Automatic muscle segmentation from MR images using deformable registration and deep learning approaches

William Harvey Henson

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

INSIGNEO institute for in silico Medicine
Department of Mechanical Engineering
Faculty of Engineering
The University of Sheffield

2023

Summary

One in four people in the UK currently have one or more musculoskeletal disorder, impacting both work and social lives of individuals with them and incurring a burden on the national health service. Musculoskeletal disorders can affect one or both of the skeletal or the muscular system and where there is a substantial understanding of the mechanisms underpinning skeletal disorders, far less is currently understood regarding disorders affecting the muscular system. The challenge hindering our understanding of the mechanisms underpinning muscle disorders lies in the difficulty in measuring the physiological status of muscle tissue.

Muscle disorders vary widely in many ways, such as the causes, muscles affected, rate of progression, and even the treatment strategies for these disorders. Not only do muscle disorders differ in these areas when comparing each one to the others, but also, people with specific muscle disorders respond to them in different ways. For these reasons, subject-specific, quantitative characterisation of the muscles within subject measured *in vivo* could enhance the current diagnosis and treatment strategies for muscle disorders. Moreover, quantitative tools to measure the response of the muscle tissue to new treatments for muscle disorders within clinical trials would grant a more informed analysis of the efficacy of treatments. Quantitative analysis of muscle tissue has not yet been adopted into clinical practice but could impact both our understanding and ability to treat people with muscle disorders.

The aim of this thesis was to build, test, and analyse methods to automatically characterise the muscles from medical imaging data. Four methods have been detailed and explored to address the limitations associated with the current gold standard approach used to characterise the muscles from medical images. The outcome of the thesis is a general and complete overview of existing and novel methods to characterise muscles from medical images.

Future work should analyse the methods presented in this thesis and adopt that which is best suited to their study. The motivation and ambition behind the thesis are that the studies presented facilitate future research seeking to understand muscle disorders in a quantitative manner. In the long-term, the work presented in this thesis could promote clinical adoption of computational tools for characterising muscle disorders, leading to enhanced diagnosis and monitoring of such disorders.

Acknowledgements

I would firstly like to acknowledge my supervisory team: Prof. Claudia Mazzà, Dr. Enrico Dall'Ara, and Dr. Xinshan Li.

In particular, I would like to thank Claudia for her continued professional and personal support. With her soft (and sometimes very stern) words of encouragement, her guidance has shaped me into a professional, confident, and rigorous young scientist with a critical and optimistic mind. I would specifically like to thank Enrico for his unwavering scientific influence and direction. He is one of the most brilliant scientists I have had the pleasure of working with, and my hope is that at least some of his knowledge and manner have rubbed off on me. Thirdly, I would like to acknowledge the scientific and personal aid that Shannon has selflessly offered within the last 6-9 months of my PhD journey.

I would like to thank my friends & colleagues, without whom, I would never have begun or made it through my PhD journey. I would like to thank my friends (you know who you are), who instilled me with confidence that I could complete this journey and maintained my optimism on the difficult days that any PhD student must combat. Also, I would like to thank my colleagues that have been there to encourage me when times have been hard. On a more positive note, I would like to thank you all for celebrating the many successes we have all had throughout all of our journeys.

I would like to extend specific appreciation to my Mother, Father, and Sister for their emotional support not just within my PhD journey, but all the time leading up to it. Without these three special people, I would not be in the position I am today, for too many reasons to list – So thank you all. And yes, Mum and Clo, I know that a PhD doesn't make you a real Doctor!

Finally, I would like to express my gratitude to my partner, Isobel, who has been there for me every single day over the past 3 years. The emotional strain of a PhD can be severe, but with her constant support and open ear, we have arrived at this point together and I'm excited for our next journey together which I hope will not cause her quite as many headaches!

Publications & contributions

Journal entries

In print

1) E. Montefiori, B. M. Kalkman, **W. H. Henson**, M. A. Paggiosi, E. V. McCloskey, C. Mazzà. "MRI-based anatomical characterisation of lower-limb muscles in older women", PLoS One (December 2020), doi: 10.1371/journal.pone.0242973

2) **W. H. Henson**, C. Mazzà, E. Dall'Ara. "Deformable image registration based on single or multi-atlas methods for automatic muscle segmentation and the generation of augmented imaging datasets", PLoS One (March 2023), doi: 10.1371/journal.pone.0273446

In preparation

1) **W. H. Henson** & L. Dowling, X. Li, C. Mazzà, E. Dall'Ara & J. Walsh. Tentative title: "Comparison of fat infiltration and muscle functional capacity in healthy, obese, and dynapenic abdominal obese adults".

2) **W. H. Henson**, C. Mazzà, X. Li, E. Dall'Ara. Tentative title: "Traditional and novel Deep learning-based muscle segmentation from MR images enhanced with deformable image registration-generated augmented imaging datasets".

Conference abstracts

Oral presentations

1) **W. H. Henson**, E. Dall'Ara, C. Mazzà. "Automatic muscle segmentation through enhanced elastic image registration", BioMedEng 2021, University of Sheffield, United Kingdom.

2) **W. H. Henson**, E. Dall'Ara, C. Mazzà. "Automatic muscle segmentation through deformable image registration", CMBBE 2021, University of Bonn, Germany.

3) **W. H. Henson**, C. Mazzà, E. Dall'Ara. "Automatic muscle segmentation with deformable image registration from MR images of human lower limbs and the generation of augmented MR imaging data", ESB 2022, University of Porto, Portugal.

4) Zhicheng L., **W. H. Henson**, C. Mazzà, E. Dall'Ara, Guo L. "Deep learning-based automatic segmentation of skeletal muscle", ESB 2023, Maastricht, the Netherlands.

Poster presentations

1) **W. H. Henson**, C. Mazzà, E. Dall'Ara. "Efficient muscle segmentation from MRI data: preliminary data using elastic image registration", CAMS KNEE 2020, ETH Zürich, Switzerland.

2) L. van Gelder, **W. H. Henson**, C. Mazzà. "Calculating autosymmetry based on accelerometer data", ISB 2021, Stockholm, Sweden.

3) **W. H. Henson**, E. Dall'Ara, C. Mazzà. "Automatic muscle segmentation with deformable image registration from MR images of human lower limb", INSIGNEO showcase 2022, University of Sheffield, United Kingdom.

Awards

1) 1st placed team, The MultiSim OATech+ Modelathon 2020, International workshop on musculoskeletal modelling, University of Sheffield.

2) 1st place, 3-minute thesis award 2021, Faculty of Engineering, University of Sheffield.

3) GTA award 2022 for "Student centred approach in teaching", Faculty of Engineering, University of Sheffield.

4) Recognised as one of three students for excellence in research within the Department of Mechanical Engineering 2022, University of Sheffield.

Content

Chapter 1 - General introduction and thesis overview

Chapter 2 - Background

Chapter 3 - Literature review and motivation

Chapter 4 - Optimizing a muscle segmentation pipeline using deformable image registration

Chapter 5 - Enhanced inter-subject registration and the generation of an augmented imaging database

Chapter 6 - Traditional and novel deep learning-based segmentation approaches

Chapter 7 - General discussion and conclusions

Nomenclature

MR	Magnetic Resonance
CT	Computerised Tomography
US	Ultrasound
EM	Electromagnetic
RF	Radio Frequency
EWGSOP19	European Working Group on Sarcopenia in Older People (2019)
2D	Two-dimensional
3D	Three-dimensional
CoV	Coefficient of Variation
MSK	Musculoskeletal
NMSK	Neuromusculoskeletal
BMI	Body Mass Index
SSM	Statistical Shape Modelling
ShIRT	Sheffield Image Registration Toolkit
NS	Nodal Spacing
f	Fixed image
m	Moving image
CNN	Convolutional Neural Network
ANOVA	Analysis of Variance

Declaration

I, William H. Henson, confirm that the work presented in this thesis is a product of my studies. Where information is portrayed to the reader that was derived from other sources, I confirm that this is indicated within the thesis.

Chapter 1:

General introduction and thesis overview

1.1. Introduction

Musculoskeletal disorders are becoming ever more numerous in modern society, due to a larger population reaching greater ages [1]. Estimates suggest that one in five people of working age in the United Kingdom (UK), have been diagnosed with one or more musculoskeletal disorders [2]. Such disorders are the current leading cause of disability in the UK, with chronic symptoms like joint pain [3] and muscle weakening [4], as well as increased risk of fall [5], and early mortality [4]. Medical imaging is capable of enhancing diagnostic regimes, qualitative and quantitative monitoring of disease progression, and the measurement of intervention strategies for musculoskeletal disorders [6-8]. Advancements in innovative imaging techniques and the availability of advanced imaging analysis tools has improved our understanding of musculoskeletal disorders and influenced the medical response to them [8, 9].

Medical image segmentation provides quantitative, spatially structured details of the inner anatomy of individuals, allowing specific biomarkers to be isolated and characterised [10, 11]. Many scientific breakthroughs within the clinical research domain have been presented using medical image segmentation [12-14]. For example, in 2020, Hollon et al. [15] presented a method to segment brain tumours intraoperatively, allowing near real time visualisation of brain tumour tissue. A second example lies in the characterisation of osteoarthritis, particularly osteoarthritis within the knee joint [16]. Medical image segmentation has been used to measure the volume of the cartilage (behaves as lubricant between two moving bones), and further, the change in volume over time within the knee joints of subjects [17]. Through studies such as these, quantitative investigations into the change in cartilage volume in response to different treatments is currently under investigation [18]. Thirdly, medical image analysis is heavily involved throughout the process of joint replacement, both in the planning of the surgery and its long-term assessment, which maximises the chance of success and longevity of the implant [19].

The musculoskeletal system is a multi-faceted system consisting of two major parts: the skeleton, where rigid components (bones) are connected by pivoting joints, and the muscles, which supply forces to the bones in order to produce movement. This system is responsible for allowing movement of the human body, and its performance therefore directly affects mobility and locomotion [4, 13], including a person's ability to work and carry out daily activities [2]. Through advancements in medical imaging analysis, our understanding of bone diseases such as osteoarthritis [20] and osteoporosis [21] have been greatly enriched. However, this cannot be said for disorders that affect the muscles [22], for example sarcopenia, the age-related degradation of skeletal muscle tissue [23-25]. Estimates by the European Working Group on Sarcopenia in Older Individuals (EWGSOP19) suggested that between 10% and 40% of individuals aged over 60 have either sarcopenia, or probable sarcopenia [26, 27]. This disease affects people who suffer from it, as they are more at risk of fall and therefore major bone fractures [23], with poor gait characteristics which can further impact normal function of the skeletal system [23, 27]. Sarcopenia also limits people's mobility compounding the issues listed and encouraging the advancement of those aforementioned disorders [26, 27]. The reason for the lack of clinical understanding of this disease and others alike can be attributed to the difficulty in characterising the muscles within the human body [27]. As suggested previously, medical image analysis has advanced significantly in recent years, bridging the gap between reality and computer visualisation for the inner anatomy, but our understanding of muscle disorders at the current time, have not advanced in line with other tissue groups [28].

Skeletal muscles enable all voluntary movements of the human body, and their structural health is essential for everyday life. Muscle characteristics such as volume, geometry and length, or the level of fat infiltration have been established to affect the functional capacity of individual muscles [29-31], and it is these characteristics that are altered as a symptomatic response to muscle disorders [31, 32]. However, isolating muscle properties from medical images is still a challenge [28]. Though, quantifying the characteristics of individual muscles in a subject specific manner through segmentation could provide insight into the areas of the muscular system that limit a person's capacity to perform movement tasks [28]. Further, muscle segmentation could have the capacity to enhance diagnosis, monitoring, or identify individuals at risk of certain muscle disorders through quantification of individual muscle characteristics.

Driven by the motivations outlined, the overarching aims of this thesis was to develop and test automatic approaches to segment muscles from MR images. Fulfilling this aim should guide future research to adopt new methods to capture muscle characteristics, with the ambition of catalysing future research into the effects and treatments of muscle disorders.

1.2. Thesis overview

The thesis consists of 7 chapters, as shown in Figure 1.1. Chapter 1 presents an overview of the topic, the motivation of the thesis, and introduces the general content of the thesis. Chapter 2 provides the technical background knowledge required to contextualise the later content of the thesis, such as, an explicit explanation of the lower limb anatomy, details of muscle function, introducing muscle disorders and how they affect muscle function, and detailing the process of collecting medical images. Chapter 3 first presents the uses and limitations of the gold standard approach used to segment the muscles. Thereafter, a summary of the current literature surrounding automatic muscle segmentation is presented, highlighting the gaps that should be addressed.

Chapters 4, 5, and 6 contain the methods, results and critical appraisal of tools developed here to automatically segment muscles from medical imaging data. Chapter 4 focuses on the development and initial testing of an automatic segmentation tool based on deformable image registration, first by segmenting one limb using the contralateral limb as the reference, and thereafter, segmenting one subject whilst using a different subject as the reference. Chapter 5 builds on this method, making use of morphological image processing techniques to boost the accuracy of this technique. Chapter 5 also uses a more advanced segmentation method named “multi-atlas” segmentation, to further enhance the segmentation pipeline. Finally in this chapter, deformable image registration method was used to generate unique “new/augmented” images in order to boost the number of datasets at our disposal. Chapter 6 explored deep learning-based methods, both traditional and novel, for individual muscle segmentation. Chapter 6 built on the knowledge and results (specifically the augmented image database), from Chapter 5. The final chapter, Chapter 7, summarises the findings, contributions, and limitations of the thesis, presenting an overall conclusion and recommendations for future work.

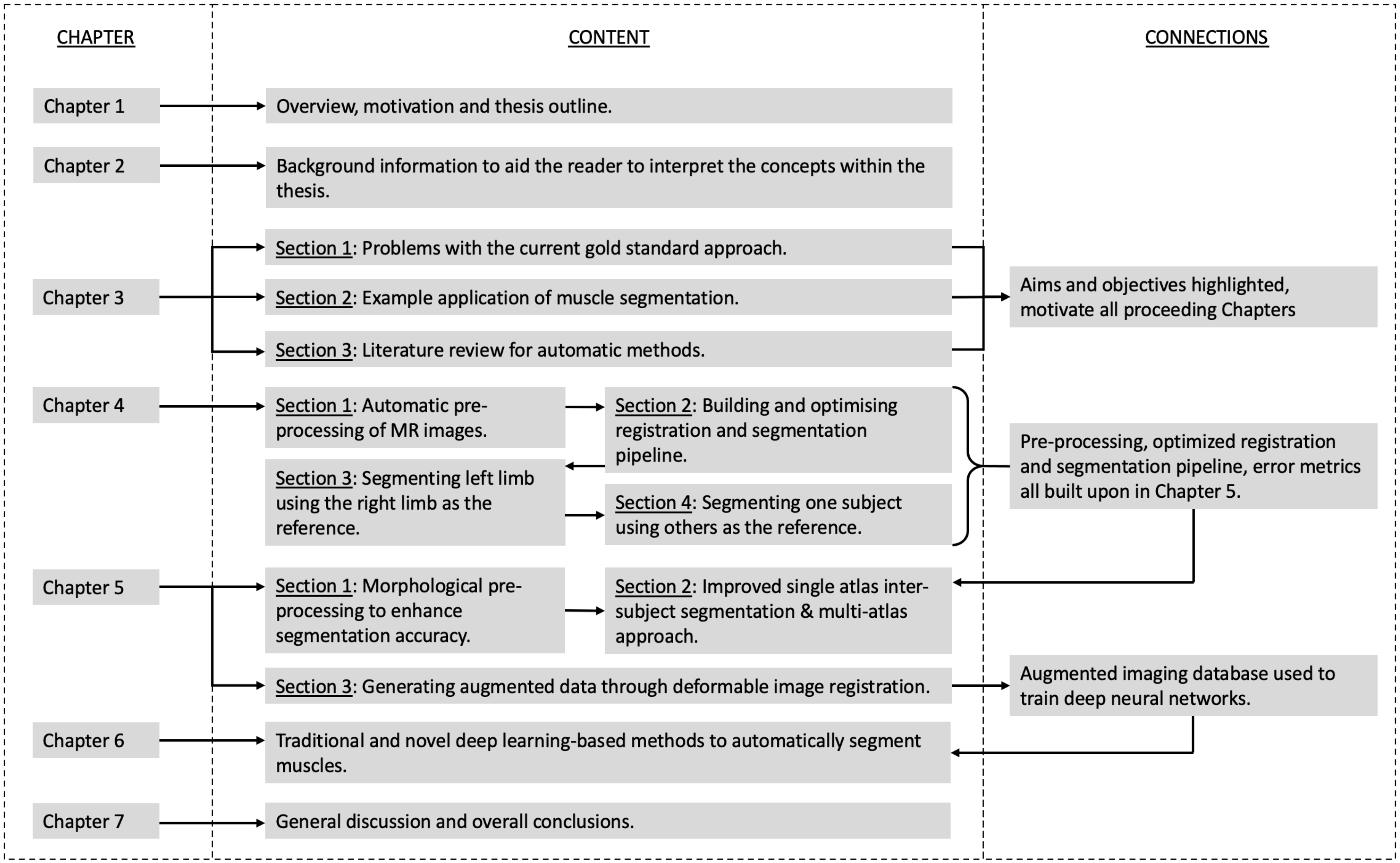

CHAPTER
CONTENT
CONNECTIONS
Chapter 1
Overview, motivation and thesis outline.
Chapter 2
Background information to aid the reader to interpret the concepts within the thesis.
Chapter 3
Section 1: Problems with the current gold standard approach.
Section 2: Example application of muscle segmentation.
Section 3: Literature review for automatic methods.
Aims and objectives highlighted, motivate all proceeding Chapters
Chapter 4
Section 1: Automatic pre-processing of MR images.
Section 2: Building and optimising registration and segmentation pipeline.
Section 3: Segmenting left limb using the right limb as the reference.
Section 4: Segmenting one subject using others as the reference.
Pre-processing, optimized registration and segmentation pipeline, error metrics all built upon in Chapter 5.
Chapter 5
Section 1: Morphological pre-processing to enhance segmentation accuracy.
Section 2: Improved single atlas inter-subject segmentation & multi-atlas approach.
Section 3: Generating augmented data through deformable image registration.
Augmented imaging database used to train deep neural networks.
Chapter 6
Traditional and novel deep learning-based methods to automatically segment muscles.
Chapter 7
General discussion and overall conclusions.

Figure 1.1: Schematic representation of thesis structure highlighting the main contents of each chapters and their inter-connections.

Chapter 2:

Background

2.1. The lower-limb musculoskeletal system

The musculoskeletal system is the combination of the human skeletal and muscular systems. The skeletal system is comprised of rigid components connected by pivoting joints, and the muscular system supplies force to the bones through contraction to produce movement. Through this mechanism, these subsystems operate in tandem to enable movement of the human body.

Three sections of the human body will be focussed on in this work: the hip, thigh, and shank. The main bones contained within these sections are the pelvis, femur, patella, tibia, and fibula. The femoral head sits within the pelvis in a ball-and-socket joint, allowing motion of the hip, explicitly: flexion/extension, abduction/adduction, and internal/external rotation. The femur is connected to the tibia by a hinge joint, allowing the extension and flexion of the knee. The knee joint also allows for a small amount of abduction/adduction, internal/external rotation, and translation in the three spatial directions to add stability and dampen large forces acting through the knee [33]. The bones that lie within the region explored within this thesis are presented in Figure 2.1 below, wherein the bones and joints are labelled. The 37 muscles that lie within the legs, encircling these bones and allowing the motions outlined previously are also presented in Figure 2.1, showing the relative size and anatomical location of these muscles, within the human body. There are some stabilising muscles within the feet, but these are not considered in the thesis. Additionally, the muscles that are explored in this thesis are summarised in Table 2.1, where the bones that they are attached to and motion(s) that they allow are explicitly noted.

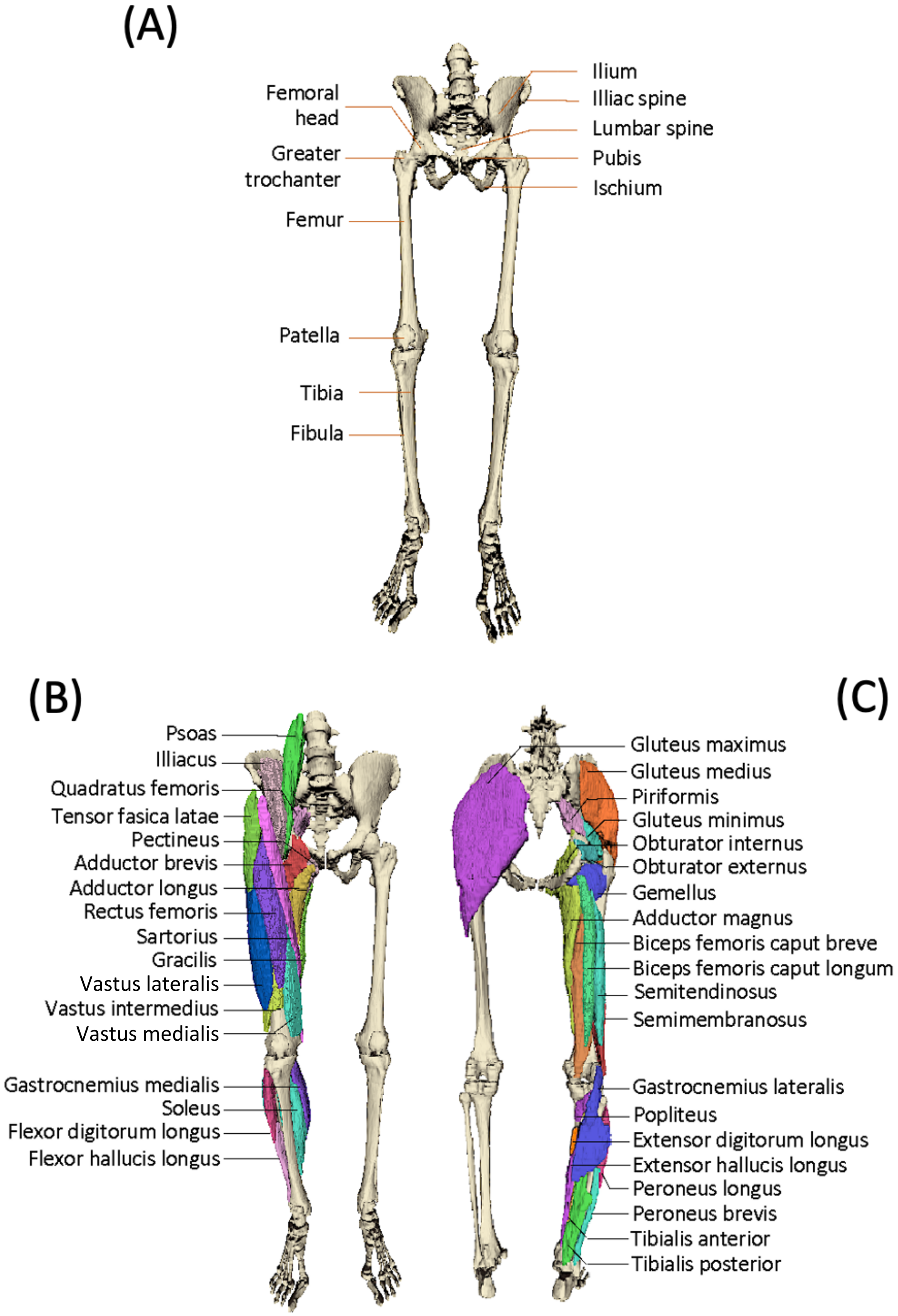
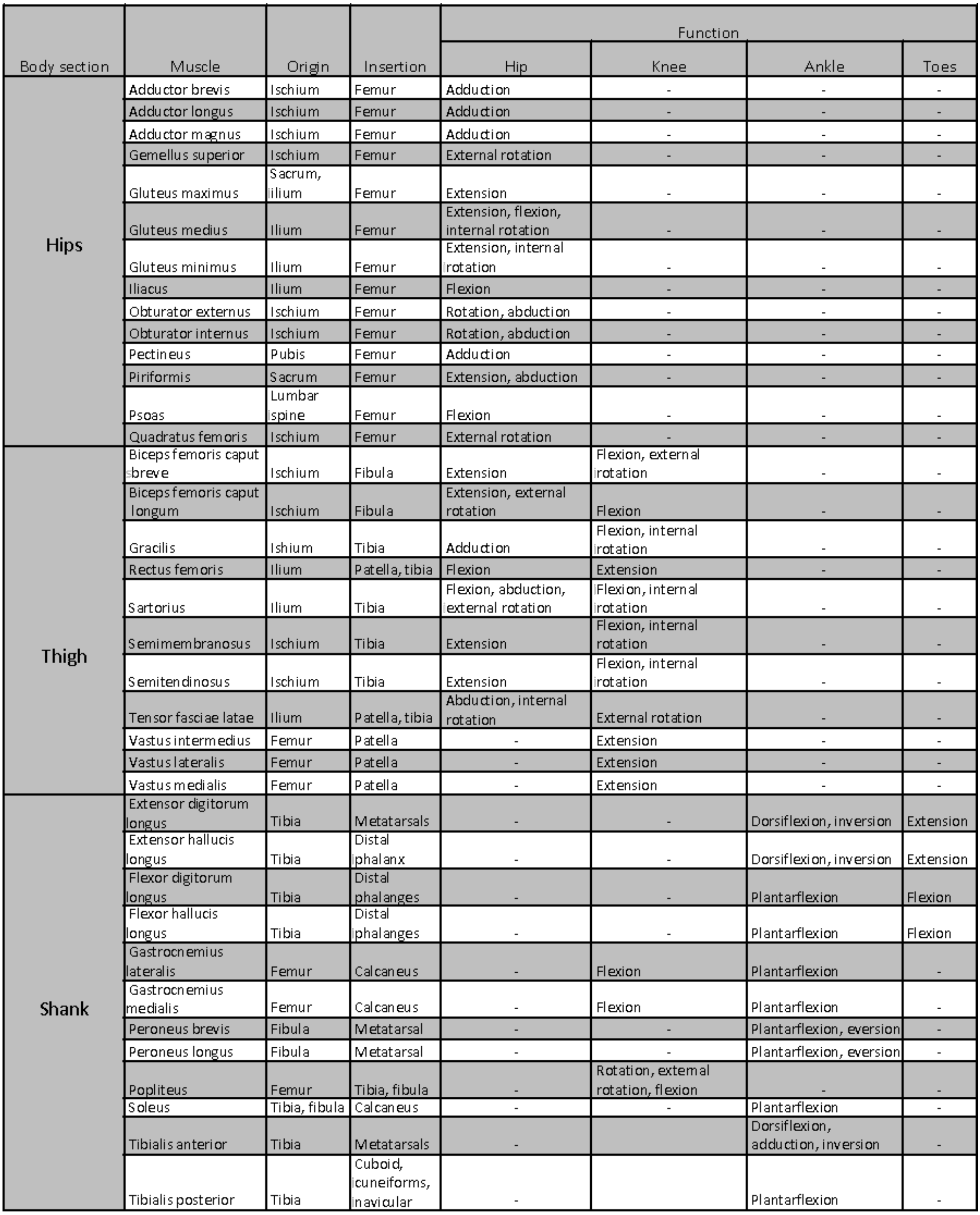


Figure 2.1: The lower limb bones (A) and muscles (B, C).



Body section	Muscle	Origin	Insertion	Function			
				Hip	Knee	Ankle	Toes
Hips	Adductor brevis	Ischium	Femur	Adduction	-	-	-
	Adductor longus	Ischium	Femur	Adduction	-	-	-
	Adductor magnus	Ischium	Femur	Adduction	-	-	-
	Gemellus superior	Ischium	Femur	External rotation	-	-	-
	Gluteus maximus	Sacrum, ilium	Femur	Extension	-	-	-
	Gluteus medius	Ilium	Femur	Extension, flexion, internal rotation	-	-	-
	Gluteus minimus	Ilium	Femur	Extension, internal rotation	-	-	-
	Iliacus	Ilium	Femur	Flexion	-	-	-
	Obturator externus	Ischium	Femur	Rotation, abduction	-	-	-
	Obturator internus	Ischium	Femur	Rotation, abduction	-	-	-
	Pectineus	Pubis	Femur	Adduction	-	-	-
	Piriformis	Sacrum	Femur	Extension, abduction	-	-	-
	Psoas	Lumbar spine	Femur	Flexion	-	-	-
	Quadratus femoris	Ischium	Femur	External rotation	-	-	-
Thigh	Biceps femoris caput breve	Ischium	Fibula	Extension	Flexion, external rotation	-	-
	Biceps femoris caput longum	Ischium	Fibula	Extension, external rotation	Flexion	-	-
	Gracilis	Ishium	Tibia	Adduction	Flexion, internal rotation	-	-
	Rectus femoris	Ilium	Patella, tibia	Flexion	Extension	-	-
	Sartorius	Ilium	Tibia	Flexion, abduction, external rotation	Flexion, internal rotation	-	-
	Semimembranosus	Ischium	Tibia	Extension	Flexion, internal rotation	-	-
	Semitendinosus	Ischium	Tibia	Extension	Flexion, internal rotation	-	-
	Tensor fasciae latae	Ilium	Patella, tibia	Abduction, internal rotation	External rotation	-	-
	Vastus intermedius	Femur	Patella	-	Extension	-	-
	Vastus lateralis	Femur	Patella	-	Extension	-	-
	Vastus medialis	Femur	Patella	-	Extension	-	-
Shank	Extensor digitorum longus	Tibia	Metatarsals	-	-	Dorsiflexion, inversion	Extension
	Extensor hallucis longus	Tibia	Distal phalanx	-	-	Dorsiflexion, inversion	Extension
	Flexor digitorum longus	Tibia	Distal phalanges	-	-	Plantarflexion	Flexion
	Flexor hallucis longus	Tibia	Distal phalanges	-	-	Plantarflexion	Flexion
	Gastrocnemius lateralis	Femur	Calcaneus	-	Flexion	Plantarflexion	-
	Gastrocnemius medialis	Femur	Calcaneus	-	Flexion	Plantarflexion	-
	Peroneus brevis	Fibula	Metatarsal	-	-	Plantarflexion, eversion	-
	Peroneus longus	Fibula	Metatarsal	-	-	Plantarflexion, eversion	-
	Popliteus	Femur	Tibia, fibula	-	Rotation, external rotation, flexion	-	-
	Soleus	Tibia, fibula	Calcaneus	-	-	Plantarflexion	-
	Tibialis anterior	Tibia	Metatarsals	-		Dorsiflexion, adduction, inversion	-
	Tibialis posterior	Tibia	Cuboid, cuneiforms, navicular	-		Plantarflexion	-

Table 2.1: 37 major lower limb muscles partitioned into the three sections of the body under investigation in this thesis. The origin, insertion and function of each muscle is stated explicitly.

