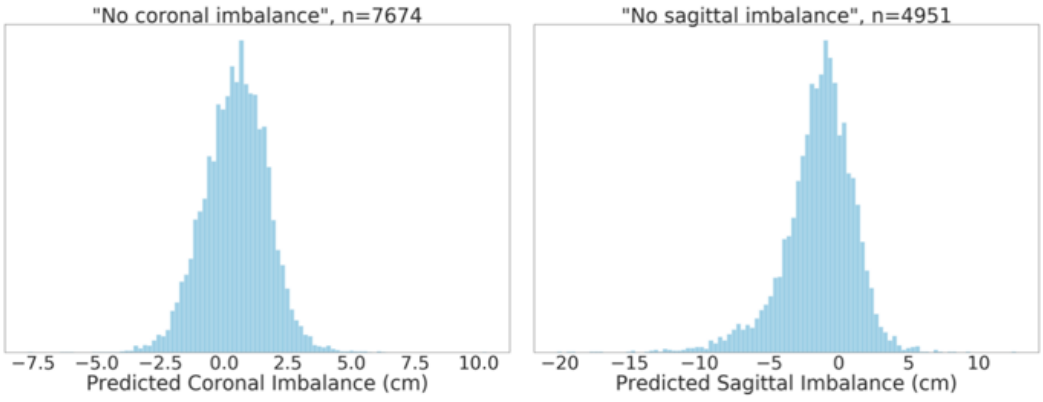


Figure 8.13 Predicted coronal and sagittal imbalance measurements for images whose report stated “no coronal imbalance” or “no sagittal imbalance”.

8.6 Conclusion

Future work will focus on: (1) Testing and documenting the landmark algorithms' robustness to domain shift and identifying new failure modes. Specifically, automatic vertebral midline extraction and shape modeling will be tested on radiographs pooled from several collaborating institutions. Moreover, individual researchers are encouraged to test the algorithm online and provide feedback, reporting cases of success or failure. This follows guidelines outlined in a recent review of machine learning for scoliosis by Chen et al.[278] calling for "heterogenous test sets" for spine deformity research and evaluation of ML tools by multidisciplinary teams. (2) Using a curated subset of pre-operative AIS, ASD images and surgical outcomes, determining if specific partitions of the 3D shape space may have more favorable surgical response.

In summary, this study describes a new method for automatic extraction of the vertebral midline from biplanar radiographs and a method describing 3D spine shapes through 8 interpretable shape modes. Deployed institution-wide, this method has the potential to enhance clinical assessment of spine deformities in AIS and ASD.

References

1. Cieza A, Causey K, Kamenov K, et al. 2021. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 396:2006-2017.
2. Mow VC, Gu WY, Chen FH. 2005. Structure and function of articular cartilage and meniscus. Basic orthopaedic biomechanics and mechano-biology 3:181-258.
3. Antons J, Marascio MGM, Nohava J, et al. 2018. Zone-dependent mechanical properties of human articular cartilage obtained by indentation measurements. J Mater Sci-Mater M 29.
4. Muir H, Bullough P, Maroudas A. 1970. The distribution of collagen in human articular cartilage with some of its physiological implications. J Bone Joint Surg Br 52:554-563.
5. Buckwalter JA, Mankin HJ. 1997. Articular cartilage .1. Tissue design and chondrocyte-matrix interactions. Journal of Bone and Joint Surgery-American Volume 79a:600-611.
6. Askew MJ, Mow VC. 1978. Biomechanical Function of Collagen Fibril Ultrastructure of Articular-Cartilage. J Biomech Eng-T Asme 100:105-115.

7. Flowers SA, Zieba A, Ornros J, et al. 2017. Lubricin binds cartilage proteins, cartilage oligomeric matrix protein, fibronectin and collagen II at the cartilage surface. Sci Rep-Uk 7.

8. Langer F, Gross AE. 1974. Immunogenicity of Allograft Articular-Cartilage. Journal of Bone and Joint Surgery-American Volume A 56:297-304.

9. Hardingham TE, Fosang AJ. 1995. The structure of aggrecan and its turnover in cartilage. J Rheumatol Suppl 43:86-90.

10. Buckwalter J, Mankin H. 1997. Instructional course lectures, The American Academy of Orthopaedic Surgeons-articular Cartilage. Part I: Tissue design and chondrocyte-matrix interactions. Jbjs 79:600-611.

11. Buckwalter JA, Mankin HJ. 1998. Articular cartilage: Tissue design and chondrocyte-matrix interactions. Aaos Instr Cours Lec 47:477-486.

12. Gilmore RS, Palfrey AJ. 1987. A Histological Study of Human Femoral Condylar Articular-Cartilage. Journal of Anatomy 155:77-85.

13. Hoemann CD, Lafantaisie-Favreau CH, Lascau-Coman V, et al. 2012. The cartilage-bone interface. J Knee Surg 25:85-97.

14. Buckwalter JA, Mankin HJ. 1998. Articular cartilage: Degeneration and osteoarthritis, repair, regeneration, and transplantation. Aaos Instr Cours Lec 47:487-504.

15. Pan J, Zhou XZ, Li W, et al. 2009. In Situ Measurement of Transport between Subchondral Bone and Articular Cartilage. Journal of Orthopaedic Research 27:1347-1352.

16. Buckwalter JA, Mow VC, Ratcliffe A. 1994. Restoration of Injured or Degenerated Articular Cartilage. J Am Acad Orthop Surg 2:192-201.

17. Fox AJS, Wanivenhaus F, Burge AJ, et al. 2015. The Human Meniscus: A Review of Anatomy, Function, Injury, and Advances in Treatment. Clin Anat 28:269-287.

18. Petersen W, Tillmann B. 1998. Collagenous fibril texture of the human knee joint menisci. Anat Embryol 197:317-324.

19. Bullough PG, Munuera L, Murphy J, et al. 1970. The strength of the menisci of the knee as it relates to their fine structure. J Bone Joint Surg Br 52:564-567.

20. Danzig L, Resnick D, Gonsalves M, et al. 1983. Blood-Supply to the Normal and Abnormal Menisci of the Human Knee. Clin Orthop Relat R:271-276.

21. Mow VC, Arnoczky SP, Jackson DW. 1992. Knee Meniscus: Basic and Clinical Foundations: Raven Press;

22. Marchand F, Ahmed AM. 1990. Investigation of the laminate structure of lumbar disc anulus fibrosus. Spine (Phila Pa 1976) 15:402-410.

23. Yu J, Fairbank JCT, Roberts S, et al. 2005. The elastic fiber network of the anulus fibrosus of the normal and scoliotic human intervertebral disc. Spine 30:1815-1820.

24. Eyre DR, Muir H. 1976. Types I and II collagens in intervertebral disc. Interchanging radial distributions in annulus fibrosus. Biochem J 157:267-270.

25. Iatridis JC, MacLean JJ, O'Brien M, et al. 2007. Measurements of proteoglycan and water content distribution in human lumbar intervertebral discs. Spine 32:1493-1497.

26. Lyons G, Eisenstein SM, Sweet MBE. 1981. Biochemical-Changes in Intervertebral-Disk Degeneration. Biochim Biophys Acta 673:443-453.

27. Roberts S, Menage J, Urban JP. 1989. Biochemical and structural properties of the cartilage end-plate and its relation to the intervertebral disc. Spine (Phila Pa 1976) 14:166-174.

28. Hassler O. 1969. Human Intervertebral Disc - a Micro-Angiographical Study on Its Vascular Supply at Various Ages. Acta Orthopaedica Scandinavica 40:765-+.

29. Urban J, Holm S, Maroudas A. 1978. Diffusion of small solutes into the intervertebral disc: as in vitro study. Biorheology 15:203-223.

30. Horner HA, Urban JPG. 2001. 2001 Volvo Award winner in basic science studies: Effect of nutrient supply on the viability of cells from the nucleus pulposus of the intervertebral disc. Spine 26:2543-2549.

31. Sharabi M, Wade K, Haj-Ali R. 2018. The Mechanical Role of Collagen Fibers in the Intervertebral Disc. Biomechanics of the Spine: Basic Concepts, Spinal Disorders and Treatments:105-123.

32. Burr DB, Akkus O. 2014. Bone Morphology and Organization. Basic and Applied Bone Biology:3-25.

33. Oftadeh R, Perez-Viloria M, Villa-Camacho JC, et al. 2015. Biomechanics and Mechanobiology of Trabecular Bone: A Review. J Biomech Eng-T Asme 137.

34. Maeda K, Mochizuki T, Kobayashi K, et al. 2020. Cortical thickness of the tibial diaphysis reveals age- and sex-related characteristics between non-obese healthy young and elderly subjects depending on the tibial regions. J Exp Orthop 7:78.

35. Niimi R, Kono T, Nishihara A, et al. 2015. Cortical thickness of the femur and long-term bisphosphonate use. J Bone Miner Res 30:225-231.

36. Ritzel H, Amling M, Posl M, et al. 1997. The thickness of human vertebral cortical bone and its changes in aging and osteoporosis: A histomorphometric analysis of the complete spinal column from thirty-seven autopsy specimens. Journal of Bone and Mineral Research 12:89-95.

37. Boyle C, Kim IY. 2011. Three-dimensional micro-level computational study of Wolff's law via trabecular bone remodeling in the human proximal femur using design space topology optimization. Journal of Biomechanics 44:935-942.

38. Lovejoy CO, Meindl RS, Ohman JC, et al. 2002. The maka femur and its bearing on the antiquity of human walking: Applying contemporary concepts of morphogenesis to the human fossil record. Am J Phys Anthropol 119:97-133.

39. Bibby SRS, Jones DA, Ripley RM, et al. 2005. Metabolism of the intervertebral disc: Effects of low levels of oxygen, glucose, and pH on rates of energy metabolism of bovine nucleus pulposus cells. Spine 30:487-496.

40. Buckwalter JA. 1998. Articular cartilage: Injuries and potential for healing. J Orthop Sport Phys 28:192-202.

41. Battie MC, Videman T, Gill K, et al. 1991. 1991 Volvo Award in Clinical Sciences - Smoking and Lumbar Intervertebral-Disk Degeneration - an Mri Study of Identical-Twins. Spine 16:1015-1020.

42. Bierma-Zeinstra SM, van Middelkoop M. 2017. Osteoarthritis: In search of phenotypes. Nat Rev Rheumatol 13:705-706.

43. Jeon OH, Kim C, Laberge RM, et al. 2017. Local clearance of senescent cells attenuates the development of post-traumatic osteoarthritis and creates a pro-regenerative environment. Nat Med 23:775-781.

44. Courties A, Berenbaum F, Sellam J. 2019. The Phenotypic Approach to Osteoarthritis: A Look at Metabolic Syndrome-Associated Osteoarthritis. Joint Bone Spine 86:725-730.

45. Carragee EJ, Don AS, Hurwitz EL, et al. 2009. 2009 ISSLS Prize Winner: Does Discography Cause Accelerated Progression of Degeneration Changes in the Lumbar Disc A Ten-Year Matched Cohort Study. Spine 34:2338-2345.

46. Vergroesen PPA, Kingma I, Emanuel KS, et al. 2015. Mechanics and biology in intervertebral disc degeneration: a vicious circle. Osteoarthr Cartilage 23:1057-1070.

47. Hunter DJ, Bierma-Zeinstra S. 2019. Osteoarthritis. Lancet 393:1745-1759.

48. Aebi M. 2005. The adult scoliosis. European Spine Journal 14:925-948.

49. Bijlsma JWJ, Berenbaum F, Lafeber FPJG. 2011. Osteoarthritis: an update with relevance for clinical practice. Lancet 377:2115-2126.

50. Firner S, Zaucke F, Michael J, et al. 2017. Extracellular Distribution of Collagen II and Perifibrillar Adapter Proteins in Healthy and Osteoarthritic Human Knee Joint Cartilage. J Histochem Cytochem 65:593-606.

51. Lories RJ, Luyten FP. 2011. The bone-cartilage unit in osteoarthritis. Nat Rev Rheumatol 7:43-49.

52. Ding CH, Martel-Pelletier J, Pelletier JP, et al. 2007. Meniscal tear as an osteoarthritis risk factor in a largely non-osteoarthritic cohort: A cross-sectional study. Journal of Rheumatology 34:776-784.

53. Pauli C, Grogan SP, Patil S, et al. 2011. Macroscopic and Histopathologic Analysis of Human Knee Menisci in Aging and Osteoarthritis. Osteoarthr Cartilage 19:S202-S202.

54. Urban JPG, Roberts S. 2003. Degeneration of the intervertebral disc. Arthritis Research & Therapy 5:120-130.

55. Adams MA, McNally DS, Dolan P. 1996. 'Stress' distributions inside intervertebral discs - The effects of age and degeneration. Journal of Bone and Joint Surgery-British Volume 78b:965-972.

56. Mimura M, Panjabi MM, Oxland TR, et al. 1994. Disc degeneration affects the multidirectional flexibility of the lumbar spine. Spine (Phila Pa 1976) 19:1371-1380.

57. Tsantrizos A, Ito K, Aebi M, et al. 2005. Internal strains in healthy and degenerated lumbar intervertebral discs. Spine (Phila Pa 1976) 30:2129-2137.

58. Capossela S, Schlafli P, Bertolo A, et al. 2014. Degenerated Human Intervertebral Discs Contain Autoantibodies against Extracellular Matrix Proteins. Eur Cells Mater 27:251-263.

59. Diebo BG, Shah NV, Boachie-Adjei O, et al. 2019. Adult spinal deformity. Lancet 394:160-172.

60. Sparrey CJ, Bailey JF, Safaee M, et al. 2014. Etiology of lumbar lordosis and its pathophysiology: a review of the evolution of lumbar lordosis, and the mechanics and biology of lumbar degeneration. Neurosurg Focus 36.

61. Enomoto M, Ukegawa D, Sakaki K, et al. 2012. Increase in Paravertebral Muscle Activity in Lumbar Kyphosis Patients by Surface Electromyography Compared With Lumbar

Spinal Canal Stenosis Patients and Healthy Volunteers. Journal of Spinal Disorders & Techniques 25:E167-E173.