2.1.1. Muscle tissue and function

There are three types of muscle tissue within the human body: smooth, cardiac and skeletal muscle tissue. Smooth muscle tissue exists within the walls of hollow organs, such as within the stomach, bladder, and intestines. Additionally, this muscle tissue is found within blood vessels, urinary tract, and respiratory system [34]. Smooth muscle is involuntary, non-striated, and is therefore not capable of producing significant force through contraction, but enables and assists normal bodily functions, such as the constriction and dilation of the blood vessels [34]. Cardiac muscle tissue, on the other hand, is striated and exists only within the heart. Cardiac muscle tissue is designed for rhythmic contraction, requires no voluntary control, and does not experience fatigue [35]. Skeletal muscle, distinct from both smooth and cardiac muscle tissue, is the most abundant type of muscle tissue within the human body, comprising approximately 50% of body weight. Skeletal muscle tissue is innervated, striated, voluntary, and experiences fatigue with high energy requirements.

2.1.1.1. Skeletal muscle

The body of skeletal muscle itself can be assessed at increasing levels of magnification, from the whole muscle down to a subcellular level. Each sub-component of muscle tissue at increasing magnification are grouped into fractal-style bundles wherein each child structure appears like a replica of the parent structure but at smaller scale (Figure 2.2). This repeating, fractal like structure is similar to steel wire rope, an engineered product designed to withstand exceptional tensile and bending stresses. Skeletal muscle tissue is highly organised in this repeating fashion to exhibit similar desirable material properties. Skeletal muscle is required by the body to generate tensile force and does so using contractile mechanisms on a subcellular scale.

These repeating structures are important for muscle force generation but require sophisticated and targeted imaging methods to be visualised. The macro-structure of the muscle: volume, length, geometry, and fat infiltration, are important factors for muscle force generation [28, 30, 31]. Changes in these characteristics are often responses to changes in the micro-structure of the muscles [36, 37].

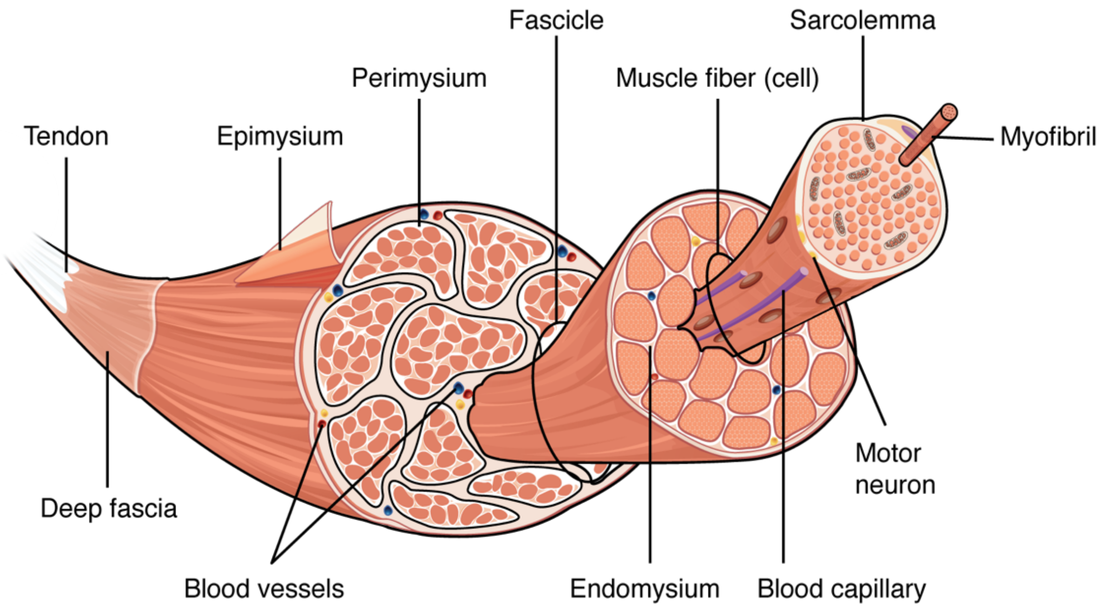


Figure 2.2: The transition from macro (left) to subcellular-structure (right) of the muscle tissue [38].

2.1.1.2. Force-length relationship

There are three types of contraction that muscle tissue is capable of: isometric, concentric, and eccentric. Isometric contractions stabilise joints, maintaining one position in response to joint loading, such as when holding a weight at constant position. Concentric contractions occur upon shortening of the muscle, such as the motion involved in a bicep curl. Finally, Eccentric contractions involve lengthening of a contracted muscle and are used to decelerate or control load. The muscle force generating capacity of these three different types of contraction in order, from highest to lowest, are: eccentric, isometric, then concentric. Figure 2.3 further describes this relationship. Initially, the sarcomere length has an increasing relationship with the tensile force achievable, reaching a maximum at the optimal point where the thick filaments are all able to attach to the thin filaments. Thereafter, the tensile force decreases as the sarcomere length increases, due to a reduced number of attachment locations or cross bridges, until the amount of overlap reduces to zero. This relationship has been proven experimentally within vertebrates [39, 40] and it is accepted that skeletal muscle acts in the same manner across all vertebrates [41].

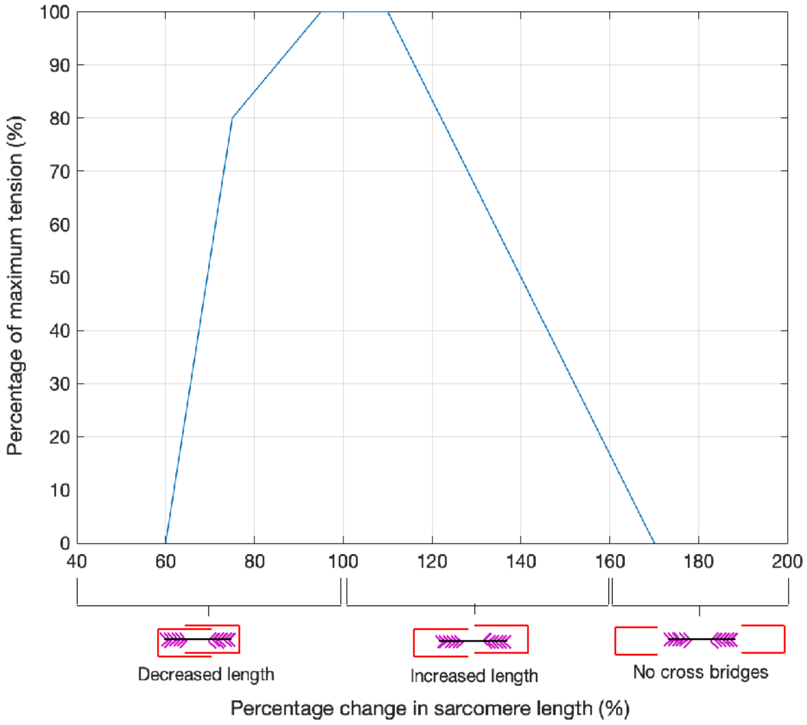


Figure 2.3: The force-length relationship of sarcomere elements within muscle tissue found experimentally within vertebrates by Gordon, Huxley and Julian 1966 [40].

2.1.1.3. Force-velocity and power-velocity relationship

The speed at which a muscle changes length impacts the force and power output of the contraction (Figure 2.4). The tensile force generated by a contraction depends on the number of cross bridges formed between thin and thick filaments within sarcomeres [40]. Though, the formation of cross bridges does not occur immediately upon contraction, meaning that if filaments slide over one another at a faster rate, the creation of cross bridges is reduced. Therefore, with faster contraction, the tensile force output of a muscle is reduced. At maximum velocity, no cross-bridges can form, meaning zero force is generated and no power is produced. Opposingly, when the muscle is contracted at minimal velocities, maximum force is generated but with such low speed, no power is produced. Maximum power is produced at approximately one third of maximum contractile velocity. These relationships were established through *in vivo* experiments by Edman et al. on a variety of vertebrates [39].

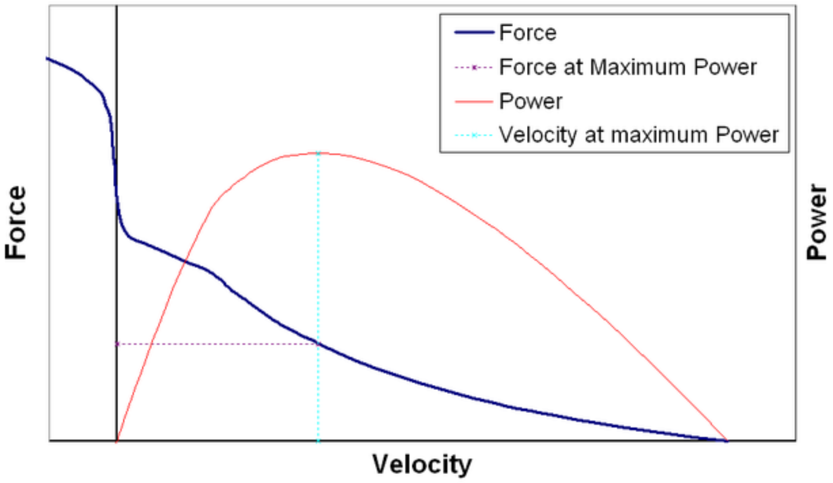


Figure 2.4: Force-velocity and velocity-power relationship of muscle contractions, showing the force at maximum power and velocity at maximum power.

2.1.1.4. Neurological control

Skeletal muscles contract when given a nerve impulse originating from motor neurons within either the brain, brain stem (upper motor neurons), or the spinal cord (lower motor neurons). As Figure 2.5 shows, motor neurons send information directly to muscle fibres through axons, which carry signals to designated areas within muscles. Muscle fibres contract when a nerve impulse is transmitted to them at a highly specialised contact points within the fibre, named neuromuscular junctions. The neuromuscular junction connects the terminal end of a motor neuron to many muscle fibres through unmyelinated terminal branches (unmyelinated meaning lacking a myelin sheath, a nerve coating allowing fast and efficient transmission of electrical impulses), as shown in Figure 2.5. Within voluntary muscle, the nerve impulse originates either in the brain or in the brain stem, enabling neural control. At the neuromuscular junctions, nerve action potential (like electrical potential) triggers the release of chemical transmitters that cause specialised proteins surrounding the muscle fibre to initiate contraction of the muscle fibres. The neuromuscular junction is a part of the nervous system that is prone to disease. Deficiencies within the neuromuscular junction are the result of many NMSK disorders, which are described in further detail in Section 2.2.

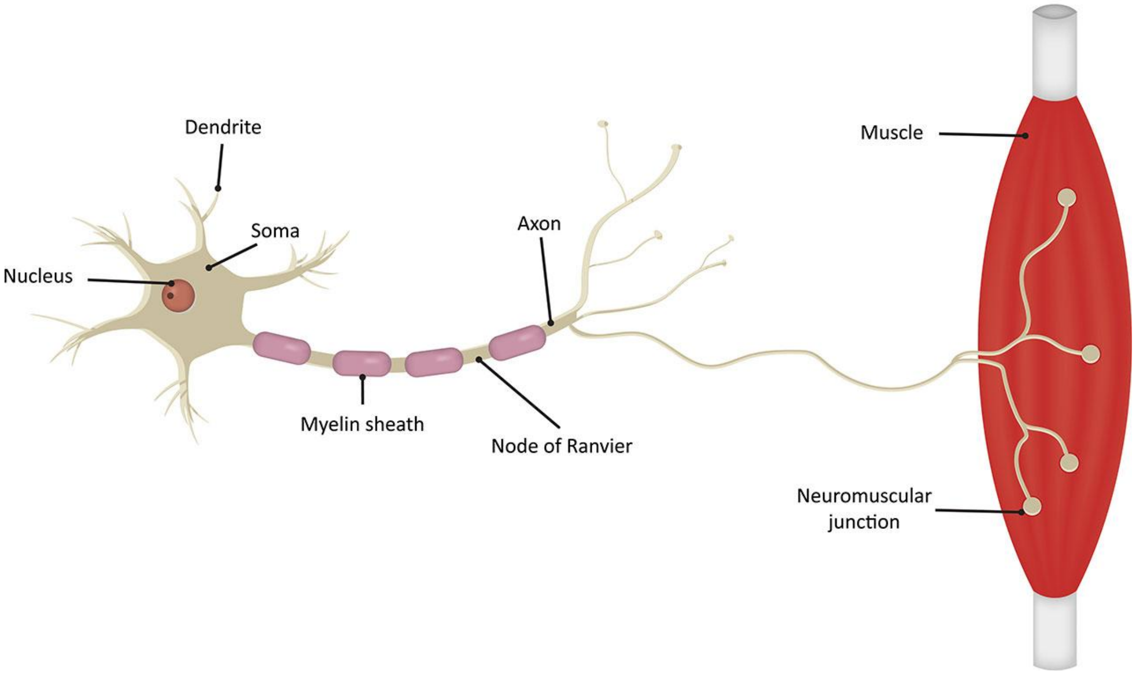


Figure 2.5: Schematic of the neurological connection between the central nervous system and muscle tissue. The dendrite which lies within the central nervous system transmits nerve impulses to the muscle fibres through axons which branch off and connect to many muscle fibres through the neuromuscular junction.

2.1.1.5. Force application to the skeleton

The body of skeletal muscles is typically connected to one or more bones (depending on the function of the muscle) by tendons, a dense connective tissue made mostly of collagen. Tendons are mechanically tough, with relatively high-tensile-strength and viscoelastic; hence they are well designed to repetitively transfer large amounts of force between connected components, i.e., from muscle to bone. The interfaces between tendons and muscles are called myotendinous junctions, while the interfaces between tendons and bones are called entheses, highlighted as TBI in Figure 2.6. The soft tissue of the muscle gradually forms into the stiff tissue of tendon via the myotendinous junction, extending the length of the transition between the two tissues and increasing the strength of the bond between the two tissues. The entheses, like the myotendinous junctions, are areas of transition between the collagen rich tendons and calcium rich bones wherein the tendinous tissue gradually becomes calcified.

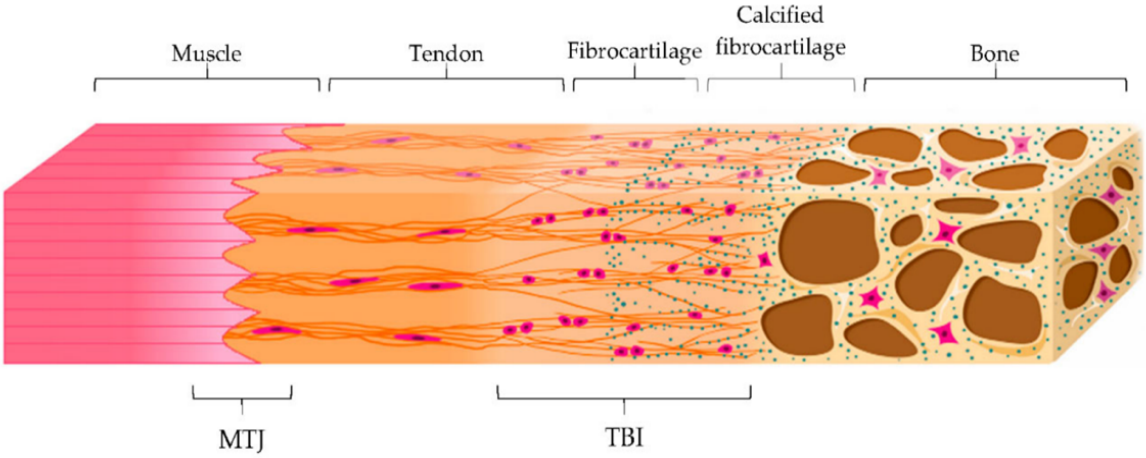


Figure 2.6: The transition from muscle tissue, to tendon and bone. The muscle tissue transitions to tendinous tissue at the myotendinous junction (MTJ). The tendon then transitions to bone through gradual calcification at the tendon-to-bone interface (TBI). Image acquired from Bianchi et al [42].

The architecture of muscles is dictated by the direction of contraction of muscle fibres relative to the axis of force-generation, which determines the characteristics of force transmission from the contractile muscle tissue to the bones. There are two main types of muscle architecture: parallel, and pennate. Parallel muscle architectures are those that the muscle fibres are organised parallel to the force generating axis. This architecture of muscle is well suited to fast contractions, or those that are required to act over a large range of motion, such as the sartorius, the longest muscle in the human body. Additionally, parallel muscles can appear in three sub-categories: strap (sartorius), fusiform (biceps femoris caput breve and longum), or fan-shaped (gluteus maximus). Pennate muscles, on the other hand, are those where the fibres are organised at an angle to the force generating axis. These muscles can generate more force than parallel muscles, as the muscles contain a greater number of fibres. Pennate muscles like parallel muscles, can be divided into sub-categories: unipennate (extensor digitorum longus), bipennate (rectus femoris), and multipennate (deltoids). Though there are no multipennate muscles within the lower limbs, it has been noted for completion. Interpretations of all noted muscle architectures are presented in Figure 2.7 below.

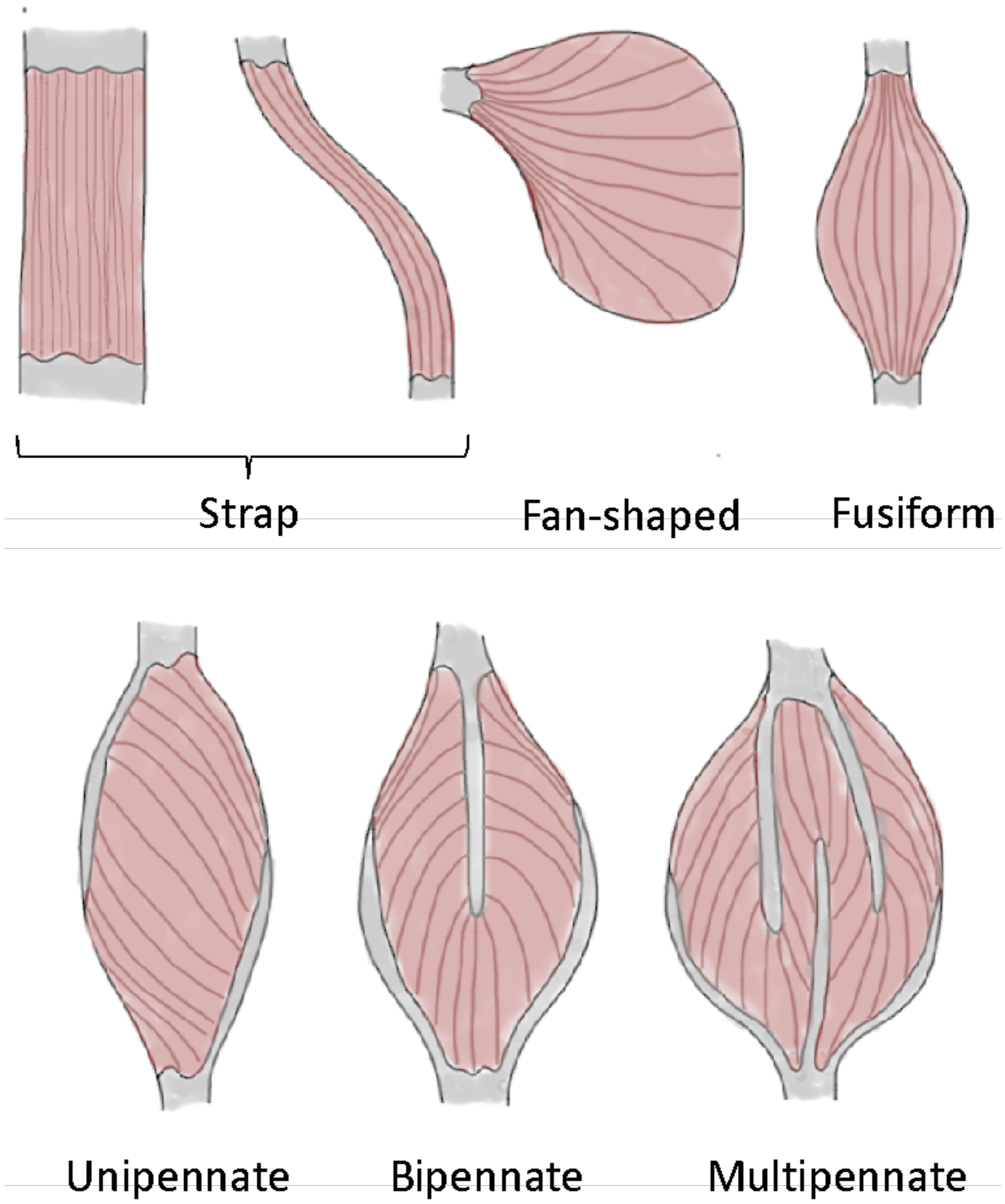


Figure 2.7: The six skeletal muscle architectures present within the lower limb muscles assessed in this thesis. The top and bottom row represent the 3 sub-structures of parallel muscles, and the 3 sub-structures of pennate muscles, respectively.

The collaborative function of the muscular, skeletal, and nervous systems creates movement of the human body [43]. Through neural control, either conscious or sub-conscious, muscles contract, generating tension which is enacted upon the skeleton. As forces are applied to the skeleton, the rigid components (bones) of the skeleton are rotated about one another at articulated joints. With simple movement such as standing or walking, many joints pivot simultaneously, requiring the fine neural control of many muscles concurrently [44].

2.1.1.6. Hypertrophy, atrophy, and fatigue

Muscles are adaptive, living tissue that react to stimulation, or lack thereof. Stimulation of the muscles can occur in response to two different types of exercise: aerobic, or anaerobic. Muscles require a constant flow of oxygenated blood to fuel movement and to remove lactic acid and other waste products that are produced upon activation of the muscle. Where the flow of blood is sufficient to supply the required amount of oxygen, and remove waste material, the stimulus of the muscle results in aerobic exercise (examples are walking, cycling, or jogging). Conversely, in cases wherein the flow of blood is not sufficient, the result is anaerobic exercise (examples are weightlifting or sprinting). Strength training, typically involving anaerobic exercise, triggers an adaptive response in the muscles where the muscle fibres increase in volume. The exact mechanisms that underpin this adaptation in muscle tissue, though they are known to be both neurological and muscular, are not currently well understood. However, it is widely known that through progressively overloading the muscles, one's physical strength can be greatly increased. This response of the muscle tissue is called hypertrophy. Note that aerobic exercises are not linked with muscle hypertrophy and do not result in the increase in muscle volume, but rather strengthen neural links with muscles and improve the supply of oxygenated blood, enhancing the removal of waste products from the muscle. Muscle atrophy is the opposite adaptive response of muscle and describes the loss of skeletal muscle volume. Atrophy is caused mainly by long periods of immobility but can be attributed to other factors, such as MSK disorders. Many musculoskeletal and neuromusculoskeletal disorders result in muscle atrophy, acting under a wide variety of physiological mechanisms, but the result is consistent: muscle weakness and a reduced physical capacity. Some of these mechanisms will be explored in the following sections.

Muscle fatigue is another response that muscle tissue has upon specific types of activation. Anaerobic exercise results in the build-up of lactic acid, causing a feeling of pain within the muscle. Where it was once believed that the build-up of lactic acid was the cause of muscle fatigue, this assumption is now under question within the biochemical research community [45]. The reduction in performance could be attributed to neural fatigue, where the nerve signals weaken when the muscle is required to perform powerful contractions near the limit of the muscle's ability to generate force. Secondly, muscle fatigue can be attributed to the shortage of substrates, or fuel, to the activated muscle fibres, or the accumulation of metabolites, which hinder the activation of individual sarcomeres [46].

2.2. Musculoskeletal and neuromusculoskeletal disorders

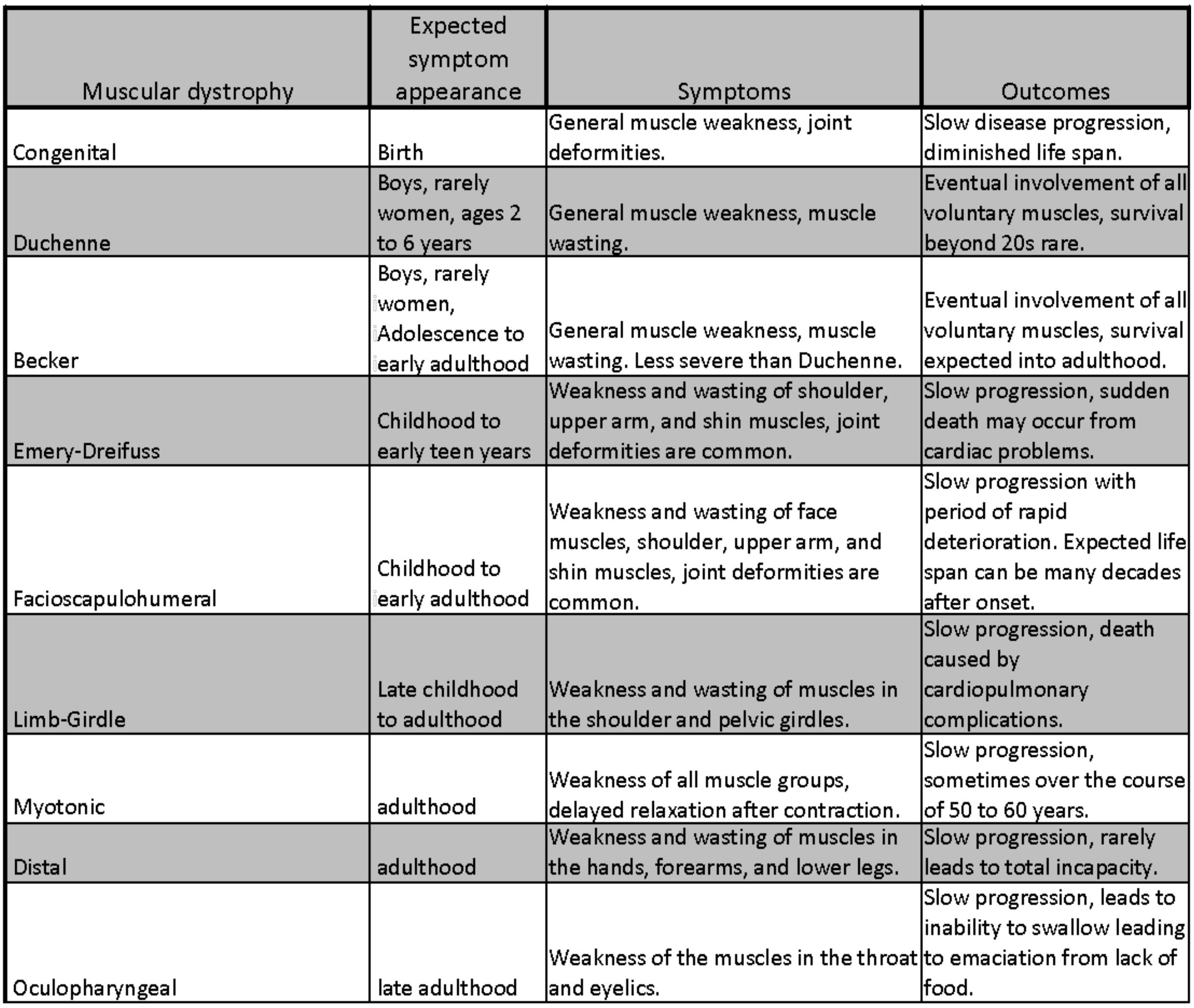
Musculoskeletal and neuromusculoskeletal disorders are injuries or disorders affecting the muscles, nerves, tendons, joints, and other areas such as cartilage (the lubricant present between connected bones), or spinal disks. Many (if not all) MSK and NMSK disorders affect one or more of these areas, as they are all intrinsically linked to one another. The list of known MSK and NMSK disorders is exceptionally long, and so in this section the focus will remain on the most common muscle disorders that affect the lower limb musculoskeletal system.