62. Wu H-L, Ding W-Y, Shen Y, et al. 2012. Prevalence of vertebral endplate modic changes in degenerative lumbar scoliosis and its associated factors analysis. Spine 37:1958-1964.

63. Acaroglu RE, Dede O, Pellise F, et al. 2016. Adult spinal deformity: a very heterogeneous population of patients with different needs. Acta Orthop Traumato 50:57-62.

64. Veldhuizen AG, Wever DJ, Webb PJ. 2000. The aetiology of idiopathic scoliosis: biomechanical and neuromuscular factors. Eur Spine J 9:178-184.

65. Wren TA, Beaupre GS, Carter DR. 1998. A model for loading-dependent growth, development, and adaptation of tendons and ligaments. J Biomech 31:107-114.

66. Crijns TJ, Stadhouder A, Smit TH. 2017. Restrained Differential Growth: The Initiating Event of Adolescent Idiopathic Scoliosis? Spine (Phila Pa 1976) 42:E726-E732.

67. Bibby SR, Fairbank JC, Urban MR, et al. 2002. Cell viability in scoliotic discs in relation to disc deformity and nutrient levels. Spine (Phila Pa 1976) 27:2220-2228; discussion 2227-2228.

68. Roberts S, Menage J, Eisenstein SM. 1993. The cartilage end-plate and intervertebral disc in scoliosis: calcification and other sequelae. J Orthop Res 11:747-757.

69. Balague F, Mannion AF, Pellise F, et al. 2012. Non-specific low back pain. Lancet 379:482-491.

70. Altaf F, Gibson A, Dannawi Z, et al. 2013. Adolescent idiopathic scoliosis. BMJ 346:f2508.

71. Bushberg JT, Boone JM. 2011. The essential physics of medical imaging: Lippincott Williams & Wilkins;

72. Elgazzar AH, Kazem N. 2015. Biological effects of ionizing radiation. The pathophysiologic basis of nuclear medicine: Springer; pp. 715-726.

73. Griffiths HJ. 1989. Tissue Substitutes in Radiation Dosimetry and Measurement. No. 4. Radiology 173:202-202.

74. Kellgren JH, Lawrence JS. 1957. Radiological assessment of osteo-arthrosis. Ann Rheum Dis 16:494-502.

75. Neumann G, Hunter D, Nevitt M, et al. 2009. Location specific radiographic joint space width for osteoarthritis progression. Osteoarthr Cartilage 17:761-765.

76. Oak SR, Ghodadra A, Winalski CS, et al. 2013. Radiographic joint space width is correlated with 4-year clinical outcomes in patients with knee osteoarthritis: data from the osteoarthritis initiative. Osteoarthr Cartilage 21:1185-1190.

77. Kinds M, Vincken K, Hoppinga T, et al. 2012. Influence of variation in semiflexed knee positioning during image acquisition on separate quantitative radiographic parameters of osteoarthritis, measured by Knee Images Digital Analysis. Osteoarthr Cartilage 20:997-1003.

78. Hoeffner E, Mukherji S, Srinivasan A, et al. 2012. Neuroradiology back to the future: spine imaging. American journal of neuroradiology 33:999-1006.

79. Glassman SD, Bridwell K, Dimar JR, et al. 2005. The impact of positive sagittal balance in adult spinal deformity. Spine (Phila Pa 1976) 30:2024-2029.

80. Cobb J. 1948. Outline for the study of scoliosis. Edwards JW ed Instructional Course Lectures, The American Academy Orthopaedic Surgeons 5:261–275.

81. Lenke LG, Edwards CC, Bridwell KH. 2003. The Lenke classification of adolescent idiopathic scoliosis: how it organizes curve patterns as a template to perform selective fusions of the spine. Spine 28:S199-S207.

82. Schwab F, Ungar B, Blondel B, et al. 2012. Scoliosis Research Society—Schwab adult spinal deformity classification: a validation study. Spine 37:1077-1082.

83. Terran J, Schwab F, Shaffrey CI, et al. 2013. The SRS-Schwab adult spinal deformity classification: assessment and clinical correlations based on a prospective operative and nonoperative cohort. Neurosurgery 73:559-568.

84. Levitt MH. 2013. Spin dynamics: basics of nuclear magnetic resonance: John Wiley & Sons;

85. Bezci SE, Werbner B, Zhou MH, et al. 2019. Radial variation in biochemical composition of the bovine caudal intervertebral disc. Jor Spine 2.

86. Hunter DJ, Guermazi A, Lo GH, et al. 2011. Evolution of semi-quantitative whole joint assessment of knee OA: MOAKS (MRI Osteoarthritis Knee Score). Osteoarthr Cartilage 19:990-1002.

87. Peterfy C, Guermazi A, Zaim S, et al. 2004. Whole-organ magnetic resonance imaging score (WORMS) of the knee in osteoarthritis. Osteoarthr Cartilage 12:177-190.

88. Pfirrmann CW, Metzdorf A, Zanetti M, et al. 2001. Magnetic resonance classification of lumbar intervertebral disc degeneration. Spine 26:1873-1878.

89. Modic MT, Steinberg PM, Ross JS, et al. 1988. Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. Radiology 166:193-199.

90. Meiboom S, Gill D. 1958. Modified spin-echo method for measuring nuclear relaxation times. Review of scientific instruments 29:688-691.

91. Nishioka H, Hirose J, Nakamura E, et al. 2012. T1rho and T2 mapping reveal the in vivo extracellular matrix of articular cartilage. J Magn Reson Imaging 35:147-155.

92. Wei B, Du XT, Liu J, et al. 2015. Associations between the properties of the cartilage matrix and findings from quantitative MRI in human osteoarthritic cartilage of the knee. Int J Clin Exp Patho 8:3928-3936.

93. Rautiainen J, Nissi MJ, Salo EN, et al. 2015. Multiparametric MRI assessment of human articular cartilage degeneration: Correlation with quantitative histology and mechanical properties. Magnetic Resonance in Medicine 74:249-259.

94. Li X, Cheng J, Lin K, et al. 2011. Quantitative MRI using T1ρ and T2 in human osteoarthritic cartilage specimens: correlation with biochemical measurements and histology. Magnetic resonance imaging 29:324-334.

95. Son M, Goodman SB, Chen W, et al. 2013. Regional variation in T1rho and T2 times in osteoarthritic human menisci: correlation with mechanical properties and matrix composition. Osteoarthritis Cartilage 21:796-805.

96. Dunn TC, Lu Y, Jin H, et al. 2004. T2 relaxation time of cartilage at MR imaging: Comparison with severity of knee osteoarthritis. Radiology 232:592-598.

97. Razmjoo A, Caliva F, Lee J, et al. 2021. T2 analysis of the entire osteoarthritis initiative dataset. J Orthop Res 39:74-85.

98. Nieminen MT, Toyras J, Rieppo J, et al. 2000. Quantitative MR microscopy of enzymatically degraded articular cartilage. Magn Reson Med 43:676-681.