2.2.1. Muscle disorders

Muscle disorders cause muscle weakness as a direct result of dysfunctional muscle fibres. Muscle disorders can arise from inherited genetic mutations from one or both parents, known as muscular dystrophies (MDs), or they can be acquired within life, characterised as muscle atrophy. They differ in the muscles affected, disease mechanism, cause, rate of progression, and physiology, such as age and sex. The details of each of the 9 types of muscular dystrophy are elaborated on in Table 2.2, with a visual aid of the areas of the body typically affected in Figure 2.8. The most well-known MD within this group is Duchenne muscular dystrophy affecting (typically) males in the developmental stage. The prevalence of this disease in the UK has been reported between 2-11 per 100,000 people [47]. Duchenne MD is a genetic disorder. It is characterised by progressive degeneration and atrophy within the upper body muscle groups and the muscles within the thigh in response to a genetic mutation of a particular gene within the X chromosome, hence the significantly higher prevalence within males [47]. As a compensatory mechanism, children with Duchenne muscular dystrophy are visualised to have hypertrophy within the calves, to maintain normal healthy movement as much as possible. Children and people with Duchenne muscular dystrophy, up until very recently, did not typically survive beyond their teen years due to the degradation of the cardiac and respiratory systems. Fortunately, with advances in the capacity for care in these areas, life expectancy can now be extended into the 20s and 30s. There is no known treatment or cure for this form of MD though many clinical trials are being explored at present. One other, significantly less severe genetic muscle disorder noted in Table 2.2 is myotonic dystrophy, affecting 11 per 100,000 people (in Northern England) [48]. Myotonic dystrophy can affect almost any muscle group in the body, based on the genetic mutation that causes the disorder [48]. Myotonic dystrophy prevents the relaxation of muscle after contraction, causing damage to the muscle fibres. Given that in most cases, MDs are genetic disorders, they currently have no cure, but management strategies reduce the life-threatening aspect of these disorders, as they begin to affect the respiratory and cardiac systems [48, 49].

Acquired muscle disorders, on the other hand, are those that arise not due to genetics, but are developed due to other factors. The common physiological symptom of acquired muscle disorders is muscle atrophy, a loss of skeletal muscle mass [50]. Skeletal muscle acts as a storage unit for amino acids (a group of proteins) used in energy production in the case that demand is high or when supply is low. Muscle atrophy occurs when demand for these amino acids stored within the muscles outweighs the ability to synthesise these proteins, the functional muscle mass is lost [50, 51]. The mechanism for this muscle loss depends directly on the cause of the imbalance between synthesis and demand of these amino acids. Explicitly, there are two scenarios that result in imbalance: the demand for amino acids exceeding the rate of synthesis, or secondly, the rate of synthesis being reduced [50].

There are a great number of acquired muscle disorders that occur for a multitude of reasons. To highlight one that is likely to affect the cohort focussed on throughout this thesis: sarcopenia is the age-related degradation of skeletal muscle [27, 52, 53]. The prevalence of sarcopenia is reported by the European Working Group on Sarcopenia in Older People (EWGSOP) as 22.6% and 26.8% in women and men respectively [27], considering individuals above the age of 60. Where the prevalence is significantly increased to 31.0% and 52.9% (in men and women respectively) when considering individuals above the age of 80 [27]. Sarcopenia is characterised by degenerative loss of muscle mass, muscle quality, and functional strength, particularly within the lower limb muscle groups. The outcomes, therefore, of this disease are related to an increased risk of fall incurring greater risk of fracture, and a decline in functional capacity [27]. Sarcopenia is currently recommended to be diagnosed through (amongst other methods) measurement of grip strength by the EWGSOP2 (the second iteration of the EWGSOP) [27]. Though, the physiological symptoms, reported by EWGSOP2, of this disease are muscle atrophy and a decrease in muscle quality [25, 53]. The muscle quality referred to is the replacement of functional muscle fibres with fat deposits, causing fatty infiltration of the muscle (myosteatosis) [31]. In turn, the desaturation of functional muscle tissue reduces the functional capacity of the muscles and therefore limits mobility of individuals with sarcopenia. Furthermore, this is a degenerative disease, and so with time, the severity of muscle atrophy and myosteatosis worsens. Management strategies for sarcopenia centre around targeted exercise of the muscles affected and an increase in dietary protein included in the diet, to stimulate hypertrophy and protein synthesis, combatting atrophy and myosteatosis [54].



Muscular dystrophy	Expected symptom appearance	Symptoms	Outcomes
Congenital	Birth	General muscle weakness, joint deformities.	Slow disease progression, diminished life span.
Duchenne	Boys, rarely women, ages 2 to 6 years	General muscle weakness, muscle wasting.	Eventual involvement of all voluntary muscles, survival beyond 20s rare.
Becker	Boys, rarely women, Adolescence to early adulthood	General muscle weakness, muscle wasting. Less severe than Duchenne.	Eventual involvement of all voluntary muscles, survival expected into adulthood.
Emery-Dreifuss	Childhood to early teen years	Weakness and wasting of shoulder, upper arm, and shin muscles, joint deformities are common.	Slow progression, sudden death may occur from cardiac problems.
Facioscapulohumeral	Childhood to early adulthood	Weakness and wasting of face muscles, shoulder, upper arm, and shin muscles, joint deformities are common.	Slow progression with period of rapid deterioration. Expected life span can be many decades after onset.
Limb-Girdle	Late childhood to adulthood	Weakness and wasting of muscles in the shoulder and pelvic girdles.	Slow progression, death caused by cardiopulmonary complications.
Myotonic	adulthood	Weakness of all muscle groups, delayed relaxation after contraction.	Slow progression, sometimes over the course of 50 to 60 years.
Distal	adulthood	Weakness and wasting of muscles in the hands, forearms, and lower legs.	Slow progression, rarely leads to total incapacity.
Oculopharyngeal	late adulthood	Weakness of the muscles in the throat and eyelics.	Slow progression, leads to inability to swallow leading to emaciation from lack of food.

Table 2.2: Summary of the 9 main muscular dystrophies, with their expected onset, symptoms, and outcomes summarised. Table adapted from John Hopkins medicine library [55].

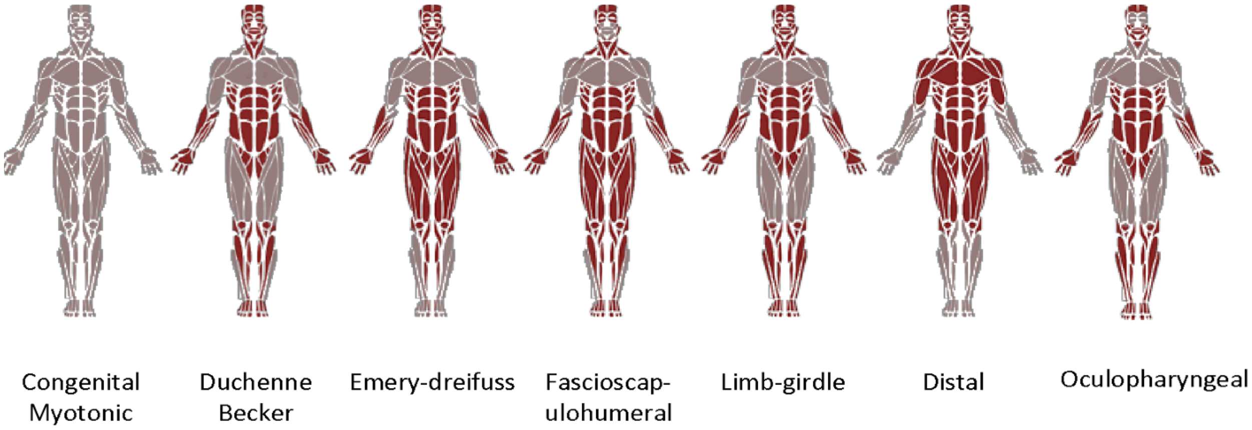


Figure 2.8: Visual representation of the areas of the body affected by each of the nine muscular dystrophies expanded upon in Table 2.2. Red areas represent unaffected areas, grey areas represent areas of weakening or wasting due to each type of dystrophy.

2.2.2. Neuromuscular disorders

The lines blur between strictly muscular disorders and neuromuscular disorders as the central nervous system and the muscular system are closely tied. Many muscle disorders like those mentioned in the previous section incur damage to the nervous system through disuse, as neural links to the muscles become weaker [56]. Many disorders, however, affect the nervous system before presenting symptoms within the muscles. Neuromusculoskeletal disorders, such as Motor Neuron Disease (MND), neuropathy, and Multiple Sclerosis (MS) are all diseases of great severity, initially causing pain, discomfort, and muscle twitching, and leading to immobility, respiratory problems, and a reduced life expectancy [57]. These disorders all affect, for widely varying reasons that will not be explored in this thesis, the nervous system which supplies the nerve impulse causing muscles fibres to contract [56]. Over time, when muscle fibres are not supplied with nerve impulses, they begin to die leaving no known route to recovery. As portrayed in Section 2.1.1.4, neuromuscular junctions attach individual nerves to many muscle fibres. Damage to one nerve can therefore affect many muscle fibres. Compounding the effect of these neuromuscular disorders, muscle fibres themselves are relatively large, multinucleated cells, with many active sarcomeres within each fibre. Logically, these neuromuscular disorders can act rapidly depending on disease and the site that they affect (whether it is the nerves or the neuromuscular junction itself). For many neuromuscular disorders, there are no known cures, but treatments such as physiotherapy attempt to maintain the neurological connection between the nervous system and the muscles [58, 59]. Though difficult to treat, neuromuscular disorders are typically well diagnosed with symptoms often visualised through medical imaging techniques for both the brain and the muscle tissue that could be affected [60]. Though beyond the scope of this thesis, the tools built and explored could benefit the understanding of such disorders, as the distribution of muscle tissue, or change in its volume in certain areas is used as an indicator for diagnosis with many neuromuscular disorders [61].

2.3. Techniques for measuring the lower limb anatomy

2.3.1. Anatomical planes of the body

There are 3 universally accepted planes and axes of the body. Firstly, to define the three spatial axes, known as the: frontal, sagittal, and longitudinal axes. The frontal axis can be thought of as the x-axis in a cartesian coordinate system and is parallel to a line passing from one shoulder to the other. The sagittal axis, the y-axis, can be thought of as parallel with a line passing from the back to the chest and is orthogonal to the frontal axis. Lastly, the longitudinal axis, the z-axis, is orthogonal to both the sagittal and frontal axes and is parallel to a line passing from head to feet. The three planes of the body: the frontal, sagittal, and transverse planes, are built across these axes. The frontal plane contains the frontal and longitudinal axes, the sagittal plane contains the sagittal and longitudinal axes, and the transverse plane contains the frontal and sagittal axes, as shown in Figure 2.9. Most visualisations of medical images of the body within this thesis are shown in the transverse plane. Lastly, one noteworthy feature of the medical images used in this thesis (to observe the inner anatomy), is that the anatomical left and right are visualised on the right and left of the images, respectively.

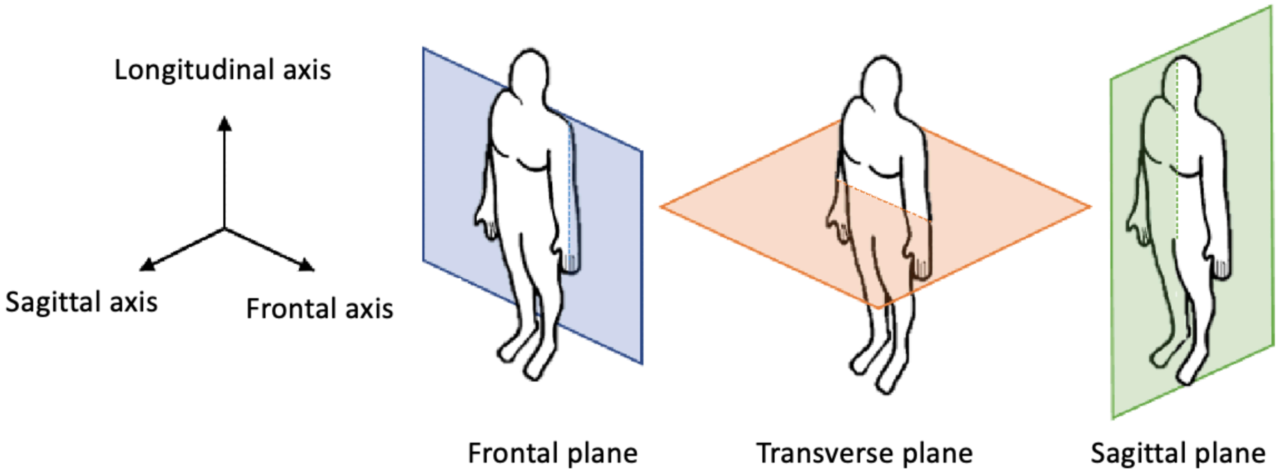


Figure 2.9: Anatomical planes and axes of the body.

2.3.2. Introduction to measurement techniques

Many techniques have been explored to investigate the musculoskeletal system within humans. These techniques range in sophistication from dissection of *ex vivo* subjects or palpating the features of interest, to more advanced options such as medical imaging. The limitations of dissection or manual palpation are obvious and so typically medical imaging is the preferred method used to visualise the lower limb musculoskeletal system, along with many other areas of interest within the body. There are several medical imaging techniques that have both benefits and limitations, but all offer computerised visualisation of the internal anatomy of humans.

2.3.3. Computer tomography

Computer Tomography (CT) is an imaging technique capable of capturing high-resolution three-dimensional (3D) images of the internal anatomy [7]. An example image acquired from the middle of the thigh is presented in Figure 2.10. CT scanners operate through rotating an X-ray source and detector about the subject, measuring the attenuation of the X-rays as they pass through the body. The attenuation measured can be attributed to either absorption or scattering of the X-rays by the tissues within the body, which is correlated to the density and size of the body. The rotation of the source and detector about the body allows the 3D tissue structures within the body to be captured (for 2D example see Figure 2.10). The internal anatomy is captured at regular intervals or slices, which are concatenated to form a 3D representation of the subject. CT, therefore, exposes subjects to a large amount of X-ray radiation, which is known to be harmful if the exposure is excessive [62]. In particular, the imaging of the region of interest in this work (the lower limbs) through CT would require many slices to be captured, incurring large amounts of X-ray exposure. On the other hand, CT is well suited for smaller capture volumes that focus on hard tissue, such as a 2D dental scan or visualisation of a bone fracture. Hard tissues attenuate the X-rays radiation to a higher degree than soft tissues, granting higher contrast across these tissues [63]. The focus on small capture volumes aims to limit the exposure of a subject to this potentially harmful ionizing radiation. Additionally, CT is used in those cases where other imaging modalities cannot be used, such as using Magnetic Resonance (MR) imaging for subjects with metal implants.

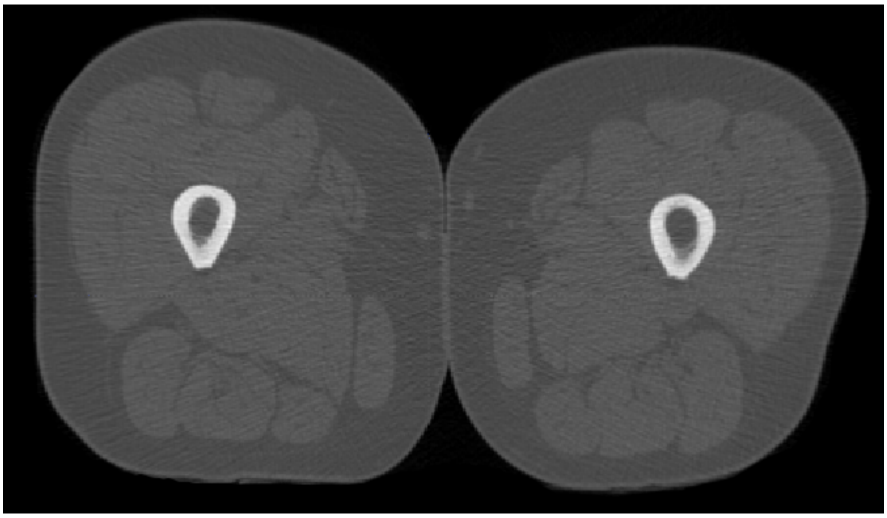


Figure 2.10: Example CT image acquired from the middle thigh. The most prominent tissue boundaries are the cortical-trabecular bone interface, and the muscle-bone interface.

2.3.4. Medical ultrasound imaging

Ultrasound (US) imaging uses high frequency sound waves often far outside the audible human hearing range (>20000 Hz) to visualise the internal body [64]. An example US image acquired in the middle of the thigh is shown below, in Figure 2.11. These high frequency sound waves are sent into the body in pulses via a probe connected to a display showcasing the images in real time. The sound waves pulsed into the body are reflected to the probe upon interaction with tissue, wherein the properties of the reflected sound wave are interpreted. The properties of the reflected sound wave are dictated by the material properties of the tissue that they are reflected by, allowing the display to present the different tissues. There are no known dangers of using US imaging to visualise the internal anatomy, and so it is widely used during pregnancy to visualise the development and physical state of the foetus. US as an imaging modality is extremely cost effective and easy to use and so is preferred in many applications to other medical imaging techniques. However, there are two main limitations of this imaging modality. Firstly, the resolution and clarity of US images is not as high as the other medical imaging techniques and the acquisition can be heavily dependent on the skill of the operator. Furthermore, the boundaries between tissues are typically very clear within images captured with US, but the boundaries within tissues, such as the boundaries between muscles, are not presented as clearly. Secondly, the images are traditionally visualised in real time to allow clinicians to seek out and isolate the region or tissue of interest. Therefore, 3D images are not typically reconstructed as a frame of reference is not defined, and this would be required for the images to be combined. For these two reasons, this imaging modality might not be well suited for capturing the lower limb anatomy.

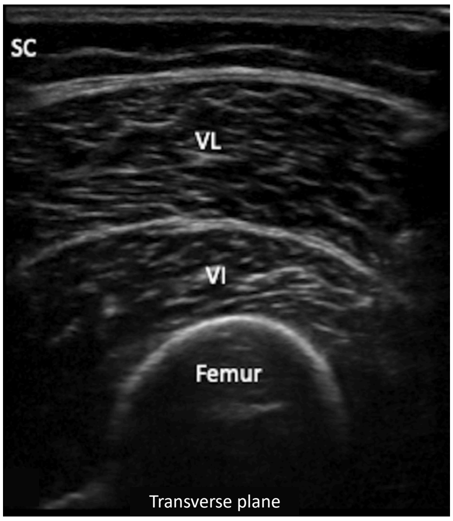


Figure 2.11: An example slice of transverse medical ultrasound image acquired from the mid-thigh. SC - subcutaneous tissue, VL - vastus lateralis, VI – Vastus intermedius. Image acquired from Albayda et al 2020 [65].

2.3.5. Magnetic resonance imaging

2.3.5.1. Basic concepts

Magnetic Resonance (MR) imaging uses strong magnetic fields and radio frequency (RF) pulses to scan the human body and capture images of the internal anatomy. An over-simplified view of MR imaging is the measurement of the density of hydrogen atoms within the different tissues inside the human body. This form of imaging is typically used in hospitals or research centres to form clinical diagnoses, categorise the stages of disease, and measure the effectiveness of disease intervention. MR imaging is better suited for soft-tissue (muscle, brain tissue, abdominal organs) than CT or US, as the contrast presented is much greater. There is very little risk associated with MR imaging as no ionising radiation is used to generate the images. Subjects are always thoroughly checked for metal implants (knee replacements, dental fillings, earrings) before proceeding to the scan as MR scanners use strong magnetic fields to acquire the imaging data. The scanning times can be long depending on the capture volume and the subject must remain stationary throughout the scan. Therefore, children, people with specific musculoskeletal disorders (for example, cerebral palsy) and elderly individuals may experience some discomfort while being scanned. Nonetheless, MR imaging was selected as the imaging modality to explore within this thesis given its advantages in visualising different muscles. In addition, retrospective data of full lower limb MR images was available from a previous study [66].

2.3.5.2. The physics of magnetic resonance imaging

MR imaging relies on the principles of nuclear magnetic resonance discovered by Purcell and Bloch (1946) [67, 68]. All atomic nucleons (protons and neutrons) have the quantum property of spin, which can be seen as analogous to angular momentum, and MR imaging measures the effect of changing the spin of atomic nuclei within the human body. Atomic nuclei with an even number of nucleons, such as helium or oxygen, have a total spin of 0 as the spin of each nucleon is balanced by the others. Alternatively, atomic nuclei with an odd number of nucleons have a non-zero total spin. One such atomic nuclei is the hydrogen nucleus, which contains only one proton (positively charged nucleon). Given the positive electric charge and angular momentum (spin) of the nucleus within the hydrogen atom, a local magnetic field is induced around the atomic nucleus. Typically, in the absence of a magnetic field surrounding the hydrogen atom, the direction of the local magnetic field is random as the direction of spin of the hydrogen atom is also random. Within a large living body such as that of a human, where hydrogen atoms are abundant within water molecules, the net magnetisation is zero. MR imaging takes advantage of the large number of hydrogen atoms (and in some cases other abundant electrically charged atomic nuclei such as those in carbon atoms) by aligning the local magnetic fields and thereby aligning the spin of the atomic nuclei through application of a strong, uniform external magnetic field (B_0). The spin of the hydrogen nuclei can have two orientations upon application of the magnetic field, B_0: along the magnetic field (low energy state) or against it (high energy state). The sum of energy within a particular volume gives the net magnetization.

The signals that are interpreted to generate MR images are created through excitation of the atomic nuclei within the body whilst under the magnetic field B_0. The excitation of the atomic nuclei (which can be thought of as lifting the energy state from low to high), occurs as an oscillating RF pulse (amplitude B_1, pulse duration t_p) is applied to the body perpendicular to the magnetic field, B_0. RF pulses are selected as electromagnetic (EM) radiation of this frequency can be readily absorbed by hydrogen nuclei. Upon application of the RF pulse, some of the hydrogen nuclei that are spinning, those with their local magnetic field aligned along B_0 (i.e. the low energy state), are excited through absorption of the energy supplied by the RF pulses. The local magnetic field induced by the nuclei aligns against the magnetic field, as the direction of spin has changed in response to the change in energy state. The oscillation frequency of the RF pulse is dictated by the strength of the magnetic field B_0 (the images used in this thesis are all gathered using a 1.5T MR scanner, requiring an oscillation frequency of 64 MHz). The process affects a large proportion of the nuclei present within a capture volume and therefore affects the net magnetisation which is aligned with the main magnetic field B_0 before the RF pulses are supplied. In this way, the net direction of the net magnetism changes in response the RF pulses. The supply of this EM radiation is maintained until the direction vector of the net magnetism reaches a user inputted value named the 'flip angle'. The flip angle used for all images investigated in this study is 10 degrees. The rate at which the excited nuclei return to the low energy state when the net magnetism reaches the flip angle and the differences in this rate between the tissues provide the contrast within the resulting images. There are two mechanisms under which the nuclei that attain the higher energy state lose energy and return to the lower energy state: firstly, through heat loss to the surrounding tissue, and secondly to surrounding nuclei as kinetic energy.

2.3.5.3. T1 and T2-weighted magnetic resonance imaging

The first type of energy loss is referred to as T1-relaxation (or spin-lattice relaxation) and depends on material properties such as the heat flux. For example, trabecular bone is far less dense than cortical bone and facilitates T1-relaxation much more readily. The T1-relaxation time is measured within the field of view and these times correspond to the colouration visible within MR images. Trabecular bone has a short T1-relaxation time due to its material and structural properties and so appears light grey colouration within T1-weighted MR images. Conversely, cortical bone has a long T1-relaxation time, incurring a dark black colouration. An example T1-weighted MR image is visualised in Figure 2.12, below. This weighting of MR images is preferred for identifying fatty tissue and obtaining morphological information of the anatomy, as the tissues within the human body have widely varying T1-relaxation times.

The second type of energy loss is referred to as T2-relaxation (or spin-spin relaxation), wherein the kinetic energy of the spinning nuclei is lost to adjacent spinning nuclei. T2-weighted images therefore measure the loss of coherence of the net magnetic direction as the nuclei that lie at the high energy state decay to the low energy state. Regions of high water content decay at a faster rate, due to the greater number of collisions of the particles in these areas, facilitating the exchange of spin between adjacent nuclei. The rate at which the net magnetic direction changes is represented within T2-weighted images as the greyscale value, where a fast T2-relaxation is shown as bright white and slow T2-relaxation is shown as dark black within the resulting images. An example T2-weighted MR image is shown in Figure 2.12, below. For this reason, T2-weighted MR images are used for identifying white matter lesions (such as tumours) or regions of inflammation as these areas have a higher water content and appear bright white in T2-weighted MR images.

In the assessment of muscle tissue, T1-weighted is the optimal weighting to be used as the contrast between muscle tissue and muscle boundaries is greater than in other weightings of MR images.

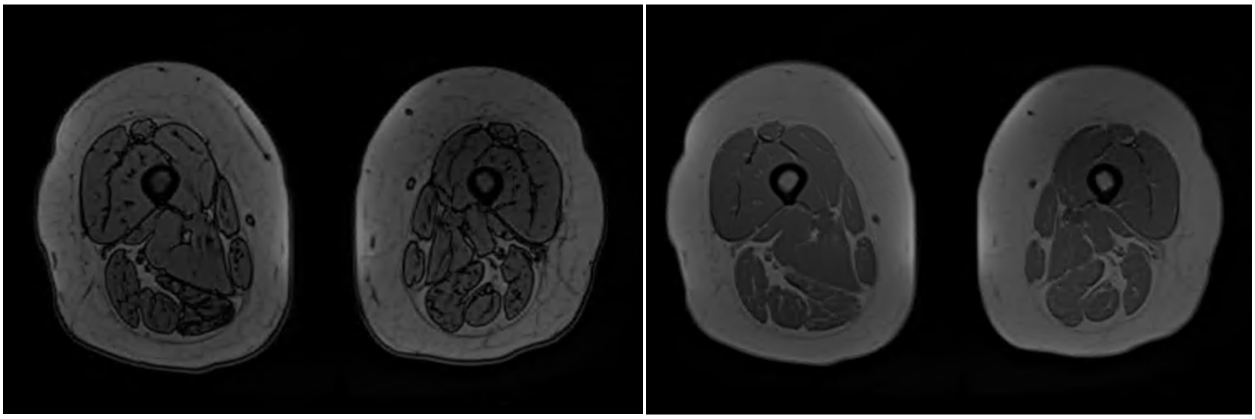


Figure 2.12: Comparison between T1 and T2-weighted MR imaging acquisition settings from images taken from within the thigh. The boundaries between and within muscle tissues are far clearer within the T1-weighted images as opposed to the T2-weighted images.

Assessing the relaxation of nucleons within the subject tissue alone does not generate an interpretable image. MR scanners extract the T1- or T2-weighted images at two dimensional (2D) intervals (referred to as slices) along the field of view of the scan, through a receiving coil. The receiving coil measures vector of current induction and therefore interprets the rate of change in the net magnetisation. The net magnetisation is measured at intervals within slices along the capture volume by encoding the volume using a frequency gradient across the two orthogonal axes of the slice. The area is thereby separated into cubes, which are represented as the voxels (3D pixels) within the resulting image. The response of the tissue within each encoded square is decoded, allowing the responses of the excited tissue to be measured with real-world spatial structure. The size of the squares (pixels) that the field of view is separated into is dictated by the frequency of the RF waves, typically between 0.5-2mm^2 for lower limbs.

2.3.5.4. Definition of magnetic resonance imaging parameters

There are many imaging parameters that can vastly alter the resolution of the outputted MR images. The following list describes each of the parameters that require setting.

1) Repetition time – The time between subsequent RF pulses applied to each capture area (slice). This value determines the amount of local magnetisation

that can recover between pulses, controlling the level of T1-relaxation. A shorter repetition time hinders T1-relaxation and therefore produces more contrast between tissues within T1-weighted images.

2) Echo time – The time between measurements of the electrical induction (caused by nuclei spinning in high energy state) by the receiving coil. The level of T2-relaxation is determined by the echo time. A longer echo time results in an enhanced response in white (tumour, inflammation) and grey matter (muscle, brain) tissues, as the amount of T2-relaxation between the different tissues is accentuated.

3) Flip angle – The amount of rotation of the net magnetisation during a single RF pulse. The flip angle selected determines how long the RF pulses are applied for.

4) Pixel size – The physical size of the elements that each slice is sectioned into. The pixel size allows conversion of the image pixels to represent real world spatial size.

5) Slice spacing – The spacing between subsequent slices captured by the MR scanner. The slice spacing has no influence on the resolution of each 2D image captured by the scanner but changes the resolution of a reconstructed 3D image if one is created.

2.4. Medical image segmentation

Medical image segmentation is a generic process in which structural anatomical information is isolated from medical images. Segmentation entails partition of homogenous materials within medical images that represent the tissues. This process has numerous applications within many medical sub-domains and has been used to isolate structural characteristics for: the muscular, skeletal, central nervous system, cardiac, and respiratory systems, as well as the organs within the abdomen, and tumours [18, 20, 69, 70]. This technique has led to many medical discoveries within all noted areas, facilitating an understanding of disease mechanisms within these areas [15, 18].

2.4.1. Muscle segmentation

Muscle segmentation is an application of medical image segmentation, where structural anatomical characteristics of the muscles are gathered *in vivo.* Muscle characteristics such as muscle volume, geometry, level of fatty infiltration, and attachment or insertion locations are all available through muscle segmentation. Through tracking these characteristics, muscle segmentation has facilitated the monitoring of the progression of musculoskeletal and neuromusculoskeletal diseases [71, 72]. Furthermore, this technique can be and has been shown to allow surveillance of the effectiveness of treatments [61, 73] and has been used for diagnostic purposes of musculoskeletal [74] and neuromusculoskeletal diseases [75]. Additionally, muscle segmentation of individual muscles has been used to inform pre-surgical planning and measure the effectiveness of a surgical intervention of specific musculoskeletal disorders [76]. Performing muscle segmentation requires labelling the contours of individual muscles from medical images, an example is shown in Figure 2.12 below. Labelling multiple images provides a three-dimensional representation of the muscles, allowing structural characteristics to be isolated. Currently, there are some limitations associated with the gold standard techniques used to isolate muscle characteristics from medical images, such as its associated repeatability issues, and lengthy time requirements. The aim of this thesis is to address these limitations through exploration of novel methods.

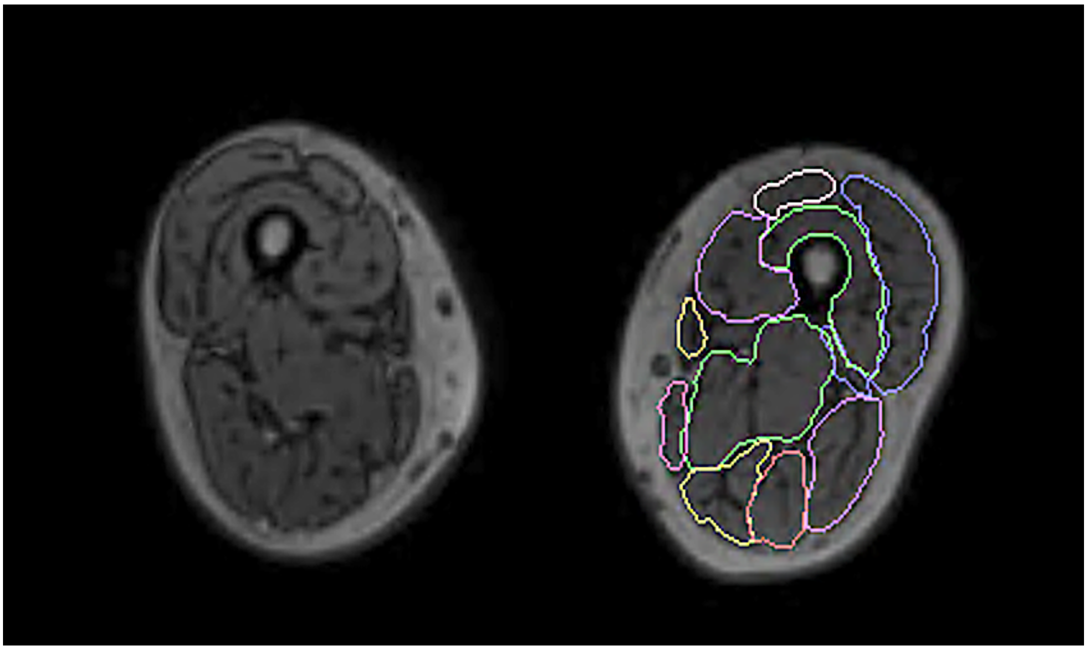


Figure 2.12: An example of a segmented 2D MR imaging slice from within the thigh. A full 3D representation of muscle segmentation is visualised in Figure 2.1.

Chapter 3:

Motivational studies and Literature review

3.1. Introduction

Chapter 3 of this thesis is split into three main sections. Section 1 discusses the current gold standard approach to perform muscle segmentation. Initially, the advantages and limitations of the gold standard approach are discussed. One of the main limitations of the gold standard approach centres around a repeatability issue of the outputs, which has been characterised for the cohort used throughout the thesis and summarised thereafter.

Section 2 outlines an application of muscle segmentation investigated by the author within a preliminary study. The objective of the preliminary study in the context of this thesis, is twofold: 1) to present an example of the scientific investigations that are available with manual muscle segmentation, and 2) to highlight the limitations of the gold standard approach, motivating the requirement of innovative segmentation methods to overcome them.

Section 3 of this chapter is a review of the literature surrounding automatic muscle segmentation, summarising the current state of the research knowledge. There are a multitude of approaches that have been investigated, and these are discussed. The conclusive remarks of the chapter motivate the further study of the two areas mainly explored within the thesis: deformable image registration, and deep learning, addressing the limitations of methods to perform automatic muscle segmentation.

3.2. The gold standard approach

Individual muscle segmentation is a process used to characterise the muscles of subjects to study the capacity, capability, and identify problematic areas of the muscular system in its ability to generate force and allow mobility [77-79]. Muscle segmentation results in the isolation of the three-dimensional (3D) geometry of the muscles directly from medical imaging data, whether it be Ultrasound (US), Computerised Tomography (CT), or Magnetic Resonance (MR) imaging.

3.2.1. Manual segmentation

Manual segmentation is the current gold standard method used to extract structural muscle characteristics from medical images, gathered from CT, US, or MRI [28]. Manual segmentation requires a trained operator to manually label the tissues of interest from 2D images. After labelling the tissue within subsequent 2D cross sections of a 3D image volume, a 3D surface can be constructed, retaining relevant spatial information and structural detail of the segmented tissue. From these 3D geometries, global characteristics of muscle such as the volume and shape, as well as more internal properties like the level of fat infiltration can be isolated [80]. These characteristics are indicative of muscle force generation, and more generally, muscle function [77-79].

The 37 muscles within the lower limbs (for details, see Table 2.1) span the longest feature of the human body. The relatively large number and size of muscles in this area cause muscle segmentation to be one of the applications of medical image segmentation that incurs the greatest time expense and operator variability [28, 66]. Explicitly, the process of manually segmenting all muscles within the lower limbs of one subject has been quoted to require upward of 24 hours [80]. With recent advances in both software (such as linear interpolation between manually segmented image slices) and hardware (such as trackpads), the time to perform manual segmentation has been reduced to approximately 10 hours per subject [66]. The second problem associated with manual segmentation is that the process remains heavily operator dependent, and this is extensively documented within the literature surrounding the process [28, 66, 81].

3.2.2. Operator dependency quoted within the literature

The operator dependency issues surrounding muscle segmentation arise due to the difficulty in visually isolating the muscles from medical images [28, 82]. Individual muscles are tightly packed together and the boundaries between them present small discontinuities in texture within medical images [83] that are very similar in appearance to intra-muscular fat infiltration [80] Furthermore, the high degree of variability of muscle shape and volume, between muscles in different subjects causes difficulty in segmenting the muscles [66]. Repeatability studies found throughout the literature highlight this, with muscle volumes and calculated from repeated segmentations consistently being over or underestimated by around 5% and up to 50%, depending on the muscles segmented and the cohort investigated [28]. The impact of this effect is wide reaching, with the potential to drastically alter the interpretation of muscle shape, volume, and length, hindering the conclusions that can be drawn from such characteristics. Repeatability issues associated with muscle segmentation are present in both inter-operator (segmentations performed by different operators) and intra-operator (repeated segmentation by one operator) analyses. In a systematic review of publications with an operator dependency study into manual muscle segmentation from MR imaging data, Pons et al. [28] found that, intra-operator reliability was good (< 5% over or underestimation of muscle volume between repeated segmentations) within 4 out of 11 included studies. Though, the inter-operator reliability was worse, being characterised as good (defined previously) to moderate (5-10% over or underestimation between repeated segmentations) within 8 of the studies. The difference between the inter and intra-operator analyses showed an accentuated operator dependency problem when comparing repeated segmentations from two different operators, rather than one operator repeating the segmentations [84, 85]. Given that there will always be a need for multiple operators to perform muscle segmentation, if using the gold standard approach, and that there are repeatability problems when comparing results found from different operators, there is a clear and significant problem with the current gold standard approach.

Moreover, there were studies that compared the repeatability issues between healthy and non-healthy cohorts [28]. One of two studies that highlight this, was that of Skorupska et al. [84], where the pelvic muscles (see Table 2.1) were segmented within a healthy cohort and a cohort with low back and leg pain. In this analysis, Skorupska et al. [84] found that the inter-operator repeatability was good for the healthy cohort, and moderate for the low back and leg pain cohort. Moreover, Springer et al. performed an intra-operator analysis for the pelvic muscles from both limbs of subjects with one hip replacement. In this study, Springer et al. [85] characterised the intra-operator repeatability and concluded that a higher variability incurred when segmenting the limb with the hip replacement (average of 1.4% more variability of the muscle volume found between repeated segmentations). Both studies highlighted that the repeatability of the manual segmentation process was lower in non-healthy cohorts. This is a significant shortcoming, as the people for whom muscle segmentation would be most impactful: people with muscle pathologies, the segmentation results were more susceptible to operator variability issues [28, 84, 85]. This is a logical conclusion, as the structure of pathological muscle bodies are far more complex in their appearance, with more visible damage within the muscles further blurring the lines between muscle boundaries, as can be seen in Figure 3.1.

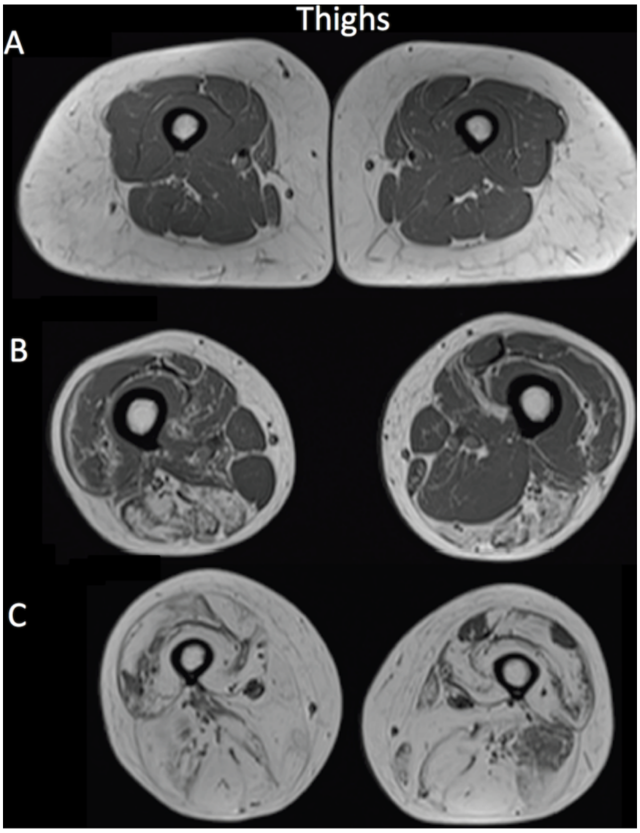


Figure 3.1: MR imaging within the thigh, showing healthy (A), selective muscle damage (B), and severe muscle damage (C). Image taken from Lareau-Trudel et al. [86], study cohort was healthy young individuals, and patients with facioscapulohumeral muscular dystrophy.

The inter-operator and intra-operator variability of manual muscle segmentation within the two cohorts used in this thesis have been characterised. Firstly, the MultiSim database, with the variability characterised in a study by Montefiori et al. [66]. Secondly, a database collected in a project titled STH21022, wherein MR images were collected from the lower limbs of 27 older women. The imaging data collected through the MultiSim project was used throughout the thesis, whereas the subjects within the STH21022 project were used only in a preliminary study and as such, further details for the STH21022 cohort are available in Section 3.3.

3.2.3. Preliminary study - Operator variability of muscle segmentation within the MultiSim cohort

The inter-operator and intra-operator variability of the Multisim cohort (used in most of this thesis) was characterised in a study by Montefiori et al., with input from the author of the thesis [66]. In this study, trained operators segmented the muscles from both lower limbs from T1-weighted MR images. The muscle segmentations gathered enabled quantification of the dissimilarity of the muscle volume and length between the dominant and non-dominant limbs of the subjects. The author of the thesis contributed to the quantification of operator variability and presented a novel automatic algorithm to quantify muscle length. The automatic algorithm for quantifying the muscle length is not used within the thesis but is explained in Appendix 1.

3.2.3.1. Participants and data acquisition for MultiSim cohort

Lower limb T1-weighted MR images from 11 post-menopausal women (mean (standard deviation)): 69 (7) years old, 66.9 (7.7) kg, 159 (3) cm) were used for this study. Images were collected using a Magnetom Avanto 1.5T scanner (Siemens, Erlangen Germany), with an echo time of 2.59 ms, repetition time of 7.64 ms, flip angle of 10 degrees. To reduce scanning time while still providing detailed geometries of the joints for use within the original study, the joints were acquired with a higher resolution (pixel size 1.05 mm^2, slice spacing 3.00 mm) than the long bone sections (pixel size 1.15 mm^2, slice spacing 5.00 mm). The study was approved by the East of England – Cambridgeshire and Hertfordshire Research Ethics Committee and the Health Research Authority (16/EE/0049).

3.2.3.2. Muscle segmentation protocol

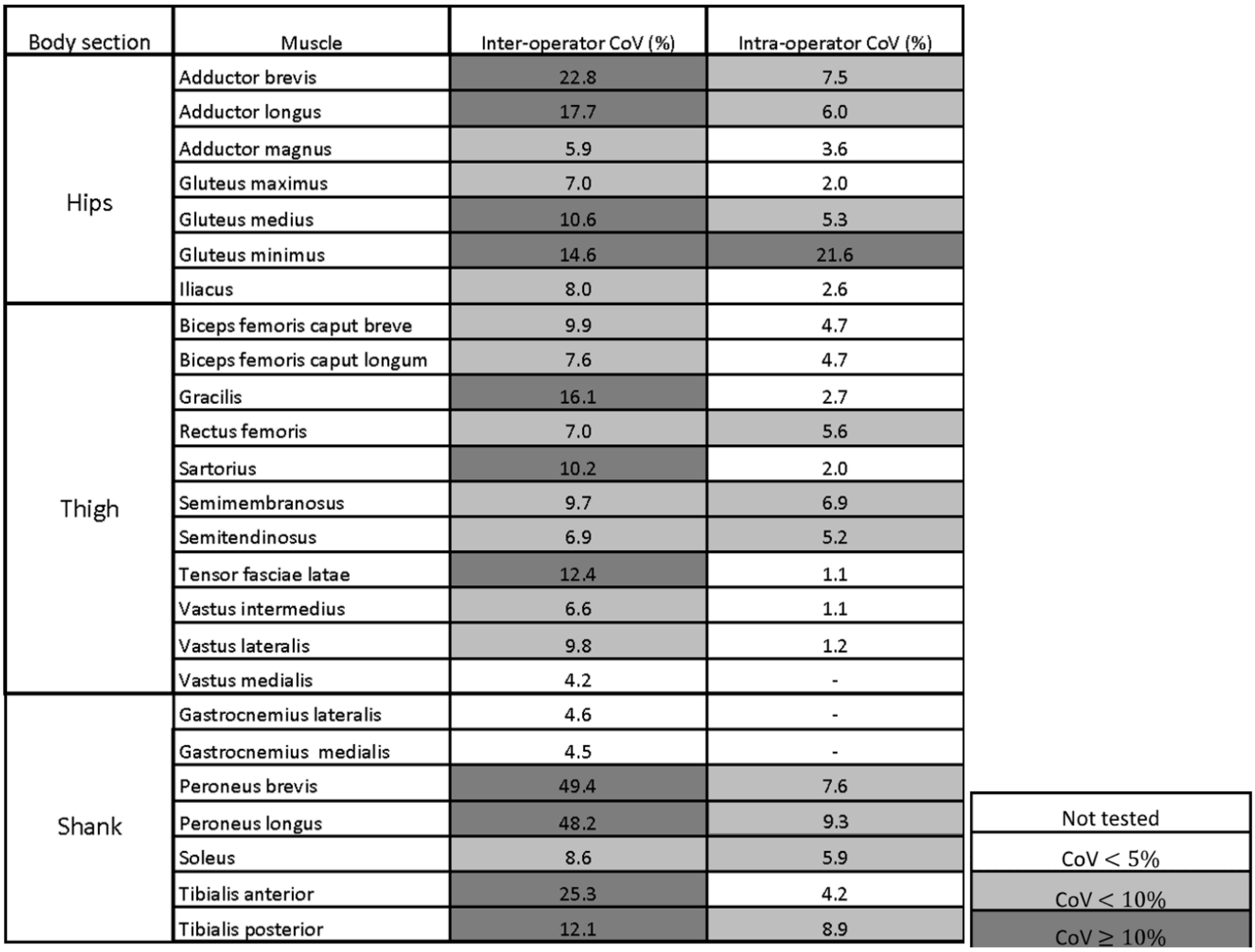
Twenty-five out of the 37 muscles outlined in Chapter 2 Section 2.1 were segmented from the MR images. The 12 muscles that were not segmented either lay outside the field of view of the images or were deemed not visible within the MR images due to the relatively low resolution of the images. The images were segmented using a semi-automatic approach (Mimics research 20.0, Materialise, Belgium). The pipeline for the semi-automatic approach began by combining the MR imaging sequences of the hips, thighs, knees, and shank. To complete this, a module within Mimics allows the user to manually locate multiple points within each of the sequences and automatically aligns them. With the aligned sequences, the user segmented the entire muscle body from the other tissues visible within the images. Thereafter, an automated, atlas-based prediction is generated, roughly locating the muscles. To finalise the segmentations, manual processing is needed. The manual processing required 10 hours on average for all muscles to be segmented per subject.

To analyse the inter-operator dependency, the Coefficient of Variation (CoV) was calculated for the muscle volumes acquired from the repeated segmentation of three subjects by three different operators. The muscles for which the CoV was less than 10% were considered acceptable [28, 66], but to be more conservative, those muscles for which the CoV in the inter-operator analysis were greater than 5% were tested for intra-operator dependency issues. The intra-operator CoV was subsequently calculated from the muscle volumes resulting from repeated segmentation of one subject three times by a single operator. All muscles with inter-operator and intra-operator repeatability greater than 10% were removed from further study, as the manual segmentations were considered not repeatable. For future analyses, these manual segmentations are the references to be compared against results obtained from automatic segmentations. Given that the muscle segmentations for some muscles are not repeatable when applying the gold standard approach, retaining these muscles would introduce bias in future results.

3.2.3.3. Inter and intra-operator repeatability results

The results for the inter and intra-operator analyses are presented in Table 3.1. The CoV resulting from the inter-operator analysis was substantially greater than the CoV found within the intra-operator analysis, for all muscles tested. The largest muscles: gluteus maximus, rectus femoris, adductor magnus, vastii, gastrocnemii, and soleus, presented the lowest CoV across both analyses, and were easily identifiable from within the images. Of the 25 muscles segmented, eleven had inter-operator CoV greater than 10% and failed to meet the outlined threshold. On the other hand, only one of the muscles tested in the intra-operator analysis presented a CoV greater than the threshold: the gluteus minimus. Additionally, the number of muscles that resulted in a good CoV (< 5%) within the intra-operator analysis (11) was considerably higher than those in the inter-operator analysis (3).

The overarching results in the context of this thesis from this investigation are that the repeatability issues associated with an inter-operator analysis are greater than those associated with an intra-operator analysis. This finding is in line with those results found in the literature, particularly with the systematic review presented by Pons et al. [28]. As a result of these analyses, the muscle segmentations considered within subsequent chapters are those gathered by one operator, ensuring consistency, and removing any inter-operator effect. Additionally, due to the results of this initial study, the gluteus minimus was removed from the analyses within the proceeding chapters, as the repeated segmentations of this muscle were found to fail the inclusion criterion (CoV < 10% within either the inter- or the intra-operator analysis). Finally, the gluteus medius was found to exceed beyond the field of view of the MR imaging sequences within several subjects and was also excluded. Therefore, the segmentations of 23 muscles were used in the analyses performed within proceeding chapters.



Body section	Muscle	Inter-operator CoV (%)	Intra-operator CoV (%)
Hips	Adductor brevis	22.8	7.5
	Adductor longus	17.7	6.0
	Adductor magnus	5.9	3.6
	Gluteus maximus	7.0	2.0
	Gluteus medius	10.6	5.3
	Gluteus minimus	14.6	21.6
	Iliacus	8.0	2.6
Thigh	Biceps femoris caput breve	9.9	4.7
	Biceps femoris caput longum	7.6	4.7
	Gracilis	16.1	2.7
	Rectus femoris	7.0	5.6
	Sartorius	10.2	2.0
	Semimembranosus	9.7	6.9
	Semitendinosus	6.9	5.2
	Tensor fasciae latae	12.4	1.1
	Vastus intermedius	6.6	1.1
	Vastus lateralis	9.8	1.2
	Vastus medialis	4.2	-
Shank	Gastrocnemius lateralis	4.6	-
	Gastrocnemius medialis	4.5	-
	Peroneus brevis	49.4	7.6
	Peroneus longus	48.2	9.3
	Soleus	8.6	5.9
	Tibialis anterior	25.3	4.2
	Tibialis posterior	12.1	8.9

Not tested
CoV < 5%
CoV < 10%
CoV ≥ 10%

Table 3.1: Inter and intra-operator repeatability of the manual segmentation procedure for the MultiSim cohort of post-menopausal women [66].

3.2.4. Applications of manual muscle segmentation

There are numerous applications of muscle segmentation and currently, they rely on the gold standard approach to manually segment muscles. As mentioned previously, there are three main outputs of muscle segmentation: muscle shape, volume, and to measure the level of fat infiltration. A change in the shape or volume of muscle can be normal, resulting from hypertrophy or atrophy, but could be a physiological response to a Musculoskeletal (MSK) or Neuromusculoskeletal (NMSK) disorder, or even injury

[61, 75, 87]. For these reasons, muscle segmentation is a tool that is very useful in areas such as sports science and clinical research. The segmentation results can either be used directly or indirectly through building subject specific dynamic MSK models [88].

Quantitative assessment of muscle shape and volume is critical for research in MSK and NMSK disorders in order to investigate how the pathology affects the muscular system [59, 60]. Muscle volume and the level of fat infiltration are effective indicators of muscle functional capacity or strength [89, 90]. Muscle atrophy and fat infiltration are physiological consequences of (N)MSK disorders, and quantification of these effects could lead to a better understanding of the mechanisms under which the disorders act. For these reasons, quantification of muscle shape and volume in cohorts with certain (N)MSK disorders has allowed a more effective understanding of the early stages of diseases and has enhanced diagnosis strategies [91, 92]. Moreover, the effects of treatments, both pharmaceutical and physiotherapeutic, could also be quantified through muscle segmentation in longitudinal studies [93, 94]. Therefore, clinical strategies to slow or reverse progression of MSK disorders could be better informed using muscle morphology and structural characteristics derived through muscle segmentation.

Many disorders that affect the muscles manifest through progressive penetration of fat into the muscle tissue, wherein contractile muscle tissue is replaced with non-contractile adipose tissue [29, 90]. Therefore, quantification of muscle fat infiltration is an important application of muscle segmentation, which would allow detailed characterisation of muscle disorders. One example is the study presented by Lareau-Trudel et al., where the level of fat infiltration was assessed through muscle segmentation in subjects with Facioscapulohumeral dystrophy (for details see Chapter 2, Section 2.1) compared with a group of healthy controls [86]. In this study, the level of intra-muscular fat was found to be much higher in the disease group (21.9 ± 20.4%) than in the healthy control group (3.6 ± 2.8%). Upon further analysis, given the high standard deviation of intra-muscular fat content in the disease group, the authors clustered 3 different imaging patterns within the disease cohort, which could represent different stages of this disease. Similarly, Wokke et al. [95] investigated the intra-muscular fat penetration in healthy subjects and subjects with Duchenne muscular dystrophy. In this study, the level of fat infiltration was measured for the individual muscles, finding a strong correlation (spearman $r = 0.89$, $P < 0.0001$) between two different approaches to determine the level of fat infiltration. The study was able to accurately measure the level of fat infiltration of many muscles within numerous subjects. A clear difference was found in the level of fat infiltration between the healthy controls and Duchenne muscular dystrophy cohort, 5.3 ± 0.98% and 29.7% ± 13.2%, respectively. These studies demonstrated that fat infiltration is a key identifier of muscle disorders. Following a muscle-specific approach could lead to a better quantitative understanding of which muscles are affected most by various disorder.

As stated previously, one more common but less severe muscle disorder is sarcopenia. Sarcopenia is the age-related degradation of muscle tissue in older individuals, which limits their physical capacity [27]. Like the other disorders described above, sarcopenia is characterised by fat infiltration of the muscle tissue [53]. There have been many studies surrounding sarcopenia since the revised clinical definition and diagnosis algorithm [27]. One such study presented by Lees et al. [96], who compared the total lean tissue mass, muscle strength, and muscle quality characterised using Computer Tomography (CT) imaging within 50 young healthy volunteers and 50 older volunteers. In this study, the authors suggested that only 2-4% of the older participants could be diagnosed with sarcopenia, but as many as 50% had low lower-body muscle quality, i.e., high levels of fat infiltration. The vast disparity between these two figures suggests that there are shortcomings in the current assessment of sarcopenia [96].

Large databases of a variety of disease cohorts would be required in order to rigorously derive a better understanding of the mechanisms underpinning muscle disorders. This quantitative understanding would be required to enhance therapeutic measures and grant earlier detection of such disorders. The following section outlines a preliminary example of one such study completed, showcasing the types of investigations into muscle disorders that could be enabled through muscle segmentation.

3.3. Fat infiltration in healthy, obese and dynapenic abdominal obese adults

The author of the thesis and Dr. Lisa Dowling contributed equally to the work presented in Section 3.3. Data acquisition was conducted by Dr. Lisa Dowling. Data pre-processing was completed by the author. The manual segmentation procedure was completed in a combined effort by the author and Dr. Lisa Dowling. Data post-processing was completed by the author, and the methods and computational tools used to process the data were built by the author. Statistical analysis was performed by the author of the thesis.

3.3.1. Study motivation

A preliminary study using manual muscle segmentation was undertaken to add to the current knowledge surrounding the effects on muscle quality of sarcopenia in older individuals. The study sought to highlight the difference in muscle quality and compare the physical capacity of the lower limb muscles between older people within three different cohorts: Normal Weight (NW), Obese (OB) and Dynapenic Abdominal Obese (DAO). Medical images were captured from the lower limbs of 26 individuals, and the maximum isometric force of knee extension and flexion were measured.

3.3.2. Methods

3.3.2.1. Subjects & imaging acquisition

Lower limb T1-weighted and Dixon method [97] MR images were acquired from 26 female subjects (Age range: 60-79; mean age: 66.6). Ethical approval was granted by the Leeds West Research Ethics Committee (REC Reference 20/YH/0274). Images were collected using a Magnetom Avanto 1.5T scanner (Siemens, Erlangen Germany), with an echo time of 2.59 ms, repetition time of 7.64 ms, flip angle of 10 degrees. The MR images were acquired in four sequences, capturing the hip, thigh, knee, and shank. To reduce scanning time, the joints were acquired with a higher resolution (pixel size 1.05 mm^2, slice spacing 3.00 mm) than the long bone sections (pixel size 1.15 mm^2, slice spacing 5.00 mm). As such, the imaging acquisition method followed that of standard clinical practise, and were not specifically selected for the purposes of this research.

The sequences were stacked in MATLAB forming one continuous 3D image containing the entire lower limb (specific details outlined in chapter 4). The subjects were split into three sub-cohorts: Normal Weight (NW, n = 10), Obese (OB, n = 9), and Dynapenic

Abdominal Obese (DAO, n = 7). The NW sub-cohort were those with Body Mass Index (BMI, kg/m^2) in the range of (18.5, 25), OB with BMI in the range of (30, 40), with both sub-cohorts performing five sit-to-stand exercises in less than 15 seconds. Those assigned to the DAO cohort had BMI in the range of (30, 40) and completed fewer than five sit-to-stand exercises in 15 seconds, following the clinical classification of DAO [27].

3.3.2.2. Muscle segmentation

To reduce segmentation time given the large study cohort and isolate the muscles of interest, two muscle groups in the dominant limb were segmented from the MR images of each subject: the knee flexors and extensors. Both groups consisted of 4 muscles that contribute most to the joint motion they permit [98]. The group of flexors consisted of the rectus femoris, vastus medialis, lateralis, and intermedius (for details see Table 2.1). The group of extensors consisted of the semimembranosus, semitendinosus, biceps femoris caput brevis, and longum. These two muscle groups were segmented from the pre-processed MR imaging sequences manually using 3DSlicer, an open-source manual segmentation software [99]. The volumes of the muscle groups were calculated from the segmentations.

The muscle groups were segmented from the MR images of the 26 subjects by two expert operators. As previously stated, there are noted repeatability issues when performing muscle segmentation [28], resulting particularly from inter-operator dependency [66]. Therefore, both the inter-operator and intra-operator repeatability were calculated for this segmentation task, through calculation of the CoV of the muscle volume of repeated segmentations. Firstly, the intra-operator repeatability of the segmentation procedure was assessed through one operator segmenting both the flexors and extensors from one subject 3 times. The inter-operator repeatability was assessed by both operators segmenting the flexors and extensors from 3 subjects. A CoV of less than 5% was deemed to be of a good level of repeatability [28, 66].

The muscle segmentations were performed in tandem with the Dixon method MR images to calculate the level of fat infiltration, as shown in Figure 3.2. The frequency-intensity histograms were plotted for the voxels containing the flexors and extensors, individually. The intensity threshold at which the muscle peak (found as the peak of greatest intensity in the frequency-intensity plots) was labelled as the point at which the gradient of the frequency intensity plot exceeded a value of 1, at the first point before the beginning of the peak, ensuring unbiased labelling (see red circle in Figure 3.2). The lean muscle volume was found through masking the muscle segmentations in response to this threshold (red circle in Figure 3.2), removing all tissue with greyscale value less than the outlined threshold. The fat infiltration (%) was defined as 100% minus the percentage of lean muscle volume over total muscle volume.