99. Antoniou J, Pike GB, Steffen T, et al. 1998. Quantitative magnetic resonance imaging in the assessment of degenerative disc disease. Magn Reson Med 40:900-907.

100. Marinelli NL, Haughton VM, Munoz A, et al. 2009. T2 relaxation times of intervertebral disc tissue correlated with water content and proteoglycan content. Spine (Phila Pa 1976) 34:520-524.

101. Paul CPL, Smit TH, de Graaf M, et al. 2018. Quantitative MRI in early intervertebral disc degeneration: T1rho correlates better than T2 and ADC with biomechanics, histology and matrix content. PLoS One 13:e0191442.

102. Weidenbaum M, Foster RJ, Best BA, et al. 1992. Correlating magnetic resonance imaging with the biochemical content of the normal human intervertebral disc. J Orthop Res 10:552-561.

103. Yang B, Wendland MF, O'Connell GD. 2020. Direct Quantification of Intervertebral Disc Water Content Using MRI. Journal of Magnetic Resonance Imaging 52:1152-1162.

104. Chen W. 2015. Errors in quantitative T1rho imaging and the correction methods. Quantitative imaging in medicine and surgery 5:583.

105. Hanninen N, Rautiainen J, Rieppo L, et al. 2017. Orientation anisotropy of quantitative MRI relaxation parameters in ordered tissue. Sci Rep-Uk 7.

106. Wheaton AJ, Dodge GR, Elliott DM, et al. 2005. Quantification of cartilage biomechanical and biochemical properties via T1rho magnetic resonance imaging. Magn Reson Med 54:1087-1093.

107. van Tiel J, Kotek G, Reijman M, et al. 2016. Is T1rho Mapping an Alternative to Delayed Gadolinium-enhanced MR Imaging of Cartilage in the Assessment of Sulphated

Glycosaminoglycan Content in Human Osteoarthritic Knees? An in Vivo Validation Study. Radiology 279:523-531.

108. Regatte RR, Akella SV, Borthakur A, et al. 2002. Proteoglycan depletion–induced changes in transverse relaxation maps of cartilage: comparison of T2 and T1ρ. Academic radiology 9:1388-1394.

109. Akella SV, Regatte RR, Gougoutas AJ, et al. 2001. Proteoglycan-induced changes in T1rho-relaxation of articular cartilage at 4T. Magn Reson Med 46:419-423.

110. Johannessen W, Auerbach JD, Wheaton AJ, et al. 2006. Assessment of human disc degeneration and proteoglycan content using T1rho-weighted magnetic resonance imaging. Spine (Phila Pa 1976) 31:1253-1257.

111. Mulligan KR, Ferland CE, Gawri R, et al. 2015. Axial T1rho MRI as a diagnostic imaging modality to quantify proteoglycan concentration in degenerative disc disease. Eur Spine J 24:2395-2401.

112. Zhao J, Li X, Bolbos RI, et al. 2010. Longitudinal assessment of bone marrow edema-like lesions and cartilage degeneration in osteoarthritis using 3 T MR T1rho quantification. Skeletal Radiol 39:523-531.

113. Zarins ZA, Bolbos RI, Pialat JB, et al. 2010. Cartilage and meniscus assessment using T1rho and T2 measurements in healthy subjects and patients with osteoarthritis. Osteoarthritis Cartilage 18:1408-1416.

114. Iriondo C, Pedoia V, Majumdar S. 2020. Lumbar intervertebral disc characterization through quantitative MRI analysis: An automatic voxel-based relaxometry approach. Magn Reson Med 84:1376-1390.

115. Choi J-A, Gold GE. 2011. MR imaging of articular cartilage physiology. Magnetic Resonance Imaging Clinics 19:249-282.

116. Bengio Y, Courville A, Vincent P. 2013. Representation learning: A review and new perspectives. Ieee T Pattern Anal 35:1798-1828.

117. Rumelhart DE, Hinton GE, Williams RJ. 1986. Learning representations by back-propagating errors. nature 323:533-536.

118. Pearson K. 1901. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2:559-572.

119. Lin T-Y, Maire M, Belongie S, et al. 2014. Microsoft coco: Common objects in context. European conference on computer vision: Springer; pp. 740-755.

120. Zbontar J, Knoll F, Sriram A, et al. 2018. fastMRI: An open dataset and benchmarks for accelerated MRI. arXiv preprint arXiv:181108839.

121. Calivá F, Leynes AP, Shah R, et al. 2020. Breaking Speed Limits with Simultaneous Ultra-Fast MRI Reconstruction and Tissue Segmentation. Medical Imaging with Deep Learning: PMLR; pp. 94-110.

122. Gebru T, Morgenstern J, Vecchione B, et al. 2018. Datasheets for datasets. arXiv preprint arXiv:180309010.

123. Winkels M, Cohen TS. 2019. Pulmonary nodule detection in CT scans with equivariant CNNs. Medical image analysis 55:15-26.

124. Bronstein MM, Bruna J, Cohen T, et al. 2021. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. arXiv preprint arXiv:210413478.

125. LeCun Y, Boser B, Denker JS, et al. 1989. Backpropagation applied to handwritten zip code recognition. Neural computation 1:541-551.

126. Krizhevsky A, Sutskever I, Hinton GE. 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25:1097-1105.

127. He K, Zhang X, Ren S, et al. 2016. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; pp. 770-778.

128. Morales Martinez A, Caliva F, Flament I, et al. 2020. Learning osteoarthritis imaging biomarkers from bone surface spherical encoding. Magnetic resonance in medicine 84:2190-2203.

129. Tomita N, Cheung YY, Hassanpour S. 2018. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Comput Biol Med 98:8-15.

130. Huang G, Liu Z, Van Der Maaten L, et al. 2017. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition; pp. 4700-4708.

131. Krogue JD, Cheng KV, Hwang KM, et al. 2020. Automatic hip fracture identification and functional subclassification with deep learning. Radiology: Artificial Intelligence 2:e190023.

132. Tolpadi AA, Lee JJ, Pedoia V, et al. 2020. Deep Learning Predicts Total Knee Replacement from Magnetic Resonance Images. Sci Rep-Uk 10.

133. Ronneberger O, Fischer P, Brox T. 2015. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention: Springer; pp. 234-241.

134. Long J, Shelhamer E, Darrell T. 2015. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition; pp. 3431-3440.

135. Isensee F, Jaeger PF, Kohl SA, et al. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 18:203-211.

136. Desai AD, Caliva F, Iriondo C, et al. 2021. The international workshop on osteoarthritis imaging knee MRI segmentation challenge: a multi-institute evaluation and analysis framework on a standardized dataset. Radiology: Artificial Intelligence:e200078.

137. Milletari F, Navab N, Ahmadi S-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 fourth international conference on 3D vision (3DV): IEEE; pp. 565-571.

138. Moskvyak O, Maire F, Dayoub F, et al. 2021. Semi-supervised Keypoint Localization. arXiv preprint arXiv:210107988.

139. Castro DC, Walker I, Glocker B. 2020. Causality matters in medical imaging. Nature Communications 11:1-10.

140. Kingma DP, Ba J. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980.

141. Larson DB, Boland GW. 2019. Imaging Quality Control in the Era of Artificial Intelligence. J Am Coll Radiol 16:1259-1266.