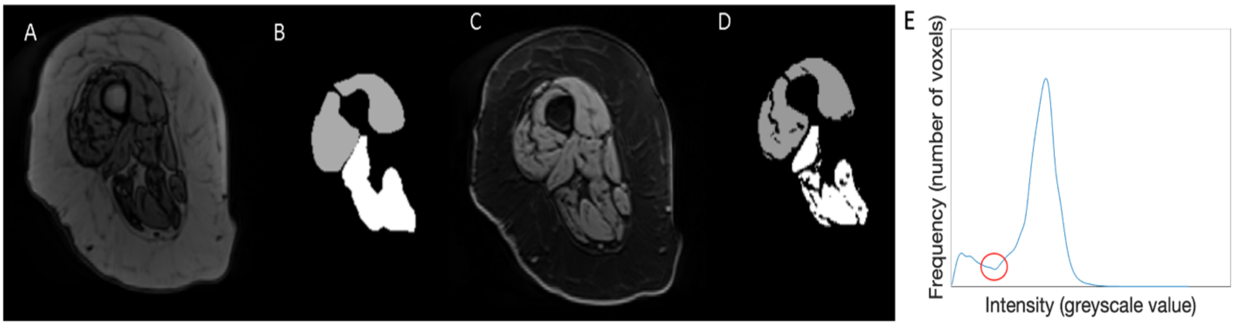


Figure 3.2: From left to right: Raw MR images from the thigh (A), the segmentation results (B) for the extensors (grey) and flexors (white), the Dixon method MR images (C), and the segmentations with the intra-muscular fat removed (D) from the single level threshold (marked by a red circle) identified by the frequency plots (E).

3.3.2.3. Maximum force output

The maximum isometric torque (τ_{max}) output of the flexors and extensors of the 26 subjects was measured using a BIODEX [100]. The knee joint was locked at a right angle by a strap placed at the ankle. The subject was then asked to flex and extend the knee with maximum effort for five times. The maximum torque measured by the BIODEX was the maximum of the 5 repeated measurements. To calculate the maximum muscle force (F_{max}) generated by the flexors and extensors, the distance (r) between the pivot and lever arm (the calf length) was measured. The calf length was calculated from the MR imaging sequences, measuring the distance from the knee to the ankle. The maximum force was calculated for both the flexors and extensors using the mechanical force-torque relationship, $F_{max} = \tau_{max}/r$. Using the maximum force and the segmented muscle volumes, the ratio between force and volume was calculated (force/volume, N/cm^3) for both the flexors and extensors of each subject, considering both total and lean muscle volumes. This metric was designed to allow direct comparison between the force output of the three different sub-cohorts normalized against the volume of the muscle groups.

3.3.2.4. Statistical analysis

A statistical analysis of the results was conducted to assess the differences in measured and calculated characteristics between the 3 different sub-cohorts. A one-sample Kolmogorov-Smirnov test was first used to test the total and lean muscle volumes, and the force/volume metrics for normality. The Kolmogorov-Smirnov test was used for each sub-cohort and muscle group independently. The nonparametric Wilcoxon rank sum test was then used to test the measured and calculated values for statistically significant differences between the 3 sub-cohorts. This test was selected as all measured and calculated values were found not to be normally distributed through the Kolmogorov-Smirnov test.

3.3.3. Results

3.3.3.1. Operator variability

Both flexors and extensor muscle groups achieved intra- and inter-operator CoV of less than 5%. Flexors intra- and inter-operator CoV were 4.8% and 1.2%, respectively. Extensors intra- and inter-operator CoV were 4.8% and 3.4%, respectively. Following the traditional characterisations of these values within the literature [66], the repeatability of the segmentation procedure was good in both analyses. The CoV found in this study are much lower than those found in the previously presented MultiSim cohort, likely due to the grouping of muscles as opposed to segmenting individual muscles.

3.3.3.2. Extensor and flexor volumes and level of intra-muscular fat

The total and lean muscle volume (cm^3), and fat content (%) calculated for both muscle groups within the three cohorts are presented in Figure 3.3.

Total volume: The total extensor volume of the three sub-cohorts were comparable, with mean ± standard deviation of 1194 ± 255 cm^3 (NW), 1109 ± 207 cm^3 (OB), and 1090 ± 181 cm^3 (DAO). Moreover, the total muscle volume of the flexor muscle groups was typically around half that of the extensors and was also comparable between the sub-cohorts, with mean ± standard deviation of 489 ± 87 cm^3 (NW), 463 ± 102 cm^3 (OB), and 456 ± 74 cm^3 (DAO).

Lean volume: No statistically significant difference was found between the lean extensor volumes of the three sub-cohorts, but the NW subjects on average had the largest lean extensor volume (1090 ± 247 cm^3), the OB subjects had the second largest (990 ± 184cm^3), and the DAO subjects had the least (952 ± 190 cm^3). Though a similar result was found for the lean flexor volume, with the NW (441 ± 86 cm^3) and DAO (362 ± 61 cm^3) sub-cohorts having the greatest and smallest average lean muscle volume, there was a significant difference found when testing the distribution of NW and DAO lean flexor volumes ($p=0.025$).

Fat infiltration: The fat infiltration (%) found between the three groups within the extensors mirrored the above results, with the NW sub-cohort with the lowest average level of fat infiltration (8.8 ± 3.6%), the OB sub-cohort with the second lowest (10.5 ± 5.0%), and the DAO sub-cohort with the most (12.9 ± 6.2%). Though, statistical tests showed that there was no evidence of differences between the three subjects. Within the flexor muscle group, the rankings were the same, with the NW subjects found to have the lowest level fat infiltration (9.9 ± 4.4%), the OB subjects with the second lowest (14.2 ± 5.7%), and the DAO subjects with the highest level of fat infiltration (17.7 ± 6.2%). Statistical tests showed that there was a significant difference between the fat infiltration of the flexors within the NW and the DAO subjects (p=0.0097).

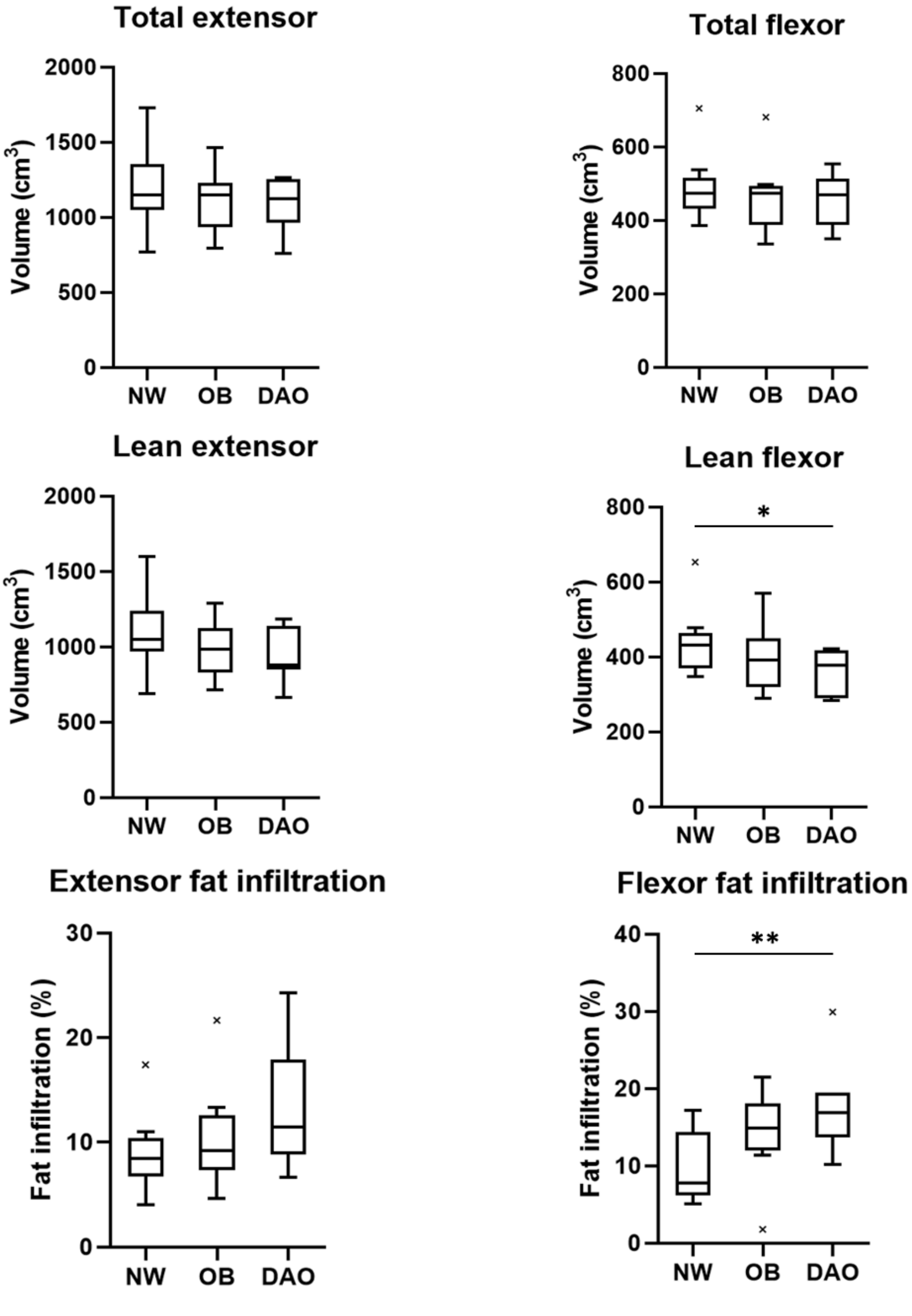


Figure 3.3: The total muscle volume (top row, cm^3), lean muscle volume (second row, cm^3), and level of fat infiltration (%), for the extensor muscle group (left) and flexor muscle group (right), across the three groups: Normal Weight (NW), Obese (OB), and Dynapenic Abdominal Obese (DAO). Lines connecting bar charts highlight statistically significant differences between the 2 connected groups, with * corresponding to $p<0.05$, ** to $p<0.01$, and *** to $p<0.001$.

3.3.3.3. Maximum isometric force per unit volume

The maximum isometric forces generated by the flexor and extensor muscle groups (measured independently using a BioDex) were normalised against both the total and lean muscle volumes of the muscle groups independently. The maximum force per unit volume generated by both muscle groups is shown in figure 3.4 for all subjects. Considering the extension and flexion forces normalised against the total extensor and flexor muscle volumes (respectively), the OB sub-cohort was able to generate the greatest force per unit volume (N/cm^3), with both the extensor muscles (0.353 ± 0.040 N/cm^3, mean ± standard deviation) and flexor muscles (0.403 ± 0.075 N/cm^3). Significant differences were measured between extensor force per unit volume of the OB and NW (0.280 ± 0.050 N/cm^3), and OB and DAO sub-cohort (0.272 ± 0.070 N/cm^3), with p = 0.0055 and 0.0229, respectively. Considering the total flexor volume, the DAO sub-cohort produced the lowest force per unit volume (0.313 ± 0.075 N/cm^3), significantly lower than the OB subjects (p = 0.0418). No statistical difference was found between the NW (0.381 ± 0.073 N/cm^3) and OB cohort considering the maximum flexion force normalised against the total muscle volume.

The extension and flexion forces normalised against the lean muscle volumes presented similar results, with the OB subjects able to generate the greatest force per unit of lean volume with both the extensor (0.395 ± 0.041 N/cm^3) and flexor (0.463 ± 0.090 N/cm^3) muscle groups. The force per unit of lean volume of the extensor muscle group was comparable between the NW (0.308 ± 0.052 N/cm^3) and DAO (0.316 ± 0.103 N/cm^3) subjects, but the NW subjects generated significantly less force per unit of lean volume than the OB subjects (p = 0.00097), where no statistical difference was found between the DAO and OB subjects. On the other hand, the force per unit of lean volume of the flexor group was similar within the three sub-cohorts (NW: 0.429 ± 0.082 N/cm^3, DAO: 0.401 ± 0.130 N/cm^3), with no significant differences.

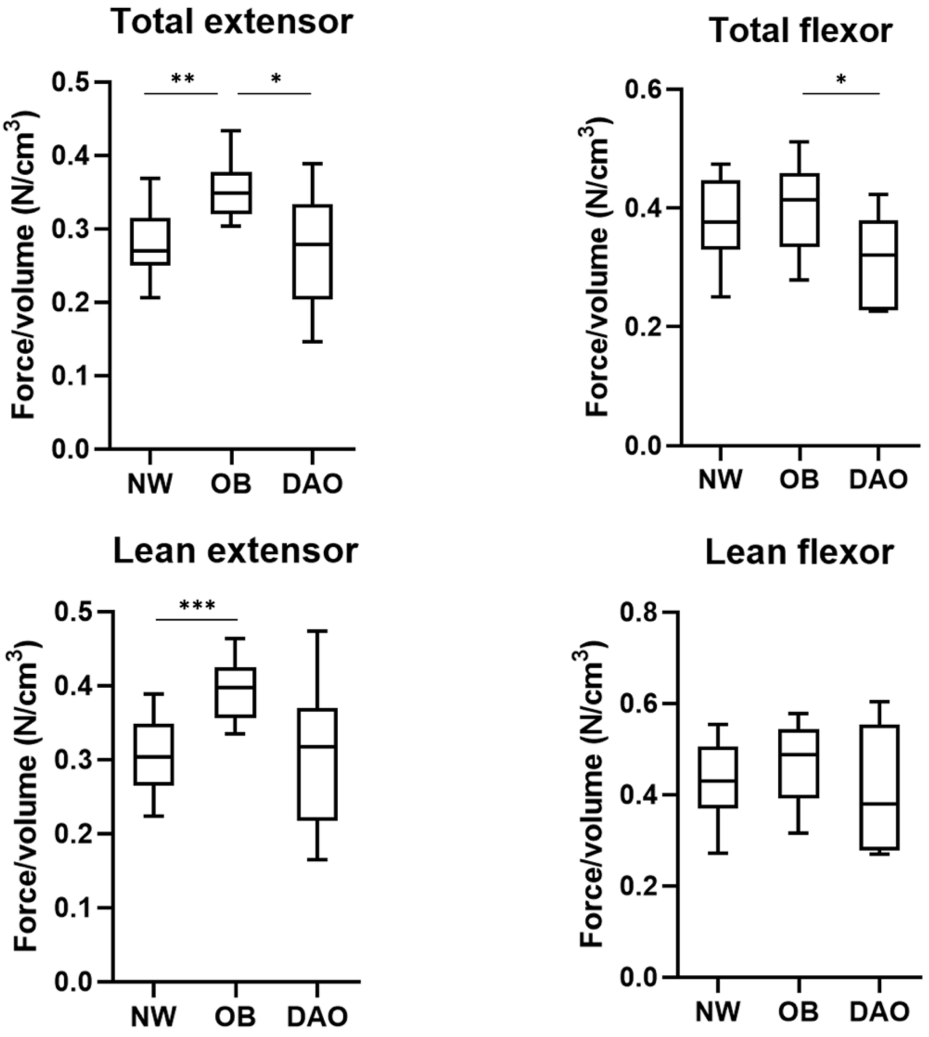


Figure 3.4: The maximum isomorphic flexion and extension force measured, normalised against the total and lean muscle group volumes. The extensor force per unit volumes are presented on the left hand side, and the flexor force per unit volumes are presented on the right hand side. Three boxplots are plotted, representing each of the three study sub-cohorts: Normal Weight (NW), Obese (OB), and Dynapenic Abdominal Obese (DAO).

3.3.4. Discussion & conclusion

In this Preliminary study, the knee flexor and extensors were segmented from MR images. An investigation into the relationship of the muscle volumes found through segmentation and the maximal isometric force was conducted and the level of fat infiltration within the three sub-cohorts was compared, examining the two muscle groups independently in both analyses.

The study found that the flexor muscle group had a higher level of fat infiltration (13.5%) than the extensor group (10.5%), particularly within the OB and DAO cohort (as shown in figure 3.3). This suggests that there is a connection between obesity and fat infiltration of muscle, but this appears to be a particular problem within DAO subjects, that have a higher level of fat infiltration into the knee extensors and flexors than the other groups. The OB subjects were able to produce a greater amount of force per unit lean volume considering both the extensors (OB: 0.395 N/cm^3, DAO: 0.316 N/cm^3, NW: 0.308 N/cm^3) and flexors (OB: 0.463 N/cm^3 NW: 0.421 N/cm^3, DAO: 0.401 N/cm^3). Note that the lean muscle volumes were comparable between the three sub-cohorts. The disparity between the three sub-cohorts could arise from such factors as neurological connections to the muscles that cannot be analysed with available MR images, or differences in the muscle fibre structure [36]. As the OB cohort had the greatest force per unit volume of muscle but had comparable muscle volumes (for both muscle groups), one alternate reason for this higher level of force production could be that the OB cohort undergoes daily strength training through carrying a greater weight, enabling the higher level of force than the normal weight cohort. Further analyses are required to make statistically significant conclusions from the results of this preliminary study. A larger cohort and a more sophisticated segmentation procedure would allow a more detailed investigation, but processing time and operator repeatability issues (particularly within a non-healthy cohort) would limit these investigations, particularly if individual muscles were to be considered.

3.4. Literature review

Manual muscle segmentation, as the results above highlighted, is a powerful tool for scientific discovery, but is limited. The drive for automating the process and alleviating these shortcomings is therefore growing in the research community. The benefits to automating the process are highlighted within the previous sections. Processing more data, more quickly, without operator variability, and segmenting individual muscles rather than muscle groups would enable more disease or disorder specific mechanisms to be understood [59, 60, 96]. Furthermore, a method to quantitatively measure muscles and their structural health and functional capacity, would be beneficial for a more targeted and subject specific intervention strategy [101]. Moreover, this quantitative tool could enhance investigations in the effects of therapeutic interventions for muscle disorders [61, 71, 74].

3.4.1. Existing automatic segmentation pipelines

Many different approaches have been explored to automatically segment muscles from different imaging modalities. These can be split into two overarching categories, 1) traditional, purely mathematical approaches based on image processing, and 2) probabilistic learning-based approaches. These two approaches have some significant benefits and limitations, and there are merits to further exploring both methods' application and addressing their current limitations. Notable studies are highlighted for both approaches in the following sections.

3.4.2. Traditional approaches

There are different traditional image processing techniques used to segment medical images that use mathematical manipulation of the images. The most basic example of this is thresholding: deciding upon a greylevel value, with all pixels within an image being accepted if they fall above or below that threshold. While this approach has been shown to be successful in some medical image segmentation applications, it is unable to perform individual muscle segmentation [102]. The muscle boundaries are not so clear within any type of medical image, and there are a variety of different tissue-tissue boundaries with different characteristics that must be isolated, meaning more sophisticated approaches are required. The two most successful traditional methods used within the literature to automate muscle segmentation are statistical shape modelling, and image registration.

Statistical shape modelling (SSM) entails the generation of an average atlas geometry, which can be scaled or deformed to fit individuals. It has been used to achieve a good agreement of automatically and manually generated segmentations of single muscles from MR images, such as the quadratus lumborum within the lower back by Engstrom et al. [103] where the Dice Similarity Coefficient (DSC), a volumetric and geometrical measure of agreement, achieved was 0.86 ± 0.08 (mean ± standard deviation). The quadratus lumborum is a muscle with a non-complex, truncated cone-like shape that is consistent between individuals and is therefore well suited to automatic segmentation using this technique. However, the large variability of muscle volume and geometry within the lower limb skeletal muscles, even between subjects with similar anthropometric characteristics, limits the application of SSM to segment other muscles. This is highlighted in a study by Andrews and Hamarneh [104], where SSM was used to segment partial sections of 11 muscles in the thigh. Using a similar SSM based method, the achieved DSC was 0.81 ± 0.07 on average, which was significantly lower than that found in the study by Engstrom et al. [103]. One of the reasons for the disparity in the accuracy of results, was that this method was less well suited to the segmentation of many individual muscles with wide variations in structures and shapes.

Image registration is the process of aligning two images. In any registration algorithm there are two inputs: a fixed image (often referred to as the target image), and a moving image (often referred to as the reference image). With image registration, the moving image is deformed, either linearly or non-linearly, to match the fixed image [105]. Image registration has been explored within the literature to perform muscle segmentation. The idea behind this process is that when registering images, features within images are aligned. If the features are known within the moving image, then they can be found through registration, in the fixed image. A simple example of the process of image registration is shown in Figure 3.5, below.

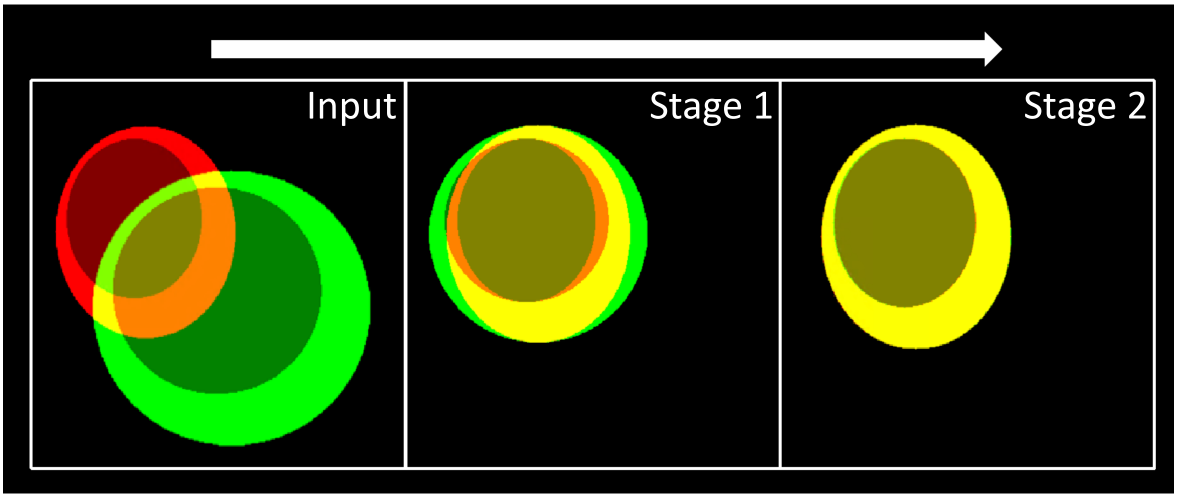


Figure 3.5: Example of registration of two example images, both with two features (circle within a circle). The moving image (shown in green) is deformed into the red image (shown in red) through registration. Through the two-step registration shown, the green image is first scaled and translated, followed by a squeeze in the horizontal direction. The yellow colouration in the resulting image (right hand side) showcases an accurate registration, as the green and red images are perfectly aligned.

All image registration algorithms require a measure of similarity between the fixed and moving images [105]. Maps that displace pixels in the moving image are then tested to maximize (or in some cases, minimize) the similarity measure of the two images. Displacement maps (i.e. for each voxel in the image) are tested iteratively, until the deformed moving image is satisfactorily similar to the fixed image.

Two-dimensional (2D) deformable image registration has been used to generate 3D muscle segmentations. In a study by Ogier et al. [83] 2D image registration was used to propagate segmentations of individual slices into partial sections of 3D muscle geometry using only a few manually segmented slices. The algorithm was used to individually segment the four knee extensor muscles (the three vastii and rectus femoris) quoting an average DSC of 0.91 across the four muscles. This method required the manual segmentation of a small number of MR imaging slices and propagated them to neighbouring images through non-linear registration. Though this method greatly reduces the number of images required to be manually segmented, the accuracy of results would be limited given an area of the body where the shape of the muscles change drastically between sequential 2D images. The authors did not account for these areas as they appear to have segmented only partial sections of the thigh muscles, not the entire muscle structure. Additionally, any inaccuracies in the manual segmentations, due to unavoidable operator variability issues, would be propagated through the 3D reconstructions, which could be a significant source of error.

3D image registration has also been used within longitudinal studies to populate MR images with partial segmentations of a small number of muscles to good effect, such as within the studies presented by Le Troter et al. [94] and Fontana et al. [93]. Both studies used segmented imaging data at an initial observation as a reference to segment the muscles from imaging data acquired at a secondary timepoint. Le Troter et al. achieved a high DSC (~0.90) in the segmentation of the knee extensor muscle group, whereas Fontana et al. achieved a similarly high DSC (~0.87), in the segmentation of the gluteus maximus, gracilis, tensor fascia latae, and sartorius. Though these longitudinal approaches provided insight into the change in muscle characteristics over time, multiple MR image sequences are required from individual subjects at two different timepoints and one dataset must be manually segmented. These limitations prevent large scale automatic segmentation of new subjects. In the literature, inter-subject registration aiming to segment the muscles of a new subject, referencing a previously segmented subject has not yet been fully explored to the best of the author's knowledge.

Image registration has also been used widely with multi-atlas approaches [81, 106-108], which combine the results of multiple different segmentations generated through independent registrations. The locations in the image where the segmentations agree, are collated forming one multi-atlas segmentation [106]. The technique seeks to gather the best aspects from different segmentation results and reduce inaccuracies [106]. Figure 3.6 below shows a mathematical representation of how multi-atlas approaches operate.

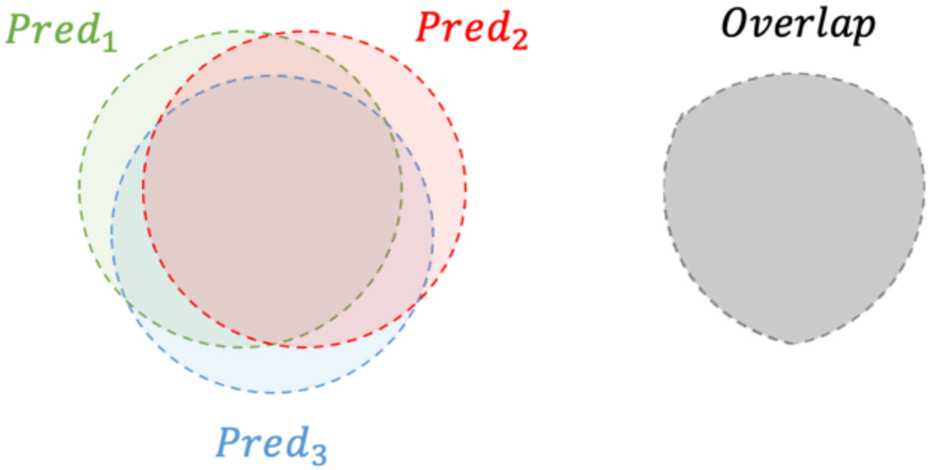


Figure 3.6: Multi-atlas approach for 3 given predictions or segmentations. The overlapping area is labelled as the multi-atlas prediction.

The most prevalent examples of multi-atlas segmentation methods are in the segmentation of brain tissues, such as the Simultaneous Truth and Performance Level Estimation (STAPLE) [106]. This method fuses labels of various areas of the brain by forming a probability map for each pixel, which defines the probability of each pixel belonging to a certain tissue. If the probability of a given pixel belonging to a certain label exceeds a user-stated threshold, the pixel is given that label. Herein lies the benefit of multi-atlas methods: the outlier pixels resulting from each registration are revoked, and the agreeing pixels are retained. This is clear particularly within the brain tissue segmentation methods noted, as the multi-atlas approaches outperform single-atlas registration approaches [109, 110]. However, there are limitations to multi-atlas methods. If the number of atlases combined within the multi-atlas approach increases, the number of disputed pixels (those for which the probability maps fall below the threshold) will also increase. Therefore, the number of atlases must be carefully selected with reference to the user-inputted threshold. Additionally, if the variability within the segmentations resulting from single-atlas registration is high, the multi-atlas results can be overly penalised and not represent the intended segmentation [106].

To the best of the author's knowledge, only two studies reported the use of multi-atlas approaches for muscle segmentation [108, 111], with only one focussing on individual muscles. In that study, Yokata et al. [108] segmented hip and thigh muscles (some grouped) from CT images and achieved an average DSC of 0.83. The study presented a method that automatically segmented the hip and thigh muscles with high accuracy but neglected the muscles within the calf. The study also used CT images, which expose subjects to unnecessary radiation.

3.4.3. Deep learning

Machine learning has been an area of great interest in recent years. In the context of images analysis, Convolutional Neural Networks (CNNs) are used to perform segmentation tasks [112, 113]. Convolutions are simple matrix-oriented operations that alter images in some way. A kernel (a matrix of numbers) is passed over an image, altering the pixel values within an image. Within CNNs, many convolutions are applied sequentially, and weights are assigned to each convolution, with high weights assigned to the convolutions that highlight key features [114]. In the context of medical image segmentation, CNNs are trained by inputting images in batches into the network, and a segmentation is predicted and compared to a ground truth segmentation label. Initial predictions are always poor, but in a process named back-propagation, the weights of each of the tested convolutions is altered. After training an algorithm of a suitable architecture, a new, unlabelled image can be segmented. This 'learning' process, by changing the weights of the convolutions is the reason for these algorithms to be named neural networks, as the system learns in a similar way to the human neural system, connecting combinations of convolutions that can highlight the relevant anatomical features. Deep learning is a more recent adaptation of CNNs, wherein the structure of the networks has many connected layers, which operate in tandem. In addition to the training aspect of CNNs, there are hyperparameters, such as the learning rate, kernel size, and patch size, which must be tuned [115]. To optimize these parameters, a validation dataset is separated from the training data. The validation data is tested regularly within the training process allowing the user to alter the hyperparameters used to train the network. A traditional split is 80-20% training to validation and this is seen throughout the CNN research community.

There are a multitude of studies available that have applied deep learning in the context of tissue segmentation. Three areas that have an extensive number of studies aiming to perform segmentation using deep learning are brain (450+), organ (500+) and bone segmentation (150+) [116]. Though, there are significantly fewer studies aiming to segment the lower limb muscles (50+) and even fewer where the input images are MR images [28]. While exploring this literature, it became clear that few studies aimed to segment all muscles individually. Mostly, studies segmented the entire muscle body, which is useful if overall muscle health and volume were to be assessed in longitudinal studies but gives limited insight on individual muscle function. There are, however, peer-reviewed studies that segmented individual muscles from 3D MR imaging data. Three examples of deep learning models used to segment muscles from MR images will be analysed in this section.