142. George A, Kuzniecky R, Rusinek H, et al. 2020. Standardized Brain MRI Acquisition Protocols Improve Statistical Power in Multicenter Quantitative Morphometry Studies. J Neuroimaging 30:126-133.

143. Ganin Y, Ustinova E, Ajakan H, et al. 2016. Domain-adversarial training of neural networks. The journal of machine learning research 17:2096-2030.

144. Tzeng E, Hoffman J, Saenko K, et al. 2017. Adversarial discriminative domain adaptation. Proceedings of the IEEE conference on computer vision and pattern recognition; pp. 7167-7176.

145. Sudre CH, Anson BG, Ingala S, et al. 2019. Let's agree to disagree: Learning highly debatable multirater labelling. International Conference on Medical Image Computing and Computer-Assisted Intervention: Springer; pp. 665-673.

146. Jungo A, Meier R, Ermis E, et al. 2018. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention: Springer; pp. 682-690.

147. Raghu M, Zhang C, Kleinberg J, et al. 2019. Transfusion: Understanding transfer learning for medical imaging. arXiv preprint arXiv:190207208.

148. Sculley D, Holt G, Golovin D, et al. 2014. Machine learning: The high interest credit card of technical debt.

149. Milletari F, Navab N, Ahmadi SA. 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. Int Conf 3d Vision:565-571.

150. Yao Y, Rosasco L, Caponnetto A. 2007. On early stopping in gradient descent learning. Constructive Approximation 26:289-315.

151. Gal Y, Ghahramani Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. international conference on machine learning: PMLR; pp. 1050-1059.

152. Roy AG, Conjeti S, Navab N, et al. 2019. Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control. Neuroimage 195:11-22.

153. Srivastava N, Hinton G, Krizhevsky A, et al. 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15:1929-1958.

154. Torio CM, Moore BJ. 2016. National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2013: Statistical Brief #204. Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville (MD).

155. Loeser RF, Collins JA, Diekman BO. 2016. Ageing and the pathogenesis of osteoarthritis. Nat Rev Rheumatol 12:412-420.

156. Robinson WH, Lepus CM, Wang Q, et al. 2016. Low-grade inflammation as a key mediator of the pathogenesis of osteoarthritis. Nat Rev Rheumatol 12:580-592.

157. Calvo E, Palacios I, Delgado E, et al. 2004. Histopathological correlation of cartilage swelling detected by magnetic resonance imaging in early experimental osteoarthritis. Osteoarthritis Cartilage 12:878-886.

158. Eckstein F, Collins JE, Nevitt MC, et al. 2015. Brief Report: Cartilage Thickness Change as an Imaging Biomarker of Knee Osteoarthritis Progression: Data From the Foundation for

the National Institutes of Health Osteoarthritis Biomarkers Consortium. Arthritis Rheumatol 67:3184-3189.

159. Wirth W, Hunter DJ, Nevitt MC, et al. 2017. Predictive and concurrent validity of cartilage thickness change as a marker of knee osteoarthritis progression: data from the Osteoarthritis Initiative. Osteoarthritis Cartilage 25:2063-2071.

160. Buck RJ, Wyman BT, Le Graverand MP, et al. 2010. Osteoarthritis may not be a one-way-road of cartilage loss--comparison of spatial patterns of cartilage change between osteoarthritic and healthy knees. Osteoarthritis Cartilage 18:329-335.

161. Wijayaratne SP, Teichtahl AJ, Wluka AE, et al. 2008. The determinants of change in patella cartilage volume--a cohort study of healthy middle-aged women. Rheumatology (Oxford) 47:1426-1429.

162. Cicuttini FM, Hankin J, Jones G, et al. 2005. Comparison of conventional standing knee radiographs and magnetic resonance imaging in assessing progression of tibiofemoral joint osteoarthritis. Osteoarthr Cartilage 13:722-727.

163. Cicuttini FM, Wluka A, Wang YY, et al. 2002. The determinants of change in patella cartilage volume in osteoarthritic knees. Journal of Rheumatology 29:2615-2619.

164. Eckstein F, Wirth W, Lohmander LS, et al. 2015. Five-year followup of knee joint cartilage thickness changes after acute rupture of the anterior cruciate ligament. Arthritis Rheumatol 67:152-161.

165. Buck RJ, Wirth W, Dreher D, et al. 2013. Frequency and spatial distribution of cartilage thickness change in knee osteoarthritis and its relation to clinical and radiographic covariates - data from the osteoarthritis initiative. Osteoarthritis Cartilage 21:102-109.

166. Kauffmann C, Gravel P, Godbout B, et al. 2003. Computer-aided, method for quantification of cartilage thickness and volume changes using MRL: Validation study using a synthetic model. Ieee T Bio-Med Eng 50:978-988.

167. Koo S, Gold GE, Andriacchi TP. 2005. Considerations in measuring cartilage thickness using MRI: factors influencing reproducibility and accuracy. Osteoarthr Cartilage 13:782-789.

168. Schneider E, Nevitt M, McCulloch C, et al. 2012. Equivalence and precision of knee cartilage morphometry between different segmentation teams, cartilage regions, and MR acquisitions. Osteoarthritis Cartilage 20:869-879.

169. Jorgensen DR, Lillholm M, Genant HK, et al. 2013. On Subregional Analysis of Cartilage Loss from Knee MRI. Cartilage 4:121-130.

170. Bowes MA, Guillard GA, Vincent GR, et al. 2019. Precision, Reliability, and Responsiveness of a Novel Automated Quantification Tool for Cartilage Thickness: Data from the Osteoarthritis Initiative. J Rheumatol.

171. Gaj S, Yang M, Nakamura K, et al. 2019. Automated cartilage and meniscus segmentation of knee MRI with conditional generative adversarial networks. Magn Reson Med.

172. Norman B, Pedoia V, Majumdar S. 2018. Use of 2D U-Net Convolutional Neural Networks for Automated Cartilage and Meniscus Segmentation of Knee MR Imaging Data to Determine Relaxometry and Morphometry. Radiology 288:177-185.

173. Ambellan F, Tack A, Ehlke M, et al. 2019. Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative. Med Image Anal 52:109-118.

174. Liu F, Zhou Z, Jang H, et al. 2018. Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging. Magn Reson Med 79:2379-2391.

175. Bromiley PA, Kariki EP, Cootes TF. 2019. Error Estimation for Appearance Model Segmentation of Musculoskeletal Structures Using Multiple, Independent Sub-models. Cham: Springer International Publishing; pp. 53-65.

176. Glaser C, Burgkart R, Kutschera A, et al. 2003. Femoro-tibial cartilage metrics from coronal MR image data: Technique, test-retest reproducibility, and findings in osteoarthritis. Magn Reson Med 50:1229-1236.

177. Hanna F, Wluka AE, Ebeling PR, et al. 2006. Determinants of change in patella cartilage volume in healthy subjects. J Rheumatol 33:1658-1661.