In 2017, a study published by Ghosh et al. [117] aimed to segment five individual muscles from a full lower limb MR imaging sequence. In the study, fat-suppressed imaging data was taken from an unspecified number of healthy athletes, resulting in a total of 700 MR images for each muscle. Ghosh et al. made use of AlexNet, one of the first deep CNNs available in the literature [118].The study involved the segmentation of 5 muscles (adductor longus, gracilis, sartorius, rectus femoris and vastus medialis) from 17 sets of MR imaging data, acquired from young healthy athletes. Although it is unclear the number of subjects that were tested, the authors quoted an 80-20% split in training and validation. The authors reported an average segmentation accuracy of 0.87 DSC; some results are visualised in Figure 3.7.

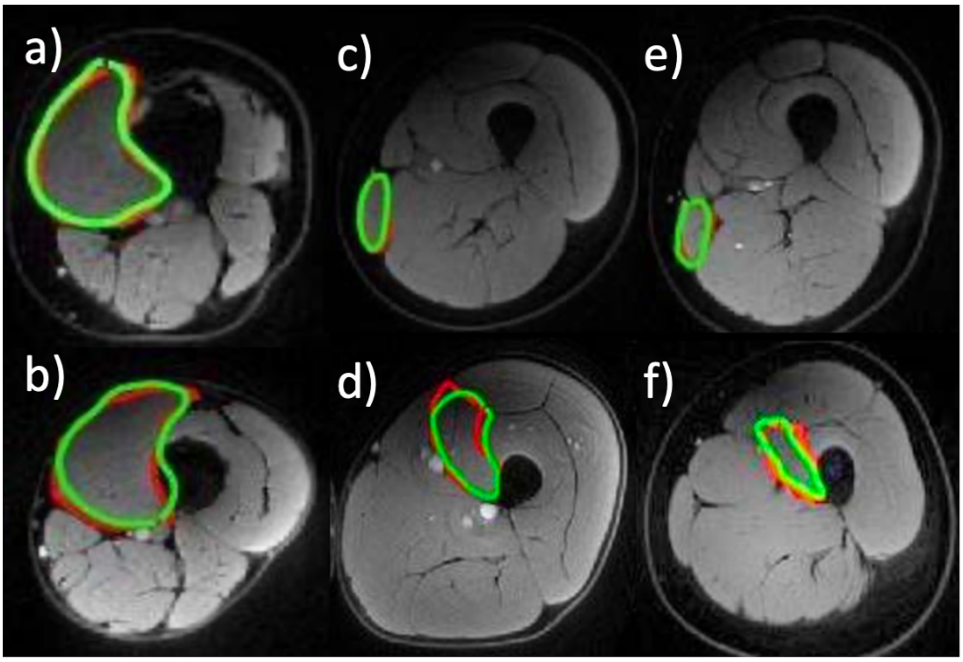


Figure 3.7: Visual results in the study by Ghosh et al., six two-dimensional image slices along the thigh, showcasing the vastus medialis (a, b), gracilis (c, e) and adductor longus (d, f). The ground truth segmentation approach is shown in red, and the predicted segmentation is shown in green. Image acquired from Ghosh et al. [117].

To conclude, this study demonstrated the application of deep learning within the problem of muscle segmentation with a moderate to high increase in accuracy. The dataset used was well suited to this task, given the high number of fat-suppressed, high-resolution images of healthy athletes. Though to note the limitations of the study, only five muscles were segmented from these images, and the CNN architecture has since become out of date, with more recent architectures enabling results of even higher accuracy within other studies.

One such study, was that published by Ding et al. [80] used the UNet structure, a more modern deep learning model [114]. The UNet is a convolutional neural network model designed specifically for problems concerning medical imaging data. It has been used for a multitude of other medical imaging segmentation tasks and has been shown to be very powerful in this application [119]. The model was built around the theory that the first half of the 'U shape', where the input images are down sampled and reduced in size, allows the neural network to recognise the larger, more global features within the images, for a given segmentation task. The latter half of the U shape, up sampling, allows training of the neural network to recognise locally important features, such as feint intermuscular boundaries. In the study by Ding et al., two pre-processed MR images with different acquisition methods are inputted into the UNet CNN as one, double channel image (see Figure 3.8). The two image acquisition methods used in this study were water-suppressed and fat-suppressed. Having both MR imaging acquisition methods allowed additional anatomical features (fat and bone) to be easily removed from the images, in a pre-processing step. Additionally, the use of two acquisition modalities allowed the neural network to learn the features within the image with two references to draw from, increasing the likelihood of the required features being meaningfully learned [120].

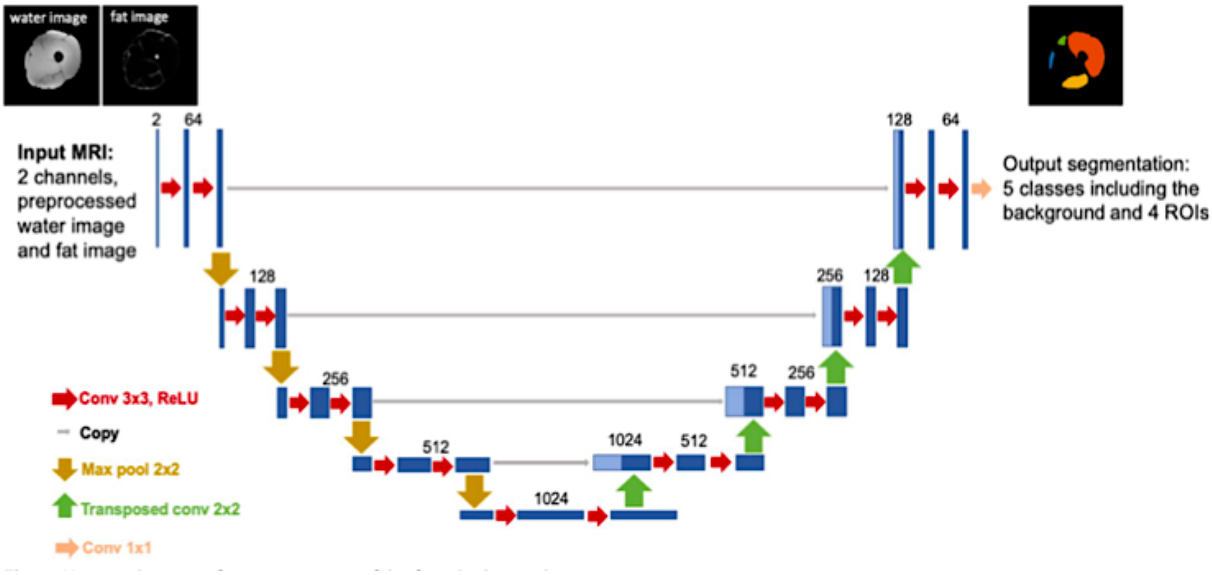


Figure 3.8: Schematic of the study by Ding et al. [80], showing the double channel MR image, with its attributed manual segmentation (left). These inputs are inputted into the U-Net structure, allowing the feature extraction to be learned.

The authors used the UNet model to segment four regions of interest: the knee extensors (the three vastii and rectus femoris), sartorius, gracilis and hamstrings (a grouped term for the semitendinosus, semimembranosus, and biceps femoris). The reason for why these muscles were selected specifically was not clear, although this group of functional muscles play major roles in the gait cycle [121]. Therefore, understanding fat infiltration for these muscles was important, but including a higher number of muscles would increase the impact of the study. The network was trained to find each of the included regions of interest within the images using a multi-coloured mask, with the 4 regions of interest each encoded in a separate channel (see Figure 3.8). This is known as multi-class segmentation, distinct from the work of Ghosh et al. [117], who trained individual networks for each muscle. The cohort of the study was derived from two different databases, with different inclusion criteria. The first dataset was an open-source public reference dataset named MyoSegmenTUM [122], consisting of 15 healthy volunteers (21.1 ± 7.7 years old) and 4 subjects with neuromuscular conditions (52.8 ± 8.9 years old). The second dataset was gathered from local clinical data, consisting of 21 subjects (56.4 ± 14.5 years old), with all but seven of these subjects being diagnosed with a muscle disorder. The inclusion of two distinct groups introduced a more challenging problem for the CNN as there is a greater number of patterns to learn, resulting in a more versatile tool. The network was trained using 23 fully segmented datasets and used to predict the segmentations of 7 individuals. The accuracy of the segmentation was stated as >0.85 in terms of DSC. Although the segmentation accuracy was high, the segmentation of 2 individual muscles was a limited result and further investigation of this method would be required to perform automatic muscle segmentation.

Ni et al. in 2019 [82] used a deep learning CNN model to individually segment all individual muscles from fat-suppressed MR images acquired from a cohort of 64 athletes (51 and 13 selected for training and testing, respectively). The authors used a similar neural network structure as the previous study. The UNet model used had some major alterations in order to accept 3D images [123] and target all major lower limb muscles. Each individual muscle was automatically segmented, as opposed to the all-at-once approach offered by Ding et al. [80], using a two-stage CNN (summarized in Figure 3.9). The first network was used to crop the raw 3D input images, producing a smaller image containing only the muscle to be segmented. This resulted in 35 datasets, each consisted of 64 3D images, one for each subject in the cohort. The second network performed the individual segmentation. CNNs were trained for each muscle, each geared toward learning the key features of individual muscles.

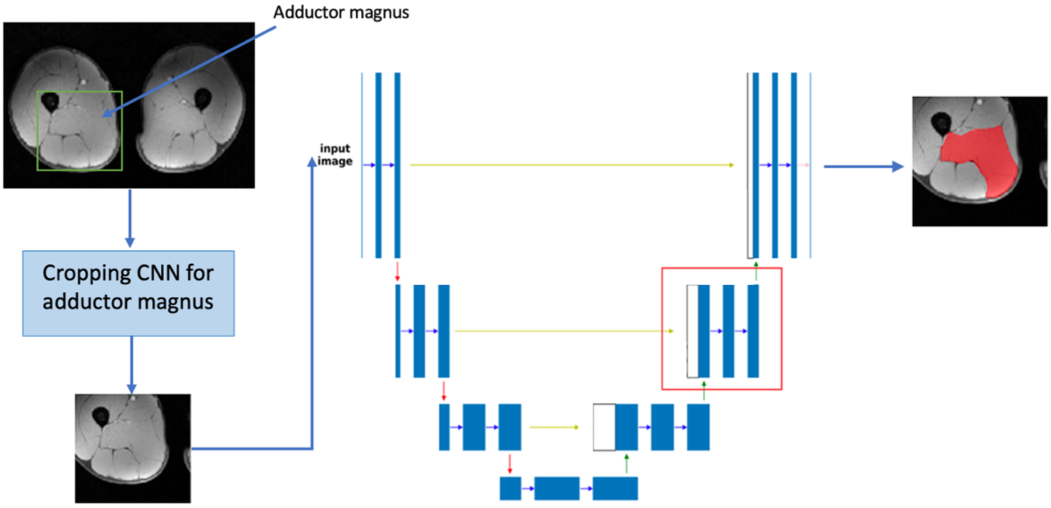


Figure 3.9: A schematic of the method used in the paper by Ni et al. The top left image shows a 2D slice of the raw input data for a given subject (the adductor magnus is highlighted), which is inputted into a cropping CNN trained for each muscle (example shown is the adductor magnus). The cropped images are inputted into the segmentation CNN, the result of which (after training) is a segmentation map, shown as the output. Image adapted from [82].

Subjectively, the results of the paper were highly impressive as shown in Figure 3.10 below. The DSC for the predictions of each of the 35 muscles included in the study were all around 0.9, which was comparable to the DSC found between segmentations performed by multiple operators (also included in the study) [80].

The muscles of the athletes were known to be 'well-developed' and 'more compact' than the general population [124]. Using a fat-suppressed MR imaging acquisition method on athletes such as these highlighted muscle boundaries. Given that the muscle body was so compact, the boundaries generally appeared homogenous in the raw images (see Figure 3.10). The authors' choice of cohort, number of subjects, imaging modality, and segmentation method all worked together to enable the high level of segmentation accuracy. An average DSC of around 0.9 was achieved across the 35 muscles segmented. Although the scope was limited, this study proved the concept that deep learning (in particular the UNet model) can be used to perform automatic muscle segmentation. The requirement of large amounts of labelled data limited the application of this approach. In addition, the impact of a mixed database, such as one containing subjects with muscle disorders, was not clear from these results. Moreover, the decision to train an individual network for each muscle within the lower limb was not well-justified. This choice incurred extra computational expense, which limited future applications using the same method. Nevertheless, the results of this study were compelling and warrant further investigation in its application to other subject groups.

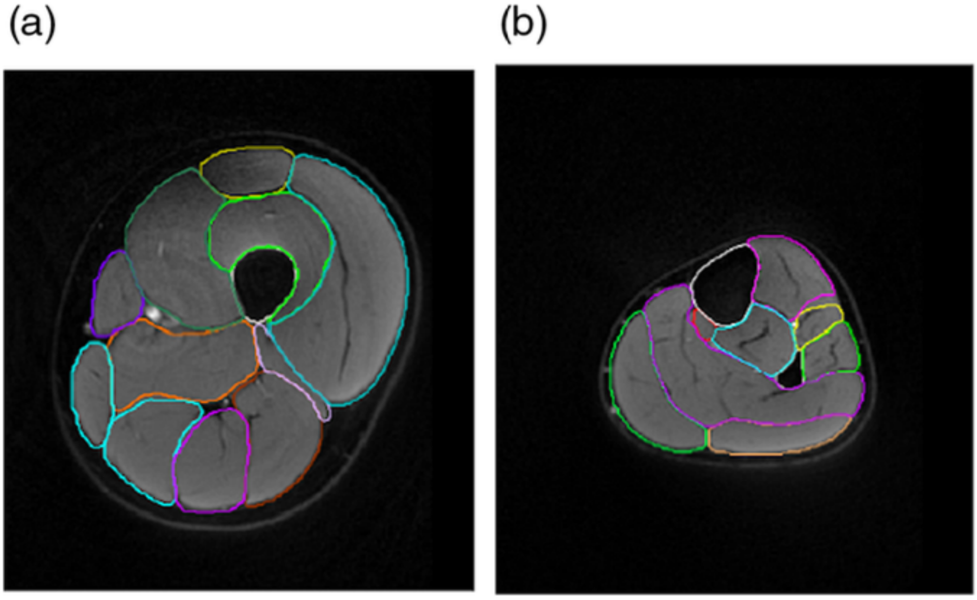


Figure 3.10: Overlapping raw MR images with segmentation results from the proposed method in the thigh (a) and calf (b). Image acquired from Ding et al. [80].

3.4.4. Summary and gap in the literature

Overall, while there have been considerable advancements in the search for a tool to automatically segment muscles, there is yet to be one solution fit for widescale use. Table 3.2 summarises the studies reviewed in this chapter, highlighting the muscles segmented and the overall accuracy reported.

Traditional approaches are yet to be fully explored. Image registration has not yet been tested for the segmentation of all individual muscles within the lower limbs. Simplistic applications may benefit greatly from the use of image registration, such as the segmentation of one limb using the contralateral limb as a reference. This application alone could halve the time taken to perform muscle segmentation (if the accuracy of the approach is similar to that of the gold standard process) and should therefore be tested. Furthermore, using image registration to segment all lower limb muscles using other, previously segmented subjects is yet to be explored. Though, noting the variability in muscle structure within subjects even from the same cohort, image registration may be limited in this respect, but this cannot be confirmed without testing.

On the other hand, deep learning-based methods are the focus of the field in its recent years and can give accurate segmentations in many medical image analysis applications. These methods, however, are not applicable for smaller cohorts, due to the requirement of large amounts of training data [125]. There are methods to overcome the requirement of large training databases, such as data augmentation, but this is yet to be fully explored within the context of muscle segmentation. Additionally, the CNN architectures currently used, particularly in the study by Ni et al. [82], incurred vast computational expense. The traditional UNet architecture has been shown effective in the segmentation of muscles from MR images but reducing the computational expense with more succinct network structures could widen the reach of these methods. A novel network architecture that is more targeted to the problem of muscle segmentation could reduce computational expense, and data augmentation could reduce requirements large manually segmented databases for training.

Explicitly, within the clinical domain, tissue segmentations must be as accurate as possible, if indeed the results were to be used to inform clinical decisions. Therefore, the aim in this context would be to achieve accuracy measures equal to the operator variability. For more challenging cohorts, such as those with muscle disorders or older individuals [28], an automatic segmentation tool should be capable of capturing the muscle volume with < 10% error, DSC > 0.85, and HD < 10mm, for each muscle, as would be satisfactory in an inter-operator repeatability analysis [28]. For less challenging cohorts, such as young healthy athletes, or children, an automatic segmentation tool should be capable of capturing the muscle volume with < 5% error, DSC > 0.9, and HD < 5mm, for each muscle, as would be expected in an inter-operator analysis.

Study	Muscles	Images	Accuracy	Method
Engstrom et al.	Quadratus lumborum	T1-weighted MR	DSC 0.86 ± 0.08	SSM
Andrews & Hamarneh	All thigh muscles (n=11)	T1-weighted MR	DSC: 0.81 ± 0.07	SSM
Ogier et al.	Knee extensors	T1-weighted MR	DSC: 0.91	2D image registration
Le trotier et al.	Knee extensors	T1-weighted MR	DSC: 0.9	longitudinal image registration
Fontana et al.	gluteus maximus, gracilis, tensor fascia latae, and sartorius	T1-weighted MR	DSC: 0.87	longitudinal image registration
Ghosh et al.	Adductor longus, gracilis, sartorius, rectus femoris and vastus medialis	Fat suppressed MR	DSC: 0.87	AlexNet
Ding et al.	Knee extensors, knee flexors, gracillis, sartorius	Fat & water suppressed MR	DSC: 0.85	2D Unet
Ni et al.	All lower limb (n=35)	Fat suppressed MR	DSC: 0.9	3D UNet

Table 3.2: Comparison between the studies highlighted within the literature review. Showcasing the authors, muscles, images, accuracy and methods used.

3.5. Aims & objectives of the thesis

Motivated by the preliminary study, problems associated with the gold standard approach, and the gaps identified within the literature surrounding muscle segmentation, the following experimental chapters of this thesis aim to test the accuracy of two algorithms, one based on image registration and the other using deep learning, to automatically segment individual muscles of the lower limb. To this aim, the following objectives were defined:

1) Build an automatic segmentation pipeline using image registration. Optimize the registration parameters to the task of segmenting muscles from MR images by registering images of the left limb of each subject to the right limb.

2) Apply the automatic segmentation pipeline using image registration to segment all lower limb muscles from subjects using other subjects as references.

3) Build deep convolutional neural networks, following state of the art architectures to perform muscle segmentation. Incorporate novel strategies to enhance the networks, addressing the need for extensive labelled databases & high computational power.

4) Compare each of the segmentation methods to inform future studies as to which method should be used for future studies.

Chapter 4:

Optimizing a muscle segmentation pipeline using deformable image registration

This chapter is partially based on a paper published in PLoS One (2023): 'Deformable image registration based on single or multi-atlas methods for automatic segmentation and the generation of augmented imaging datasets' by **W. H. Henson**, C. Mazzà, E. Dall'Ara. Doi: https://doi.org/10.1371/journal.pone.0273446

4.1. Introduction

Deformable image registration has proven useful in the application of muscle segmentation and has been shown to be capable of segmenting muscles in cases where the variability between inputs is not high, such as within longitudinal studies [93, 94]. However, image registration has not yet been used to segment all lower limb muscles from Magnetic Resonance (MR) images of new subjects, using previously segmented subjects as references. This would be of great aid for studies requiring muscle characteristics, as new subjects could be segmented automatically, with little to no operator input.

Registering the images of an atlas subject with those of a new subject (inter-subject registration) is likely to be far more difficult than it is when used in longitudinal studies. The reason for this is that the variability in the distribution of anatomical features and anthropometric characteristics between different subjects is far greater than that of one subject at two different time points [66, 93]. Even subjects within the same age range and Body Mass Index (BMI) categories have widely varying muscle volumes, shapes and structures [126] meaning that the use of image registration if used in this application would be required to overcome great differences even within the simplest of cases. If image registration was to be shown capable of overcoming this variability and capturing the muscle geometry, then new subjects could be segmented without any input from operators. Nevertheless, the application of image registration to propagate segmentations to new subjects should be explored.

Another potential application of image registration to perform muscle segmentation is to segment one limb using the opposing limb as the reference. The variability in muscle structure between left and right limbs has been shown to be significant, but far less than between different subjects [66]. Additionally, the distribution of tissues visible within medical images (e.g. muscle, fat, intramuscular fat, skin) is comparable between the left and right limbs of subjects, where this is known to be much more different between subjects [127]. Moreover, the difference in anthropometric characteristics between subjects can be great even considering subjects within the same cohort, but this would not cause issues with registration of contralateral limbs. The automatic segmentation of one of the lower limbs requiring only a segmentation of the opposing limb would half the processing time of manually segmenting the muscles from a full lower limb dataset, given that the segmentation is accurate.

There are requirements of image registration algorithms that must be met prior to use. Firstly, the muscles under investigation (all lower limb muscles) must be visible within the medical image. The images used throughout this thesis were captured in multiple sections, and these must first be stitched together, ensuring all muscles are visible in the images. The aim of the thesis is to produce an automatic tool for muscle segmentation, and therefore, this must be operator independent. Secondly, the user selected parameters used for the registration must be optimised, to maximise segmentation accuracy. These two requirements must be satisfied before testing the automatic segmentation tool.

Therefore, this chapter seeks to answer two research questions. 1) Can deformable image registration be used to segment individual muscles of the lower limb starting from the segmentations of the opposing limb of the same subject? 2) Can deformable image registration be used to segment the muscles in one subject using the segmentation maps from another subject?

4.1.1. Aims and objectives

In order to answer the research questions of this study, the study aims at developing an automatic pipeline to segment individual muscles by using deformable image registration. The specific objectives of this chapter are:

1) Pre-process raw MR image sequences automatically.

2) Develop an automatic muscle segmentation pipeline using deformable image registration.

3) Optimise the registration parameters using the intra-subject (left to right) muscle segmentation.

4) Test the accuracy of intra-subject segmentation method.

5) Test the accuracy of inter-subject registration method.

4.2. Methods

4.2.1. Objective 1: Automatic pre-processing of images

4.2.1.1. Subjects and image acquisition

Lower limb T1-weighted MR images were acquired in a previous study [66] from 11 post-menopausal women (mean (standard deviation): 69 (7) years old, 66.9 (7.7) kg, 159 (3) cm) using a Magnetom Avanto 1.5T scanner (Siemens, Erlangen Germany), with an echo time of 2.59 ms, repetition time of 7.64 ms, flip angle of 10 degrees. The study was approved by the East of England – Cambridgeshire and Hertfordshire Research Ethics Committee and the Health Research Authority (16/EE/0049) and conducted in accordance with the Declaration of Helsinki (October 2000), after gaining written informed consent. The MR images were acquired in four sequences, capturing the hips, thigh, knee, and calf. To reduce scanning time while still providing detailed geometries of the joints for use within other studies, the joints were acquired with a higher resolution (pixel size 1.05 mm^2, slice spacing 3.00 mm) than the long bone sections (pixel size 1.15 mm^2, slice spacing 5.00 mm). A sub-cohort of 5 of the 11 subjects were selected for automatic segmentation. The five subjects were chosen with the aim of creating a sub-cohort with a wide anatomical diversity, including the tallest and shortest individuals (154.0 cm, 164.2 cm), subjects with the lowest and highest BMI (21.2, 32.1), and the youngest and oldest participants (59, 83 years). The oldest participant was also the shortest participant within the cohort. Each subject was used as both a target and a reference for the image registration algorithm, creating 20 subject pairings for the inter-subject analyses.

4.2.1.2. Pre-processing of images

Firstly, each image within the sections of MR imaging data was linearly interpolated such that the resultant pixel size was $1.00 \times 1.00\, mm^2$, for all images in each of the sections, allowing concatenation of the images. The quality of the images was not altered, as the resolution for the long bones and joints were $1.15 \times 1.15\, mm^2$ and $1.05 \times 1.05\, mm^2$ respectively, resulting in small changes to the resolution of the images. Secondly, the voxel sizes were made isotropic. The image sequences capturing the long bones and joints had different slice thickness, $5.00\, mm$ for the long bones and $3.00\, mm$ for the joints. To create an isotropic representation, copies of each image were made to reduce the slice thickness to $1.00\, mm$. Through these processes, the voxel size within the different imaging sequences were homogenised to be $1.00 \times 1.00 \times 1.00\, mm^3$. Each image within a long bone section was therefore copied four times and within a joint section was copied twice as shown in Figure 4.1. The fields of view of each MR imaging sequence were homogenised, retaining the spatial location of the image contained in the metadata. Each image was wrapped in artificial blank data (grey level value of 0, equal to that of air) to match the sequence with the widest field of view, which was typically the hips.

In the areas of the body where the field of view in the longitudinal direction of the image sequences from two sections overlapped, half of the overlapping area was used from the two overlapping sections (Figure 4.2). The removal of half of the overlapping area reduced the effect of MR imaging bias [128], which can be seen in the highest and lowest images of the T1-weighted scans (Figure 4.3). Before concatenating the two sequences, the last and first images of two adjacent sequences were registered in order to enforce a simple and linear translation, using the “imregister” function in MATLAB (Image processing toolbox, version 2019b and above). The translation was then applied to each image within the lower imaging sequence to align the subject anatomical data. This simple linear registration was required to align the different image sequences, as initial attempts to combine them resulted in a linear jump at the junction of the two sequences, likely due to slight movement of the subject within the MR scanner. After application of the linear translation, all image sequences were concatenated, forming a continuous, isotropic 3D image of the lower limb.

All processes used in this section were designed to be generic (as opposed to subject specific), which allows this pre-processing method to be used for any multi-sequence MR imaging acquisition of the lower limbs.

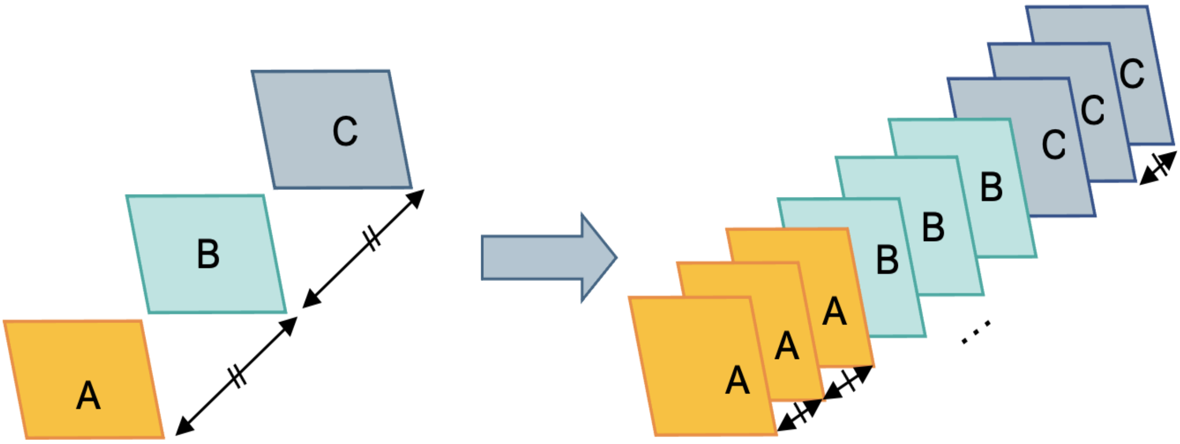


Figure 4.1: The process of creating isotropic data by generating copies of each image. In this case, the initial spacing between the slices was 3mm, resulting in two additional copies of each image.

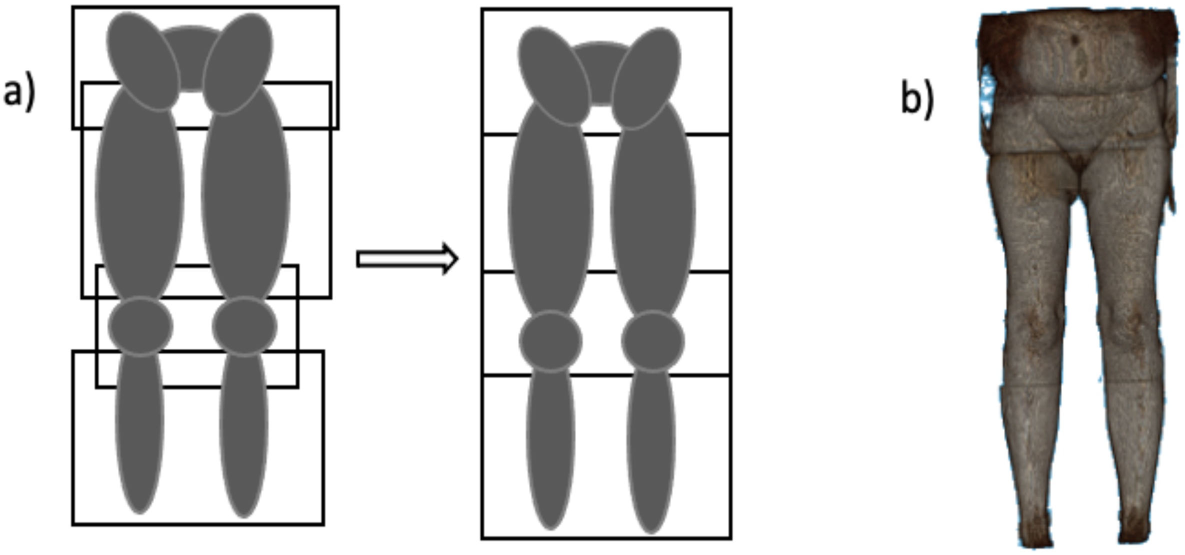


Figure 4.2: Aligning MR imaging sections. (a): Schematic to highlight the problem that all sequences (hips, thigh, knee, and calf) overlap and have different fields of view (a, left) and target solution after the pre-processing operation outlined in section 2.2 (a, right). (b): output of the operation with real MR dataset.