178. Le Graverand MPH, Buck RJ, Wyman BT, et al. 2009. Subregional femorotibial cartilage morphology in women - comparison between healthy controls and participants with different grades of radiographic knee osteoarthritis. Osteoarthr Cartilage 17:1177-1185.

179. Prieto-Alhambra D, Judge A, Javaid MK, et al. 2014. Incidence and risk factors for clinically diagnosed knee, hip and hand osteoarthritis: influences of age, gender and osteoarthritis affecting other joints. Ann Rheum Dis 73:1659-1664.

180. Adams JG, McAlindon T, Dimasi M, et al. 1999. Contribution of meniscal extrusion and cartilage loss to joint space narrowing in osteoarthritis. Clin Radiol 54:502-506.

181. Karvonen RL, Negendank WG, Teitge RA, et al. 1994. Factors Affecting Articular-Cartilage Thickness in Osteoarthritis and Aging. Journal of Rheumatology 21:1310-1318.

182. Kamnitsas K, Bai W, Ferrante E, et al. 2018. Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Brainles 2017 10670:450-462.

183. De Fauw J, Ledsam JR, Romera-Paredes B, et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 24:1342-1350.

184. Wirth W, Benichou O, Kwoh CK, et al. 2010. Spatial patterns of cartilage loss in the medial femoral condyle in osteoarthritic knees: data from the Osteoarthritis Initiative. Magn Reson Med 63:574-581.

185. Wirth W, Larroque S, Davies RY, et al. 2011. Comparison of 1-year vs 2-year change in regional cartilage thickness in osteoarthritis results from 346 participants from the Osteoarthritis Initiative. Osteoarthritis Cartilage 19:74-83.

186. Eckstein F, Kwoh CK, Boudreau RM, et al. 2013. Quantitative MRI measures of cartilage predict knee replacement: a case-control study from the Osteoarthritis Initiative. Ann Rheum Dis 72:707-714.

187. Eckstein F, Cotofana S, Wirth W, et al. 2011. Greater Rates of Cartilage Loss in Painful Knees Than in Pain-Free Knees After Adjustment for Radiographic Disease Stage Data From the Osteoarthritis Initiative. Arthritis Rheum-Us 63:2257-2267.

188. Argentieri EC, Sturnick DR, DeSarno MJ, et al. 2014. Changes to the articular cartilage thickness profile of the tibia following anterior cruciate ligament injury. Osteoarthritis Cartilage 22:1453-1460.

189. Favre J, Erhart-Hledik JC, Blazek K, et al. 2017. Anatomically Standardized Maps Reveal Distinct Patterns of Cartilage Thickness With Increasing Severity of Medial Compartment Knee Osteoarthritis. Journal of Orthopaedic Research 35:2442-2451.

190. Shah RF, Martinez AM, Pedoia V, et al. 2019. Variation in the Thickness of Knee Cartilage. The Use of a Novel Machine Learning Algorithm for Cartilage Segmentation of Magnetic Resonance Images. J Arthroplasty 34:2210-2215.

191. Price LL, Harkey MS, Ward RJ, et al. 2019. Role of Magnetic Resonance Imaging in Classifying Individuals Who Will Develop Accelerated Radiographic Knee Osteoarthritis. Journal of Orthopaedic Research 37:2420-2428.

192. Eckstein F, Charles HC, Buck RJ, et al. 2005. Accuracy and precision of quantitative assessment of cartilage morphology by magnetic resonance imaging at 3.0T. Arthritis Rheum 52:3132-3136.

193. Sharma L, Eckstein F, Song J, et al. 2008. Relationship of meniscal damage, meniscal extrusion, malalignment, and joint laxity to subsequent cartilage loss in osteoarthritic knees. Arthritis Rheum-Us 58:1716-1726.

194. Wluka AE, Wolfe R, Stuckey S, et al. 2004. How does tibial cartilage volume relate to symptoms in subjects with knee osteoarthritis? Ann Rheum Dis 63:264-268.

195. Wluka A, Stuckey S, Snaddon J, et al. 2002. The determinants of change in tibial cartilage volume in osteoarthritic knees. Arthritis Rheum-Us 46:2065-2072.

196. Wluka A, Forbes A, Wang YY, et al. 2006. Knee cartilage loss in symptomatic knee osteoarthritis over 4.5 years. Arthritis Research & Therapy 8.

197. Raynauld JP, Martel-Pelletier J, Berthiaume MJ, et al. 2006. Long term evaluation of disease progression through the quantitative magnetic resonance imaging of symptomatic knee osteoarthritis patients: correlation with clinical symptoms and radiographic changes. Arthritis Research & Therapy 8.

198. Liukkonen MK, Mononen ME, Klets O, et al. 2017. Simulation of Subject-Specific Progression of Knee Osteoarthritis and Comparison to Experimental Follow-up Data: Data from the Osteoarthritis Initiative. Sci Rep 7:9177.

199. Jones G, Glisson M, Hynes K, et al. 2000. Sex and site differences in cartilage development: a possible explanation for variations in knee osteoarthritis in later life. Arthritis Rheum 43:2543-2549.

200. Emmanuel K, Quinn E, Niu J, et al. 2016. Quantitative measures of meniscus extrusion predict incident radiographic knee osteoarthritis–data from the Osteoarthritis Initiative. Osteoarthr Cartilage 24:262-269.

201. Wenger A, Wirth W, Hudelmaier M, et al. 2013. Meniscus body position, size, and shape in persons with and persons without radiographic knee osteoarthritis: quantitative analyses of knee magnetic resonance images from the osteoarthritis initiative. Arthritis & Rheumatism 65:1804-1811.

202. Peterfy CG, Schneider E, Nevitt M. 2008. The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. Osteoarthr Cartilage 16:1433-1441.

203. Iriondo C, Liu F, Caliva F, et al. 2020. Towards Understanding Mechanistic Subgroups of Osteoarthritis: 8 Year Cartilage Thickness Trajectory Analysis. J Orthop Res.

204. Qi CR, Yi L, Su H, et al. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. arXiv preprint arXiv:170602413.

205. Chen L-Z, Li X-Y, Fan D-P, et al. 2019. LSANet: Feature learning on point sets by local spatial aware layer. arXiv preprint arXiv:190505442.

206. Pedoia V, Lee J, Norman B, et al. 2019. Diagnosing osteoarthritis from T2 maps using deep learning: an analysis of the entire Osteoarthritis Initiative baseline cohort. Osteoarthr Cartilage 27:1002-1010.

207. Taghanaki SA, Hassani K, Jayaraman PK, et al. 2020. PointMask: Towards Interpretable and Bias-Resilient Point Cloud Processing. arXiv preprint arXiv:200704525.

208. Walker BF. 2000. The prevalence of low back pain: a systematic review of the literature from 1966 to 1998. J Spinal Disord 13:205-217.

209. Katz JN. 2006. Lumbar disc disorders and low-back pain: socioeconomic factors and consequences. J Bone Joint Surg Am 88 Suppl 2:21-24.