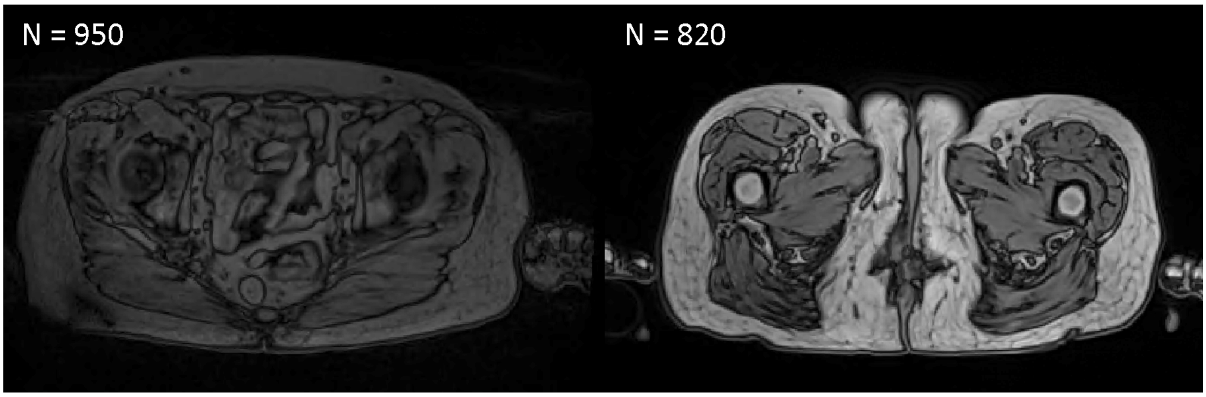


Figure 4.3: Observation of the biased field within the extrema of the thigh section of a random subject. At slice number (N) 950, the tissues appear with skewed greyscale values, where in the lower slice number (N=850), near the centre of the thigh section (along the longitudinal axis) the differences in the greyscale values between the tissues are far clearer.

4.2.1.3. Reference muscle segmentations

The 25 lower limb muscles that were visible within the pre-processed image sequences were semi-automatically segmented, using Materialise Mimics [129]. The semi-automatic tool operated by estimating approximate areas of the images that could belong to each of the muscles. The estimations were then manually altered to reflect the muscles. This operation often required significant manual input, incurring around 10 hours of operator interaction time per subject [66]. Ten of the 35 lower limb muscles were removed as they were deemed not visible within the scans, either because they lay outside the field of view, or were too small to be segmented. The repeatability of the manual segmentation process was characterised in Section 3.2.2.3 (see Table 3.1), in a study by Montefiori et al. [66]. As the manual segmentations were used in the automatic segmentation pipeline, those muscles could not be manually segmented with an acceptable level of repeatability were removed. The muscles for which both the inter and intra-operator Coefficient of Variation (CoV) across three repeated segmentations was above 10%, were removed from further study. For example, the gluteus minimus was excluded from further study. Additionally, the gluteus medius was included in the study by Montefiori et al. [66], but it was found to partially extend beyond the field of view of the images of some subjects and so was also removed from further study.

4.2.2. Objective 2: Registration and segmentation pipeline

An overview of the segmentation pipeline is presented below in Figure 4.4. There are two inputs in all registration algorithms, the target and reference images [105]. Within the registration algorithm, the reference image is deformed into the target image, via the calculation of displacement vectors at points within the reference image that are matched with similar points within the target image. These displacement vectors are usually calculated at nodes within a nodal grid defined at points within the images to limit computational expense for extremely large images. The displacement vectors at each node are isolated and applied to the muscle contours within the reference subject, wherein these contours are morphed to provide a predicted segmentation of the muscles within the target subject. The details of the registration and map application algorithms are expanded upon in the section following.

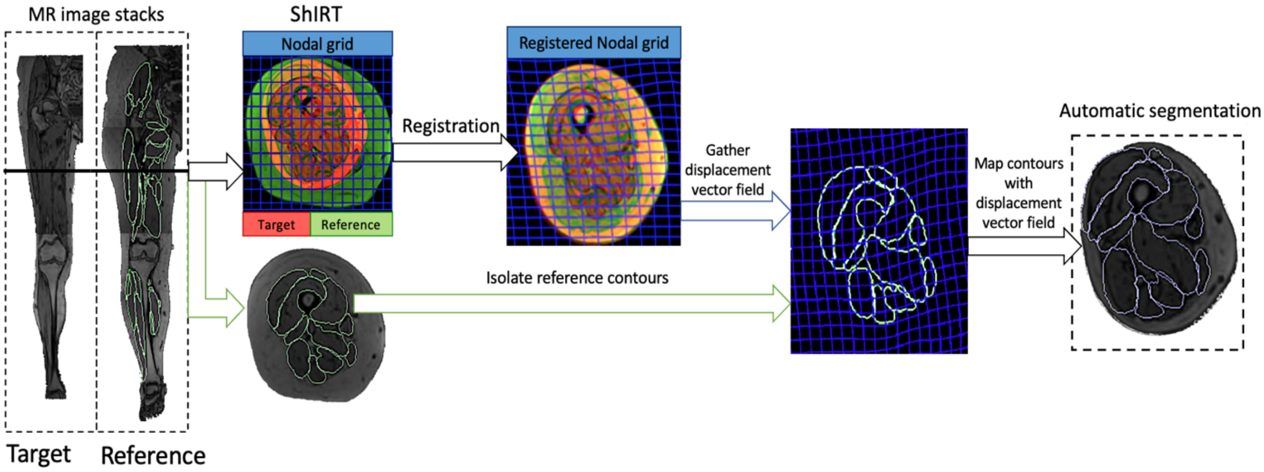


Figure 4.4: Schematic of the segmentation pipeline. After the combination operation, the target and reference imaging data were inputted into the Sheffield Image Registration Toolkit (ShIRT) and registered. The displacement vector field found through registration is applied to the muscle contours of the manually segmented reference subject, resulting in an automatic segmentation of the target subject.

4.2.2.1. Deformable image registration algorithm

Following pre-processing, subject imaging data was registered using the Sheffield Image Registration Toolkit (ShIRT) [130]. In this process a reference three-dimensional image (also referred to as the moving image) was registered to a target image (also referred to the fixed image), aligning the two images. ShIRT was used as previously developed and in the following section, a brief description of the main principles of the registration are described.

ShIRT solves the registration equations at nodes of a grid overlapped to the fixed and moved images, with distance between the nodes called the 'nodal spacing' (NS). If the grey-level intensities are given within the fixed image ($\boldsymbol{f}$) and moving image ($\boldsymbol{m}$) as $\boldsymbol{f}(x, y, z)$ and $\boldsymbol{m}(x, y, z)$ respectively, the displacement between the two intensities is:

$$\boldsymbol{f}(x,y,z)-\boldsymbol{m}(x,y,z)$$
$$=\frac{u(x)}{2}\left(\frac{\partial\boldsymbol{f}}{\partial x}+\frac{\partial\boldsymbol{m}}{\partial x}\right)+\frac{v(y)}{2}\left(\frac{\partial\boldsymbol{f}}{\partial y}+\frac{\partial\boldsymbol{m}}{\partial y}\right)+\frac{w(z)}{2}\left(\frac{\partial\boldsymbol{f}}{\partial z}+\frac{\partial\boldsymbol{m}}{\partial z}\right)-\frac{\boldsymbol{s}(x,y,z)}{2}(\boldsymbol{f}+\boldsymbol{m})$$

Equation 4.1

Where the first three terms on the right-hand side describe the spatial map from the fixed to the moving image, with respect to the three spatial directions. The partial differential terms represent the grey-level intensity gradients in the respective spatial directions. The functions u, v, and w describe the displacement field and are linearly interpolated between the nodes of the grid calculated at regular intervals that map the moving image to the fixed image. The last term ($\boldsymbol{s}$) within Equation 4.1 represents any residual differences in intensity, which accounts for potential changes in grey-levels between the two registered images. Each of the unknown quantities: u, v, w and $\boldsymbol{s}$ are found at each node, i.

$$u=\sum_{i\in N}\frac{\partial a_i}{\partial x}, v=\sum_{i\in N}\frac{\partial a_i}{\partial y}, w=\sum_{i\in N}\frac{\partial a_i}{\partial z}, s=\sum_{i\in N}\frac{\partial a_i}{\partial s}$$

Equation 4.2

Where N is the number of nodes within the cubic grid of user defined nodal spacing (NS). The functions a_i, represent the mapping functions to be found for each node i and in each of the three spatial dimensions. Equation 4.1 can then be rewritten in matrix form, as:

$$\boldsymbol{f}-\boldsymbol{m}=\boldsymbol{Ta}.$$

Equation 4.3

Where,

$$\boldsymbol{T}=\left[\left(\frac{\partial f}{\partial x}+\frac{\partial m}{\partial x}\right)\ \left(\frac{\partial f}{\partial y}+\frac{\partial m}{\partial y}\right)\ \left(\frac{\partial f}{\partial z}+\frac{\partial m}{\partial z}\right)\ -(f+m)\right],$$

$$\boldsymbol{a}=\frac{1}{2}\left[\sum_{i\in N}\frac{\partial a_i}{\partial x}\ \sum_{i\in N}\frac{\partial a_i}{\partial y}\ \sum_{i\in N}\frac{\partial a_i}{\partial z}\ \sum_{i\in N}\frac{\partial a_i}{\partial s}\right]^T$$

Equation 4.4

ShIRT was used to optimise a solution for the set of displacement vectors, a, shown in vector form in Equation 4.4. The optimisation process within ShIRT is conducted using an iterative process to reduce a sum-of-squared-differences cost function, Q, defined in Equation. 4.5. A smoothing coefficient (λ) for the displacement function was introduced into the cost function to adjust the map, where the non-linearity of the displacement field is altered in response to the smoothing coefficient. The initial optimal value of the smoothing coefficient is automatically calculated within ShIRT as the value of λ that minimizes $\boldsymbol{T}^t\boldsymbol{T} + \lambda\boldsymbol{L}^t\boldsymbol{L}$, where $\boldsymbol{L}$ is the Laplacian operator, and t refers to the matrix transpose. The optimal value of the smoothing coefficient was verified in a sensitivity analysis. The cost function is defined as follows:

$$Q = \sum_{all\ voxels} \left(\boldsymbol{f} - \boldsymbol{m}^T(\boldsymbol{a})\right)^2 + \lambda\boldsymbol{a}^t\boldsymbol{L}^t\boldsymbol{L}\boldsymbol{a}$$

Equation 4.5

Where Q is the cost function to be optimised, $\boldsymbol{f}$ as above, represents the fixed image, $\boldsymbol{m}^T(\boldsymbol{a})$ represents the moving image with the mapping, $\boldsymbol{a}$, found through registration applied, and $\boldsymbol{L}$ is the Laplacian operator. The term $(\boldsymbol{f} - \boldsymbol{m}^T(\boldsymbol{a}))^2$ quantifies the squared difference in intensity between the fixed image and the moving image after the mapping was applied. The aim of the iterative process was to reduce the cost function, Q, such that for an estimate of the map $\boldsymbol{a}_n$, the subsequent solution of the iterative process was defined

$$\boldsymbol{a}_{n+1} = \boldsymbol{a}_n + \Delta\boldsymbol{a},$$

Equation 4.6

With the two conditions:

$$\boldsymbol{f} - \boldsymbol{m}^T(\boldsymbol{a}_{n+1}) \leq \boldsymbol{f} - \boldsymbol{m}^T(\boldsymbol{a}_n),$$

Equation 4.7

$$\lambda\boldsymbol{a}_{n+1}^t\boldsymbol{L}^t\boldsymbol{L}\boldsymbol{a}_{n+1} \leq \lambda\boldsymbol{a}_n^t\boldsymbol{L}^t\boldsymbol{L}\boldsymbol{a}_n$$

Equation 4.8

Therefore, at each stage of the iterative process, there was a reduction in the cost function:

$$Q_{a_{n+1}} \leq Q_{a_n}$$

Equation 4.9

meaning a greater similarity between the fixed image, f, and the moving image with the map applied, $m^T(a)$. The iterations halt when either one of two conditions are met: 1) the number of iterations reaches a certain, user-imposed value (e.g. n=100), or 2) that the change in the average value of the displacement vector (Δa) fell below 0.1 voxels (0.1mm). The second halted the iterations in all cases, meaning that all registrations had converged.

4.2.2.2. Application of ShIRT to segment individual muscles

The 5 subjects included in the study were automatically segmented, using each of the 4 other subjects as references. To perform this operation, the displacement vector field obtained through deformable registration was applied to each of the 23 individual manually segmented muscles within the moving subject, deforming them to represent the muscles in the target subject. The manual segmentations are represented as points connected by vertices that mark the boundary of each muscle within the reference imaging data. After the reference images were registered to the target images, the point cloud representing each muscle within the reference image was then displaced in response to the registration map found through registration. Mathematically, this process is expressed as follows. Given a set of points representing an arbitrary muscle within the reference subject, R, each point $R_i = (x_i, y_i, z_i)$, gives the coordinates of a point in the space (index i identifies individual points in the set). Each element of R, is then displaced through application of the displacement vector field (or map) a, found through registration. For this, the nodes, N_j and paired vectors, a_j (where $j \in (1,2,\ldots,8)$), surrounding each point within each muscle were found and the displacement vector, $\boldsymbol{v}_i$, of each point within the muscle boundary, R_i, was found through linear interpolation (same assumption in the ShIRT algorithm), following Equation 4.10. Upon addition of the resulting displacement associated with each point $\boldsymbol{v}_i$, with the original coordinates of R_i the result is a transformed muscle boundary (P_i) representing the same muscle within the target image.

$$P_i = R_i + \boldsymbol{v}_i,$$

Equation 4.10

where,

$$\boldsymbol{v}_i = \left[v_{x,i}, v_{y,i}, v_{z,i}\right]$$

$$= \frac{1}{NS^3}\left[\begin{array}{ccc} \sum_{j=1}^{8}\left|N_{x,j} - R_{x,i}\right|a_{x,j} & \sum_{j=1}^{8}\left|N_{y,j} - R_{y,i}\right|a_{y,j} & \sum_{r=1}^{8}\left|N_{z,j} - R_{z,i}\right|a_{z,j} \end{array}\right]$$

Equation 4.11

Through the process outlined in Equation 4.11 and 4.10, displacement vectors $(\boldsymbol{v}_i)$ were found for each point (i) in the point cloud (R), and applied to define a new, deformed point cloud (P_i) representing the automatically generated muscle segmentation. These displacement vectors $(\boldsymbol{v}_i)$ were calculated by summing the contributions to each displacement vector from each of the 8 surrounding nodes (those that form a cube around each point), scaled linearly by the distance between each node and the point being deformed. The map application process outlined was performed for 23 out of the 35 visible muscles within the lower limb [66].

4.2.2.3. Error metrics

Three complementary quantitative metrics were used to test the accuracy of the automatic segmentation protocol. The relative volume error $(RVE_{j,k})$ was calculated following Equation 4.11 for each muscle (j) in each subject (k).

$$RVE_{j,k} = 100 \times \frac{V_{P_{j,k}} - V_{G_{j,k}}}{V_{G_{j,k}}}$$

Equation 4.11

Where $V_{P_{j,k}}$ and $V_{G_{j,k}}$ are the volumes of the automatic (P) muscle segmentation and ground truth segmentations (G), respectively. The subscripted terms identify the muscle (j), and subject (k) for which the RVE was calculated.

The Dice similarity coefficient (DSC) [131] was used to assess the accuracy of segmentation considering both volume and geometry, through comparison with the ground truth segmentation. The DSC varies between 0 and 1, with a value of 1 signifying that the proposed automatic segmentation and ground truth are identical. The DSC was calculated (Equation 4.12) for each muscle (j) in each subject (k), where $P_{j,k}$ and $G_{j,k}$ represent the automatic and ground truth segmentations, respectively.

$$DSC_{j,k} = \frac{2\left(P_{j,k} \cap G_{j,k}\right)}{\left|P_{j,k}\right| + \left|G_{j,k}\right|}$$

Equation 4.12

Three examples are presented in Figure 4.5, where the automatic (P) and ground truth (G) segmentations are modelled as two 2D circles of equal radius and the distance between the centres of the circles is defined as x. As the centres of the circles are shifted further apart the DSC is reduced, highlighting that the measure is able to account for spatial misplacement of the automatic segmentations in reference to the ground truth segmentations.

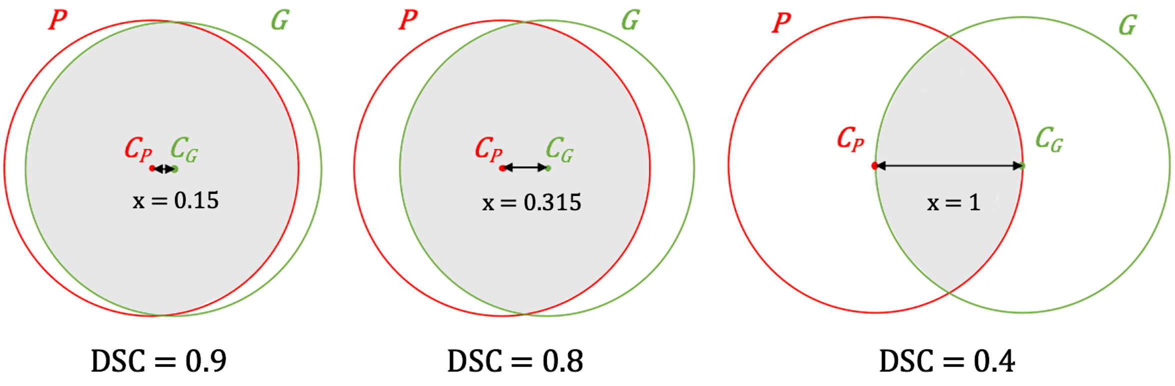


Figure 4.5: Visual representations of the Dice similarity coefficient. Two circles of radius equal to 1 are shown, representing the automatic (P) and ground truth (G) segmentations, with centres C_P and C_G, respectively. The centres are shown at three different distances apart, affecting the intersection of the circles. The DSCs are shown for the three cases.

Finally, the Hausdorff distance (HD) [132] between the proposed and ground truth muscle segmentations was calculated for each muscle in each subject, following Equation 4.13, where $p_{j,k}$ is a point within the muscle boundary $P_{j,k}$, $g_{j,k}$ is a point within the muscle boundary of $G_{j,k}$ and d is the magnitude of the greatest distance between any point $p_{j,k}$ or $g_{j,k}$ and its nearest neighbouring point in $G_{j,k}$ or $P_{j,k}$, respectively. For each subject the HD was calculated as the maximum among the minimum distances between the automatic and reference segmentations in each point.

$$HD(P_{j,k}, G_{j,k}) = max\left\{\left(d(p_{j,k}, G_{j,k})\right), \left(d(g_{j,k}, P_{j,k})\right)\right\}$$

Equation 4.13

A Hausdorff distance of zero suggests that two objects are geometrically identical, and conversely, a large Hausdorff distance implies geometrical disparity, at least in a portion of the object. The Hausdorff distance is a purely geometric measure of segmentation accuracy, considering both the shape of the two objects compared, and the difference in spatial location of the two objects. Figure 4.6 presents the HD for two arbitrary 2D surfaces modelling a proposed automatic segmentation (P) and ground truth segmentation (G).

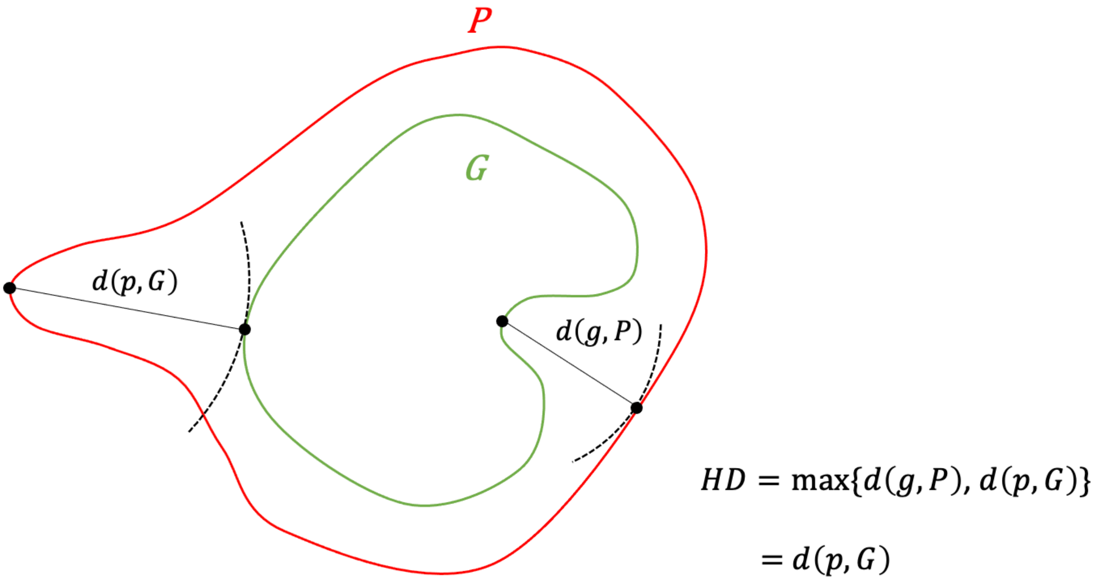


Figure 4.6: Visual representations of the Hausdorff distance. The Hausdorff distance is found for two arbitrary surfaces, P and G, as the greatest distance that can be found between a point in P and a point in G.

4.2.3. Segmentation tasks

Three segmentation tasks were outlined in this section as described in Objectives 3-5 (see Section 1.1), respectively.

4.2.3.1. Objective 3: Optimisation of registration parameters

The first task was to optimise the user inputted parameters of the registration for the segmentation of muscles from MR images through a sensitivity analysis. Two parameters were analysed, the NS and the smoothing coefficient. To perform this sensitivity analysis, imaging sequences containing the right limb were registered with sequences of the mirrored left limb (mirroring with respect to the sagittal axis was applied), for one subject chosen at random. A visual representation of this process is presented in Figure 4.7. The muscles within the right limb were automatically segmented and compared to the ground truth segmentations.

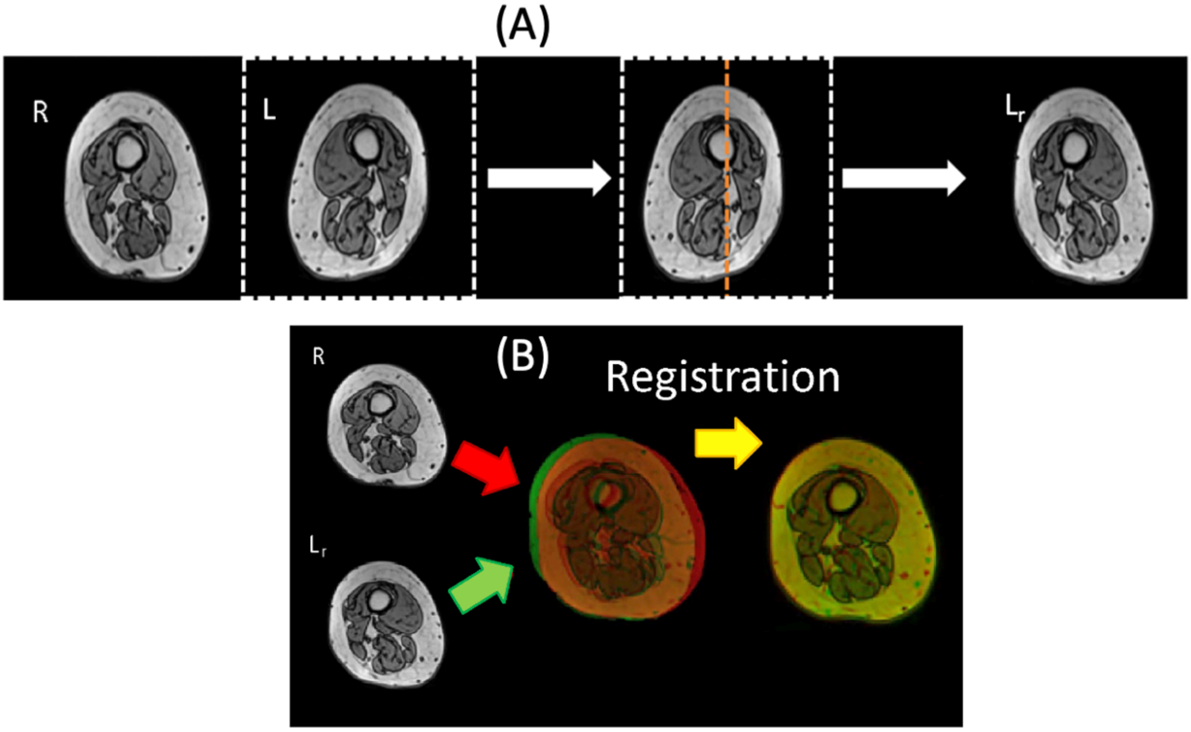


Figure 4.7: A: Separation of the anatomical right (R) and left (L) limb followed by a reflection of the left limb in the sagittal axis (L_r), shown for one example 2D slice of image. B: Registration of MR imaging data. The reflected left image (L_r) is inputted into ShIRT as the moving image (shown in green) and registered to the right image (R), inputted as the fixed image (shown in red). The registered image (right most image) shows these images after registration.

The minimum and maximum NS tested were 5mm and 40mm, with intervals of 5mm. If a NS of less than 5mm was used, the time to perform the registration would be in excess of 1.5 hours, as the computation time increases exponentially with a decrease in NS. The maximum NS tested (40mm) was selected as this was less than 33% of the diameter of the thigh within the subject considered. If a value larger than 40mm was used for the NS, the number of nodes within the body would be too few (i.e. two nodes with thigh diameter of less than 120mm) to allow the internal anatomy to be captured through registration (few degrees of freedom in the transformation).

The second user inputted registration parameter, the smoothing coefficient (λ), is optimised automatically within ShIRT [130]. This optimized value could be miscomputed and so a further sensitivity analysis was required to verify the optimal value. The automatically found smoothing coefficient (135) was computed through an initial registration of (mirrored) left and right limb of the subject. As the smoothing coefficient is not related to any physical parameter, the minimum and maximum values used within the sensitivity analysis were two orders of magnitude greater and smaller than that found automatically within ShIRT, with intervals of one order of magnitude: λ = [1.35, 13.5, 135, 1350, 13500].

The accuracy of each set of segmentations gathered by changing both registration parameters was analysed using the error metrics outlined in the previous section. This analysis was performed independently for both registration parameters, to find the value of each that provided the greatest overall segmentation accuracy. The three sets of errors found across the tested values of the NS and smoothing coefficient were independently statistically analysed. Firstly, a Kolmogorov-Smirnov test was used to test each set of segmentation errors for normality. Each set was found not to be normally distributed. Thereafter, the non-parametric Kruskal-Wallis one-way analysis of variance (ANOVA) was used to assess whether the independent variables (NS and smoothing coefficient) affected the segmentation accuracy across the three error metrics. Where there was a significant difference (p-value < 0.05) between the means of the error metrics across the tested values, a post hoc Tukey-Kramer multiple comparison test was conducted. These analyses were used to identify the optimal values for both the NS and smoothing coefficient. Those registration parameters that provided the greatest overall segmentation accuracy were used in the subsequent segmentation tasks.

4.2.3.2. Objective 4: Intra-subject registration

Registering between images of the mirrored left and right lower limb (presented in Figure 4.7) of an individual subject limits the complexity of the registration process, as there is some degree of anatomical variability between each side, but this is typically far less than the variability between subjects [66]. Therefore, the second task consisted of an intra-subject registration using the optimised registration parameters, wherein the contralateral limb of five subjects were segmented. The accuracy of each set of segmentations was analysed using the error metrics outlined in the previous section, comparing the automatically generated segmentations with their respective ground truth, manual segmentations. Kolmogorov-Smirnov test was used to assess the sets of error metrics for normality, for each of the five sets of segmentations. All sets of error metrics were deemed to be not normally distributed, meaning that a non-parametric test was required. A Kruskal-Wallis ANOVA test was used to analyse the effect of the independent variable: the subject that was segmented.

4.2.3.3. Objective 5: Inter-subject registration

The third task was to register images containing the right limb of one subject to the right limb of another one (inter-subject registration). This process simulates the automatic segmentation of the individual muscles for a new subject, using previously segmented subjects as a reference. Each subject was used as both the target and the reference within the registration algorithm, generating four segmentations for each subject. The accuracy of each set of segmentations was analysed using the error metrics outlined in the following section. The results of the inter-subject registration task were compared to that of the intra-subject registration task to assess the merits and limitations of both approaches.

To evaluate whether there was a significant difference between the results of the intra-subject and inter-subject analyses, statistical tests were conducted. The sets of segmentations generated for each target subject with the inter-subject method was treated independently and individually compared with the corresponding segmentations generated with the intra-subject analysis. To do so, first a Kolmogorov-Smirnov test was used to check the distributions of the three error metrics for the inter-subject analyses for normality. Thereafter, a Wilcoxon signed rank test was used between the errors found for each subject in the intra-subject analysis and each of the corresponding 4 sets of errors found in the inter-subject analysis.

4.3. Results

4.3.1. Objective 1: Automatic pre-processing of images

The eleven subjects for whom the imaging data was collected as part of a prior study [66] were pre-processed using the method prescribed to fulfil Objective 1. The image sequences of the hips, thigh, knees, and calf were combined, forming one continuous image containing the entire lower limb. Figure 4.8 below shows the pre-processed images of the five subjects that were selected for the registration and segmentation pipeline. The pre-processing required 109 ± 11.2 seconds (mean ± standard deviation) of computation time (Intel® Core™ i7-7700 CPU @ 3.60 GHz), depending on the size of the images being concatenated.