210. Delitto A, George SZ, Van Dillen L, et al. 2012. Low back pain. J Orthop Sports Phys Ther 42:A1-57.

211. Deyo RA, Mirza SK, Turner JA, et al. 2009. Overtreating chronic back pain: time to back off? J Am Board Fam Med 22:62-68.

212. Vergroesen PP, Kingma I, Emanuel KS, et al. 2015. Mechanics and biology in intervertebral disc degeneration: a vicious circle. Osteoarthritis Cartilage 23:1057-1070.

213. Adams MA, Roughley PJ. 2006. What is intervertebral disc degeneration, and what causes it? Spine (Phila Pa 1976) 31:2151-2161.

214. Panjabi MM. 2003. Clinical spinal instability and low back pain. J Electromyogr Kinesiol 13:371-379.

215. Pfirrmann CW, Metzdorf A, Zanetti M, et al. 2001. Magnetic resonance classification of lumbar intervertebral disc degeneration. Spine (Phila Pa 1976) 26:1873-1878.

216. Griffith JF, Wang YX, Antonio GE, et al. 2007. Modified Pfirrmann grading system for lumbar intervertebral disc degeneration. Spine (Phila Pa 1976) 32:E708-712.

217. Adams MA, Dolan P. 2012. Intervertebral disc degeneration: evidence for two distinct phenotypes. J Anat 221:497-506.

218. Stelzeneder D, Welsch GH, Kovacs BK, et al. 2012. Quantitative T2 evaluation at 3.0T compared to morphological grading of the lumbar intervertebral disc: a standardized evaluation approach in patients with low back pain. Eur J Radiol 81:324-330.

219. Michopoulou SK, Costaridou L, Panagiotopoulos E, et al. 2009. Atlas-based segmentation of degenerated lumbar intervertebral discs from MR images of the spine. IEEE Trans Biomed Eng 56:2225-2231.

220. Menezes-Reis R, Salmon CE, Carvalho CS, et al. 2015. T1rho and T2 mapping of the intervertebral disk: comparison of different methods of segmentation. AJNR Am J Neuroradiol 36:606-611.

221. Castro-Mateos I, Pozo JM, Lazary A, et al. 2014. 2D Segmentation of intervertebral discs and its degree of degeneration from T2-weighted magnetic resonance images. Proc Spie 9035.

222. Nagashima M, Abe H, Amaya K, et al. 2012. A method for quantifying intervertebral disc signal intensity on T2-weighted imaging. Acta Radiol 53:1059-1065.

223. Johannessen W, Auerbach JD, Wheaton AJ, et al. 2006. Assessment of human disc degeneration and proteoglycan content using T-1p-weighted magnetic resonance imaging. Spine 31:1253-1257.

224. Antoniou J, Epure LF, Michalek AJ, et al. 2013. Analysis of Quantitative Magnetic Resonance Imaging and Biomechanical Parameters on Human Discs With Different Grades of Degeneration. J Magn Reson Imaging 38:1402-1414.

225. Welsch GH, Trattnig S, Paternostro-Sluga T, et al. 2011. Parametric T2 and T2* mapping techniques to visualize intervertebral disc degeneration in patients with low back pain: initial results on the clinical use of 3.0 Tesla MRI. Skeletal Radiol 40:543-551.

226. Hwang D, Kim S, Abeydeera NA, et al. 2016. Quantitative magnetic resonance imaging of the lumbar intervertebral discs. Quant Imaging Med Su 6:744-755.

227. Takashima H, Takebayashi T, Yoshimoto M, et al. 2012. Correlation between T2 relaxation time and intervertebral disk degeneration. Skeletal Radiol 41:163-167.

228. Marinelli NL, Haughton VM, Munoz A, et al. 2009. T-2 Relaxation Times of Intervertebral Disc Tissue Correlated With Water Content and Proteoglycan Content. Spine 34:520-524.

229. Pedoia V, Gallo MC, Souza RB, et al. 2017. Longitudinal Study Using Voxel-Based Relaxometry: Association Between Cartilage T-1 rho and T-2 and Patient Reported Outcome Changes in Hip Osteoarthritis. J Magn Reson Imaging 45:1523-1533.

230. Pedoia V, Li XJ, Su F, et al. 2016. Fully automatic analysis of the knee articular cartilage T-1 rho relaxation time using voxel-based relaxometry. J Magn Reson Imaging 43:970-980.

231. Ben Ayed I, Punithakumar K, Garvin G, et al. 2011. Graph Cuts with Invariant Object-Interaction Priors: Application to Intervertebral Disc Segmentation. Information Processing in Medical Imaging 6801:221-232.

232. Neubert A, Fripp J, Engstrom C, et al. 2012. Automated detection, 3D segmentation and analysis of high resolution spine MR images using statistical shape models. Physics in Medicine and Biology 57.

233. Law MWK, Tay K, Leung A, et al. 2013. Intervertebral disc segmentation in MR images using anisotropic oriented flux. Med Image Anal 17:43-61.

234. Li X, Dou Q, Chen H, et al. 2018. 3D multi-scale FCN with random modality voxel dropout learning for Intervertebral Disc Localization and Segmentation from Multi-modality MR Images. Med Image Anal 45:41-54.

235. Zheng GY, Chu CW, Belavy DL, et al. 2017. Evaluation and comparison of 3D intervertebral disc localization and segmentation methods for 3D T2 MR data: A grand challenge. Med Image Anal 35:327-344.

236. Zuo J, Joseph GB, Li X, et al. 2012. In vivo intervertebral disc characterization using magnetic resonance spectroscopy and T1rho imaging: association with discography and Oswestry Disability Index and Short Form-36 Health Survey. Spine (Phila Pa 1976) 37:214-221.

237. Pandit P, Talbott JF, Pedoia V, et al. 2016. T1rho and T2 -based characterization of regional variations in intervertebral discs to detect early degenerative changes. J Orthop Res 34:1373-1381.

238. Milletari F, Rieke N, Baust M, et al. 2018. CFCM: Segmentation via Coarse to Fine Context Memory. Lect Notes Comput Sc 11073:667-674.

239. Klein S, Staring M, Murphy K, et al. 2010. elastix: A Toolbox for Intensity-Based Medical Image Registration. Ieee T Med Imaging 29:196-205.

240. Gong ZQ, Zhong P, Hu WD. 2019. Diversity in Machine Learning. Ieee Access 7:64323-64350.

241. Monu UD, Jordan CD, Samuelson BL, et al. 2017. Cluster analysis of quantitative MRI T-2 and T-1 rho relaxation times of cartilage identifies differences between healthy and ACL-injured individuals at 3T. Osteoarthr Cartilage 25:513-520.

242. Veldhuizen AG, Wever DJ, Webb PJ. 2000. The aetiology of idiopathic scoliosis: biomechanical and neuromuscular factors. European Spine Journal 9:178-184.

243. Konieczny MR, Senyurt H, Krauspe R. 2013. Epidemiology of adolescent idiopathic scoliosis. Journal of Childrens Orthopaedics 7:3-9.

244. Carter OD, Haynes SG. 1987. Prevalence Rates for Scoliosis in United-States Adults - Results from the 1st National-Health and Nutrition Examination Survey. Int J Epidemiol 16:537-544.

245. Schwab F, Dubey A, Gamez L, et al. 2005. Adult scoliosis: Prevalence, SF-36, and nutritional parameters in an elderly volunteer population. Spine 30:1082-1085.

246. Theroux J, Le May S, Hebert JJ, et al. 2017. Back Pain Prevalence Is Associated With Curve-type and Severity in Adolescents With Idiopathic Scoliosis: A Cross-sectional Study. Spine (Phila Pa 1976) 42:E914-E919.

247. Mac-Thiong JM, Transfeldt EE, Mehbod AA, et al. 2009. Can c7 plumbline and gravity line predict health related quality of life in adult scoliosis? Spine (Phila Pa 1976) 34:E519-527.

248. Terran J, Schwab F, Shaffrey CI, et al. 2013. The SRS-Schwab Adult Spinal Deformity Classification: Assessment and Clinical Correlations Based on a Prospective Operative and Nonoperative Cohort. Neurosurgery 73:559-568.

249. Tan KJ, Moe MM, Vaithinathan R, et al. 2009. Curve Progression in Idiopathic Scoliosis Follow-up Study to Skeletal Maturity. Spine 34:697-700.

250. Ascani E, Bartolozzi P, Logroscino CA, et al. 1986. Natural-History of Untreated Idiopathic Scoliosis after Skeletal Maturity. Spine 11:784-789.

251. Ferguson A. 1930. The study and treatment of scoliosis. South Med J 23:116–120.

252. Diab KM, Sevastik JA, Hedlund R, et al. 1995. Accuracy and applicability of measurement of the scoliotic angle at the frontal plane by Cobb's method, by Ferguson's method and by a new method. Eur Spine J 4:291-295.

253. Morrissy RT, Goldsmith GS, Hall EC, et al. 1990. Measurement of the Cobb angle on radiographs of patients who have scoliosis. Evaluation of intrinsic error. J Bone Joint Surg Am 72:320-327.

254. H A, Prabhu GK. 2012. Automatic quantification of spinal curvature in scoliotic radiograph using image processing. J Med Syst 36:1943-1951.

255. Pan YL, Chen QR, Chen TT, et al. 2019. Evaluation of a computer-aided method for measuring the Cobb angle on chest X-rays. European Spine Journal 28:3035-3043.

256. Zhang T, Li YF, Cheung JPY, et al. 2021. Learning-Based Coronal Spine Alignment Prediction Using Smartphone-Acquired Scoliosis Radiograph Images. Ieee Access 9:38287-38295.

257. Wu HB, Bailey C, Rasoulinejad P, et al. 2018. Automated comprehensive Adolescent Idiopathic Scoliosis assessment using MVC-Net. Medical Image Analysis 48:1-11.

258. Galbusera F, Niemeyer F, Wilke HJ, et al. 2019. Fully automated radiological analysis of spinal disorders and deformities: a deep learning approach. European Spine Journal 28:951-960.

259. Thong W, Parent S, Wu J, et al. 2016. Three-dimensional morphology study of surgical adolescent idiopathic scoliosis patient from encoded geometric models. European Spine Journal 25:3104-3113.

260. Nibali A, He Z, Morgan S, et al. 2018. Numerical Coordinate Regression with Convolutional Neural Networks. arXiv preprint arXiv:1801.07372.

261. Huang G, Liu Z, van der Maaten L, et al. 2017. Densely Connected Convolutional Networks. 30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017):2261-2269.

262. Yu F, Koltun V, Funkhouser T. 2017. Dilated Residual Networks. 30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017):636-644.

263. Bonanni PG. 2017. Contour and Angle-Function Based Scoliosis Monitoring: Relaxing the Requirement on Image Quality in the Measurement of Spinal Curvature. International Journal of Spine Surgery 11.

264. Wu H. BC, Rasoulinejad P., Li S. . 2017. Automatic Landmark Estimation for Adolescent Idiopathic Scoliosis Assessment Using BoostNet. Medical Image Computing and Computer Assisted Intervention − MICCAI 2017 Lecture Notes in Computer Science, vol 10433. .

265. Abid A, Abdalla A, Abid A, et al. 2019. Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. 2019 ICML Workshop on Human in the Loop Learning.

266. Gardner A, Archer J, Berryman F, et al. 2021. The resting coronal and sagittal stance position of the torso in adolescents with and without spinal deformity. Sci Rep 11:2354.

267. Pasha S, Flynn J. 2018. Data-driven Classification of the 3D Spinal Curve in Adolescent Idiopathic Scoliosis with an Applications in Surgical Outcome Prediction. Sci Rep 8:16296.

268. Menon KV, Kumar VPD, Thomas T. 2014. Experiments with a Novel Content-Based Image Retrieval Software: Can We Eliminate Classification Systems in Adolescent Idiopathic Scoliosis? Glob Spine J 4:13-20.

269. Nault ML, Mac-Thiong JM, Roy-Beaudry M, et al. 2014. Three-Dimensional Spinal Morphology Can Differentiate Between Progressive and Nonprogressive Patients With Adolescent Idiopathic Scoliosis at the Initial Presentation. Spine 39:E601-E606.

270. Donzelli S, Poma S, Balzarini L, et al. 2015. State of the art of current 3-D scoliosis classifications: a systematic review from a clinical perspective. J Neuroeng Rehabil 12.

271. QuinoneroCandela J, Sugiyama M, Schwaighofer A, et al. 2009. Dataset Shift in Machine Learning. Neural Inf Process S:1-229.

272. Kadoury S, Cheriet F, Labelle H. 2010. Self-Calibration of Biplanar Radiographic Images Through Geometric Spine Shape Descriptors. Ieee T Bio-Med Eng 57:1663-1675.

273. Gajny L, Ebrahimi S, Vergari C, et al. 2019. Quasi-automatic 3D reconstruction of the full spine from low-dose biplanar X-rays based on statistical inferences and image analysis. European Spine Journal 28:658-664.

274. Humbert L, De Guise JA, Aubert B, et al. 2009. 3D reconstruction of the spine from biplanar X-rays using parametric models based on transversal and longitudinal inferences. Med Eng Phys 31:681-687.

275. Deschenes S, Charron G, Beaudoin G, et al. 2010. Diagnostic Imaging of Spinal Deformities Reducing Patients Radiation Dose With a New Slot-Scanning X-ray Imager. Spine 35:989-994.

276. Somoskeoy S, Tunyogi-Csapo M, Bogyo C, et al. 2012. Accuracy and reliability of coronal and sagittal spinal curvature data based on patient-specific three-dimensional models created by the EOS 2D/3D imaging system. Spine Journal 12:1052-1059.

277. Ilharreborde B, Steffen JS, Nectoux E, et al. 2011. Angle measurement reproducibility using EOS three-dimensional reconstructions in adolescent idiopathic scoliosis treated by posterior instrumentation. Spine (Phila Pa 1976) 36:E1306-1313.

278. Chen K, Zhai X, Sun KQ, et al. 2021. A narrative review of machine learning as promising revolution in clinical practice of scoliosis. Ann Transl Med 9.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

Claudia Iriondo
Author Signature

6/6/2021
Date