This method has since been used to pre-process imaging data collected for 4 different studies (MultiSim [66] (n=11), MRI-US (n=11), Obesity study (n=26) (Section 3.3), PORTRAIT (n=7)), with complete success.

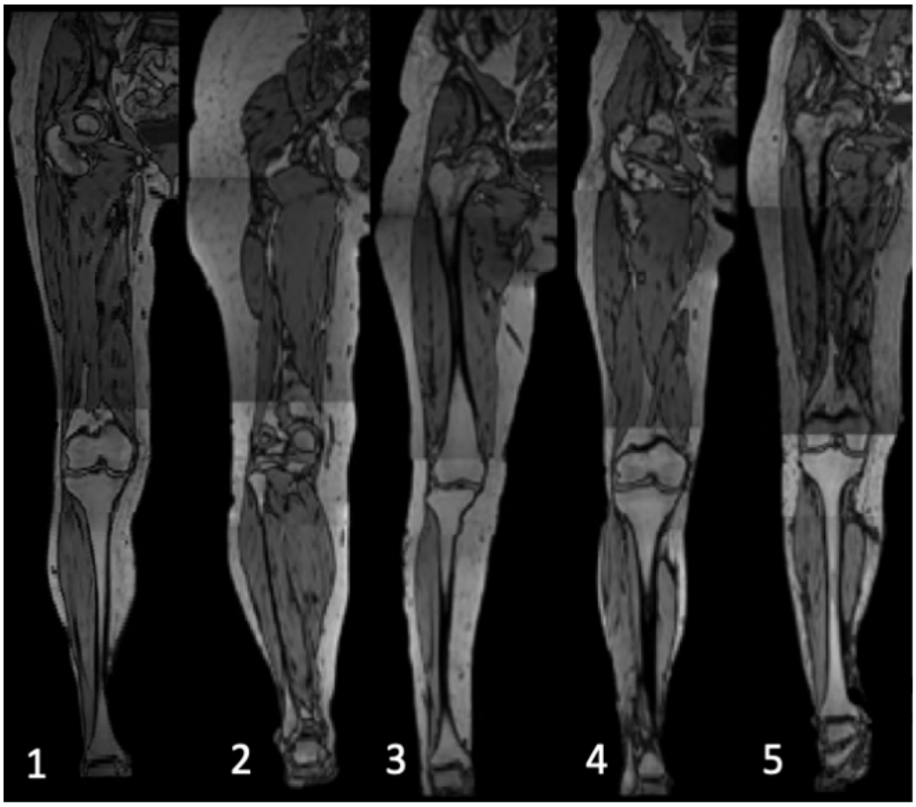


Figure 4.8: Pre-processed imaging data for the five subjects selected for segmentation using the registration and segmentation pipeline.

4.3.2. Objective 2: Registration and segmentation pipeline

All three of the outlined segmentation tasks (Objectives 3-5) were completed using the registration and segmentation pipeline. Here the computational time is reported. In the first segmentation task, 8 values of NSs were tested, NS = (5, 10, 15, ..., 40). The time requirement for the registration and segmentation pipeline to be completed for each of the eight tested values of NS are presented in Table 4.1, with the time required decreasing as the NS was increased. The smoothing coefficient had no effect on the time required to perform the registration and segmentation pipeline.

Nodal spacing (mm)	5	10	15	20	25	30	35	40
Time (min)	72	46	35	28	26	23	22	20

Table 4.1: Time requirements the registration and segmentation pipeline for each of the tested values of NS.

In the other two segmentation tasks, the registration and segmentation pipeline required between 72 and 96 minutes, to register the smallest and largest images respectively. These analyses were performed on an Intel® Core™ i7-7700 CPU @ 3.60 GHz.

4.3.3. Objective 3: Optimisation of registration parameters

4.3.3.1. Nodal spacing

The optimal values for the user inputted registration parameters were found through a sensitivity analysis. Eight values were tested for the NS (5, 10, ..., 40) within the registration algorithm and the segmentation accuracy resulting from each registration were found using three error metrics: RVE, DSC and HD. The impact of the NS on the nodal grid is illustrated in Figure 4.9. The density of the nodes reduces drastically as the NS was increased. The registered images from the lower NSs were more accurate (yellow represents overlap between the fixed and moved images reported in red and green, respectively) (Figure 4.9).

The results for the different metrics are reported in Figure 4.10. The mean RVE across all NSs were close to 0%, but NSs of 5mm and 20mm were associated with the lowest mean errors (-0.7% and 0.5%, respectively). The distributions of RVEs were statistically tested, comparing the lowest mean (NS = 20mm) against all others. No statistically significant differences were found between the means of the RVE across the 8 tested values for NS within the Kruskal-Wallis ANOVA test (p = 0.982), therefore no post hoc analysis was performed. The interquartile range and full range were the lowest with a NS of 5mm, which was considered the optimal value for RVE.

Similarly, the DSC computed for the segmentation across all NSs considered, also suggested that a NS of 5mm was optimal. The mean DSC was the greatest when using a NS of 5mm and steadily decreased as the NS increased. The mean DSCs across the 8 tested values of NS were found to be significantly different through the Kruskal-Wallis ANOVA test (p = 9.91$\times$ 10^{-12}). The post hoc Tukey-Kramer multiple comparisons test highlighted that the mean DSC found with a NS = 5mm was significantly greater than all others values, with the exception of NS = 10mm. Additionally, the post hoc test showed that the DSC found with NS = 10mm was significantly greater than those found with NS $\geq$ 25mm. All mean DSCs found from the other values of NS were not significantly different. The three muscles segmented with the lowest accuracy using NS $>$ 10mm were the biceps femoris caput breve, semitendinosus, and tensor fascia latae. However, these three muscles did not have the lowest DSC for NS $\leq$ 10 mm.

The HD found in all segmented muscles was relatively consistent across all NSs, with little to no difference between the segmentation results. This was confirmed by the Kruskal-Wallis ANOVA test, which found that the mean values of HD across the 8 tested NS values were not significantly different from one another (p=0.999). The outlier visible across all 8 NS values was the same muscle: the vastus medialis. Nevertheless, the upper quartile for NS equal to 5 mm was lower than for the other NSs.

Considering the three error metrics, it was concluded that the NS of 5mm was the optimal value of NS (within acceptable computational time) to be used in future registration tasks.

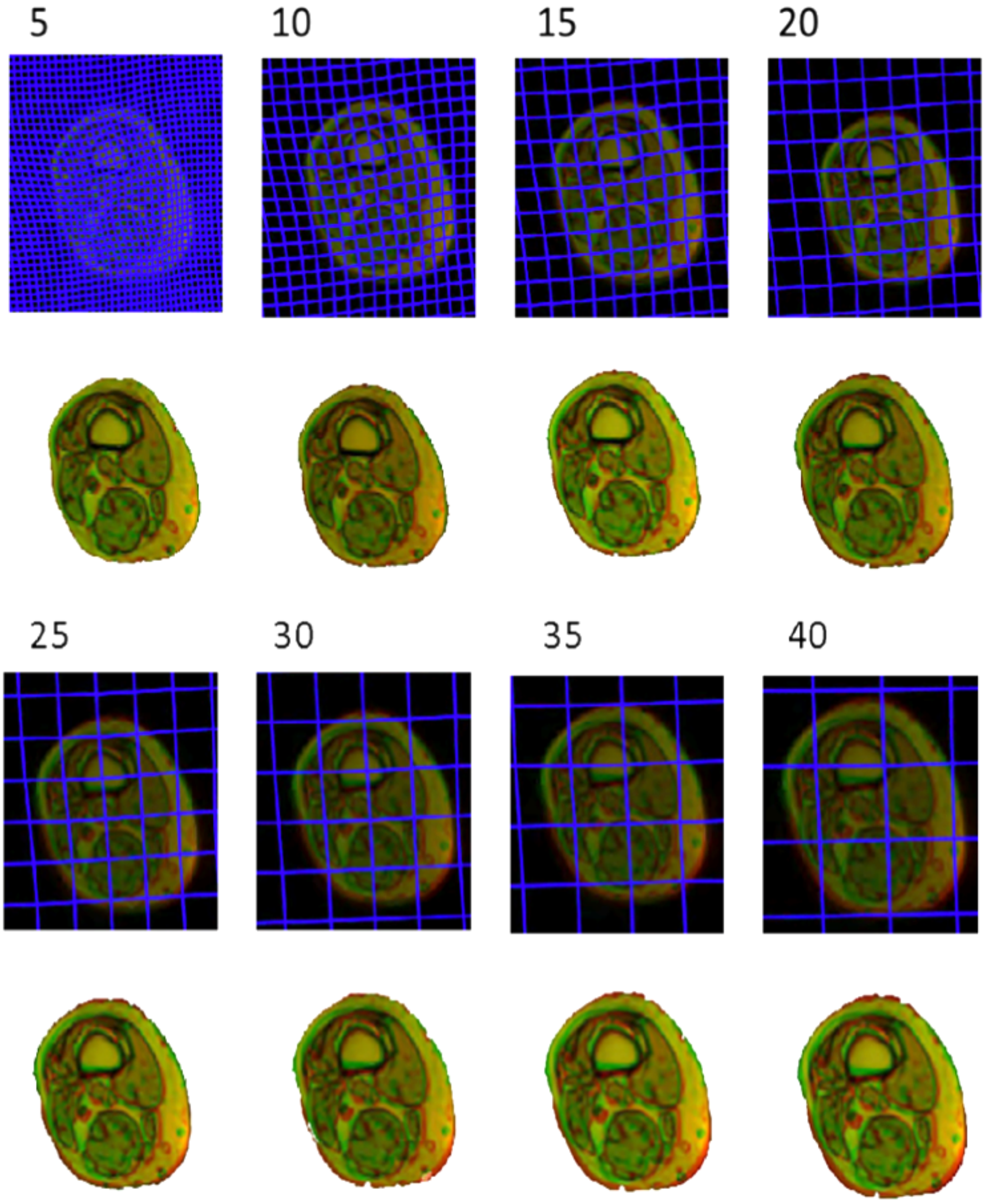


Figure 4.9: Qualitative comparison of the registration produced using the eight different values (voxels, with voxel size of 1mm) of NS chosen in the sensitivity analysis (5, 10, 15, ..., 40). The first and third rows of images show the registered nodal grid overlayed with the registered image and the second and fourth rows show the registered (green) and target (red) overlayed. Areas of yellow showcase a well registered section of the image. Conversely, areas of intense red or green showcase a poor-quality registration.

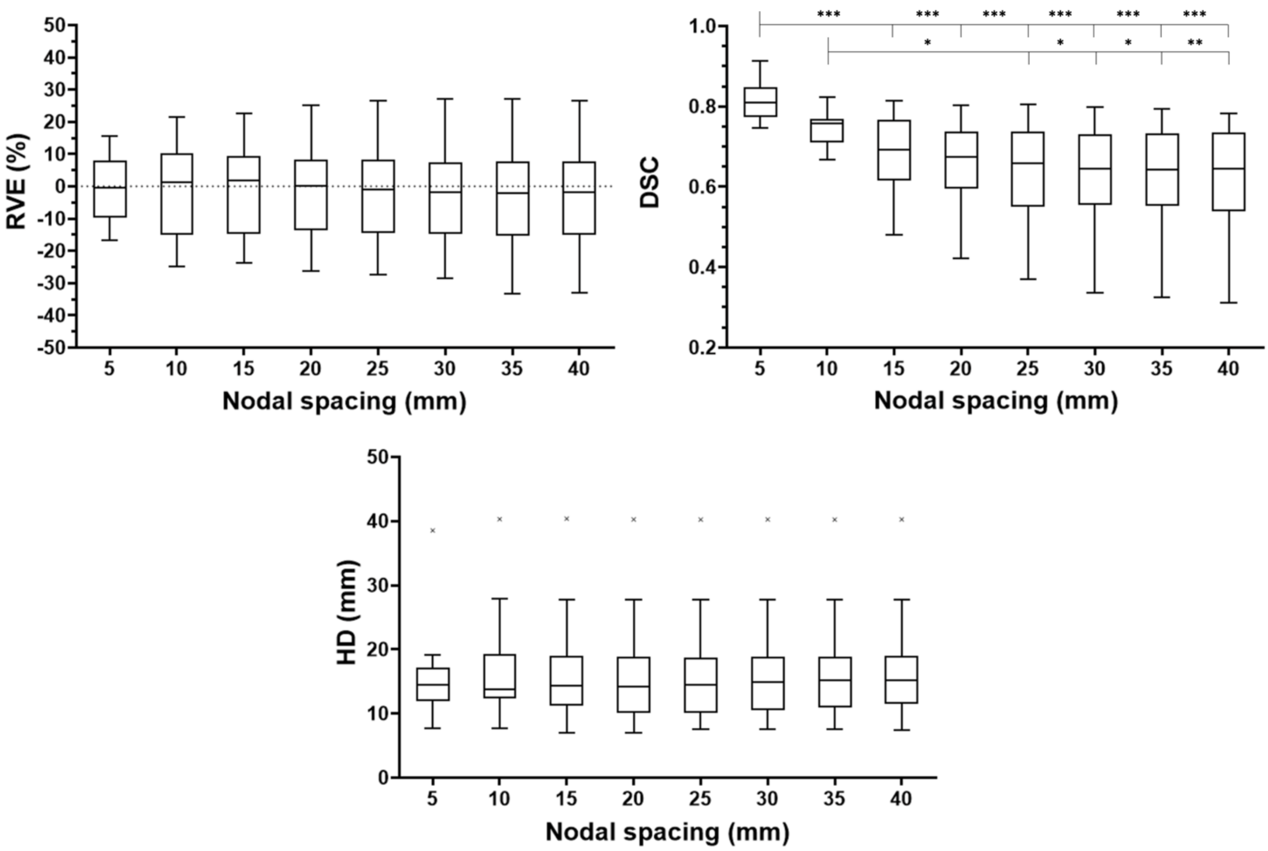


Figure 4.10: Boxplots present the RVE, DSC, and HD found for the proposed segmentations of the 23 muscles across each of the 8 values of NS analysed. A Kruskal-Wallis ANOVA test was used for each error metric (RVE: p = 0.982, DSC: p = 9.91× 10^{-12}, and HD: p = 0.999). The Tukey-Kramer multiple comparison test was used as a post hoc analysis for the DSC. Connections between bar charts (starting at extended line) shows statistically significant differences between the means of each group (* means p < 0.05, ** means p < 0.01, *** means p < 0.001).

4.3.3.2. Smoothing coefficient

A sensitivity analysis of the smoothing coefficient was also performed. The results are shown in Figure 4.11, with a visual interpretation of the effect of the smoothing coefficient shown in Figure 4.12. The optimal smoothing coefficient calculated within ShIRT for this registration task was 135. The mean RVE for smoothing coefficient equal to 135 was lower than the other tested values (>0.14% lower; p = 0.116). Similarly, the mean DSC was the greatest for smoothing coefficient equal to 135 (>0.26 higher; p = 4.79× 10^{-12}). The mean DSC found for smoothing coefficient was significantly greater than those found using other values of smoothing coefficient (p < 0.001). Finally, the mean HD found with smoothing coefficient equal to 135 was lower than the others (4.9mm lower, p = 0.0485). The mean HD was significantly lower with smoothing coefficient equal to 135 than smoothing coefficient equal to 13500 (p = 0.032). All tested values of the smoothing coefficient had one consistent outlier within the HD distributions, the vastus medialis muscle. As the optimal smoothing coefficient performed the segmentation with the highest overall accuracy, the optimal smoothing coefficient calculated within ShIRT was chosen for further analyses. Though, it must be noted that the optimal value for the smoothing coefficient depends on the images being registered and does fluctuate between different registrations.

From a qualitative standpoint, Figure 4.12 shows that the smoothing coefficient significantly alters the linearity of the deformed nodal grid, with the lowest smoothing coefficient leading to a severely non-linearly deformed nodal grid, and the greatest being smooth, with an apparent overall rotation. The optimal smoothing coefficient presented a balanced mixture of the two extreme values with a deformed grid that was mostly smooth but non-linear in specific locations owe to abrupt changes in anatomical features, such as the upper left portion within the anatomical aspect of the images.

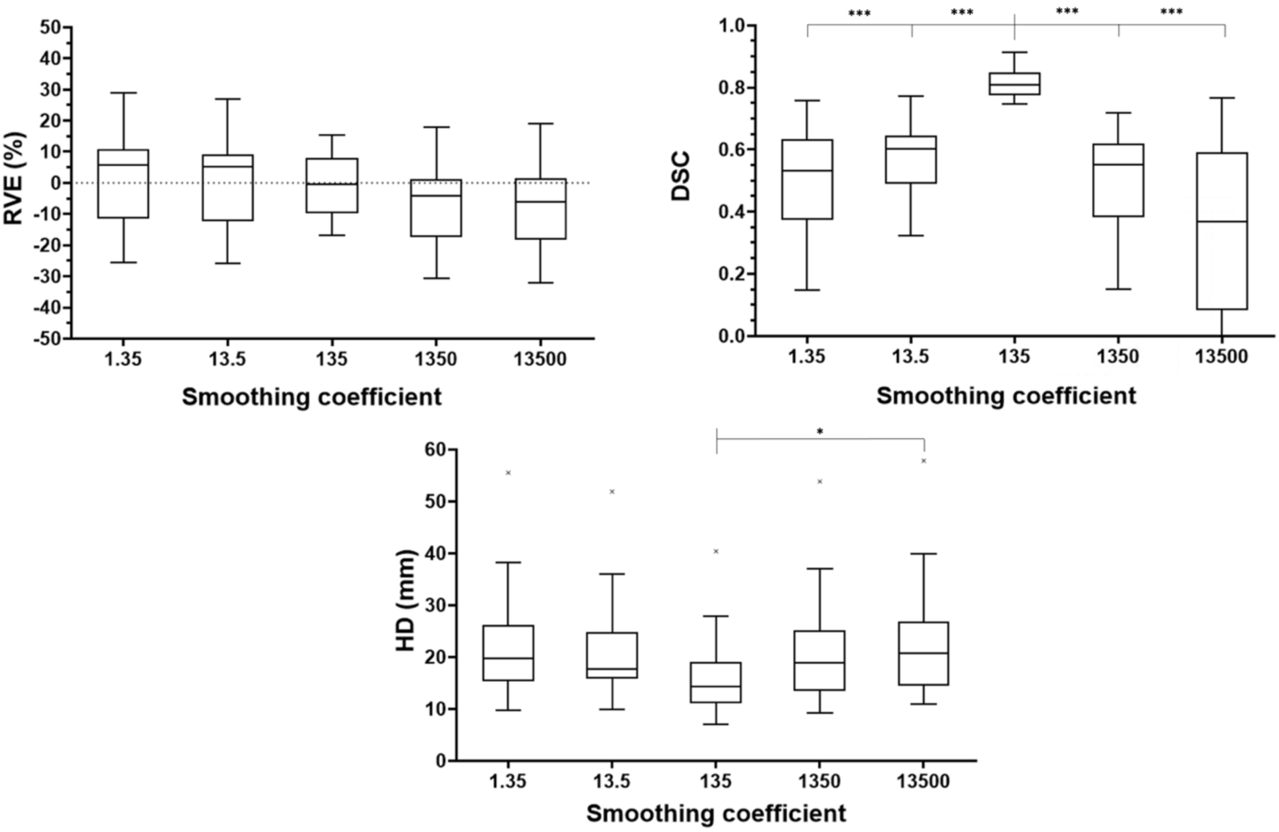


Figure 4.11: Boxplots present the RVE, DSC, and HD found for the proposed segmentations of the 23 muscles across each of the 5 values of the smoothing coefficient analysed. A Kruskal-Wallis ANOVA test was used for each error metric (RVE: $p = 0.116$, DSC: $p = 4.79\times 10^{-12}$, and HD: $p = 0.0485$). The Tukey-Kramer multiple comparison test was used as a post hoc analysis for the DSC and HD. Connections between bar charts (starting at extended line) shows statistically significant differences between the means of each group (* means $p < 0.05$, ** means $p < 0.01$, *** means $p < 0.001$).

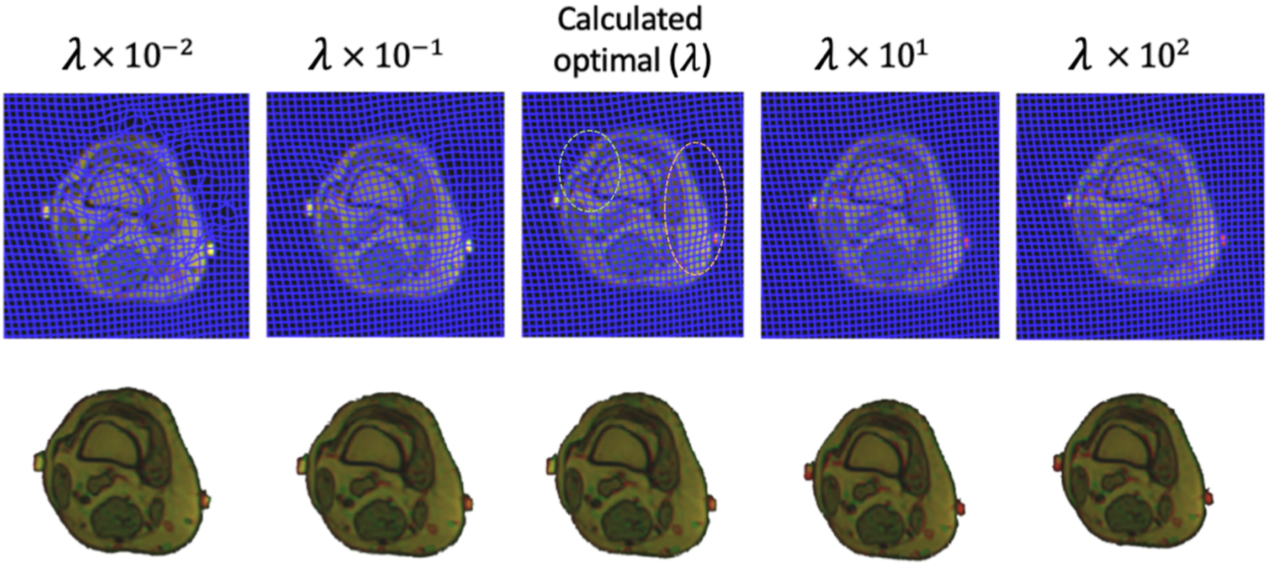


Figure 4.12: Effect of changing the smoothing coefficient on the nodal grid and resulting registration outputs. The calculated optimal smoothing coefficient (λ) in this case was 135. Areas are marked within the map found with the optimal value of λ: light green highlights an area of non-linearity, and orange highlights an area of moderate linearity. The top row of images show the registered nodal grid overlayed with the registered image and bottom row show the registered (green) and target (red) overlayed.

4.3.4. Objective 4: Intra-subject registration results

Using the optimised registration parameters (NS=5 voxels, optimal lambda), the 23 muscles considered within the right limbs of the five subjects included in the study were segmented using the left limb as the reference.

The RVE calculated across the 23 muscles within the five subjects are plotted in Figure 4.13. The mean RVE was less than ±2% across all subjects and the upper and lower quartiles were less than ±13%. The extreme values of RVE were larger, with some muscle volume being overestimated by 50%, suggesting a poor segmentation of the muscle volume in some cases. The Kruskal-Wallis ANOVA test was used on the results of the RVE and found that there was no evidence to suggest that the independent variable, the subject, influenced the RVE of the segmentations (p = 0.998).

The DSC were generally high, with the mean DSC being between 0.75 and 0.85, showing a good level of agreement between automatic and ground truth segmentations. The interquartile ranges for Subjects 1, 3 and 5 were less than 0.07, showing a consistently high DSC across the 23 muscles considered within these subjects. More than 60% of the muscle geometry was captured well within all muscles and all subjects considered (DSC > 0.6). Through a second Kruskal-Wallis ANOVA test, like the RVE, there was no evidence to suggest that changing the subject influenced the DSC of the segmentations (p = 0.528).

The HD was consistent across all muscles within each subject, with the mean HD being below 20 mm for each subject. On average, each muscle was segmented with a maximum distance between the surface of the automatic and ground truth segmentations at lower than 2 cm. However, the maximum HD within all subjects was approximately 40 mm, for muscles of greatest length, such as the sartorius. As with the RVE and DSC, the Kruskal-Wallis ANOVA test suggested that there was no evidence that the mean HDs was different amongst the five subjects (p = 0.154).

Figure 4.13 also shows an example of the ground truth and automatic segmentations for Subject 1, both in 3D and 2D image slices. Overall, the 3D representations are visually similar, with all muscles being located correctly and with each muscle being of comparable size. The results were also in line in the 2D cross sections, wherein the ground truth and automatic muscle contours had a high level of agreement.

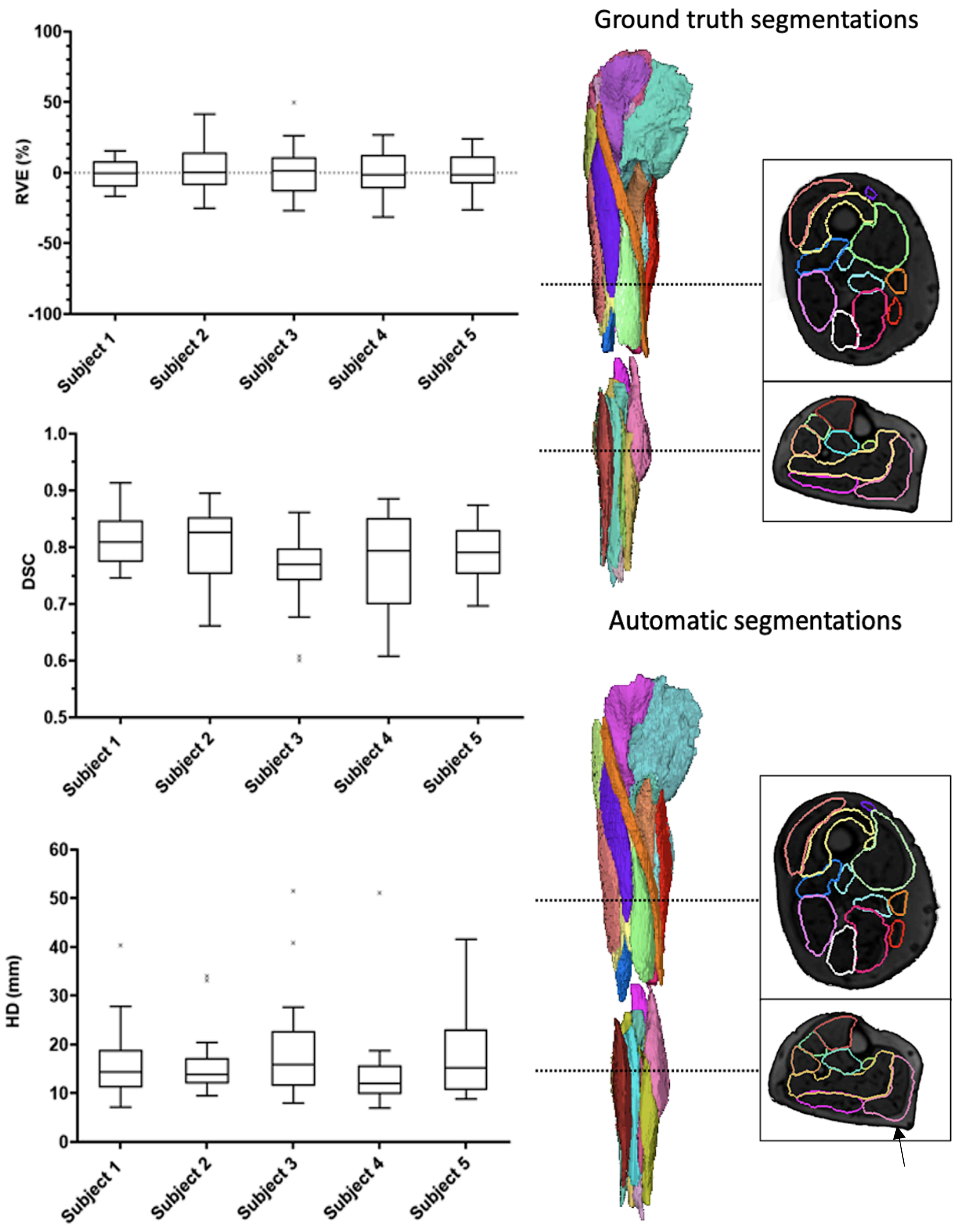


Figure 4.13: Intra-subject registration results. The left-hand side shows the RVE, DSC, and HD calculated across the 23 muscles considered within each of the 5 subjects. The right-hand side shows a visual illustration of the ground truth and automatic segmentation in both 2D and 3D for one subject. Black arrow points to the medial head of the gastrocnemius, referred to in Section 4.4.

4.3.5. Objective 5: Inter-subject registration results

Finally, the segmentation pipeline was used to segment the 23 muscles within the lower limbs of each of the five subjects, using each of the other subjects as references (20 comparisons). The resulting muscle segmentations from all pairs of inputs were compared to the associated ground truth segmentations using the three outlined error metrics to measure the accuracy of the segmentation. The resulting RVE, DSC and HD are presented in Figure 4.14, where the five groups of boxplots within each graph represent the four segmentations of each subject produced using the other 4 subjects as the references. Additionally, the results found using the contralateral limb for each subject are also reported, allowing comparison between intra-subject and inter-subject registrations.

The RVEs for the inter-subject analysis are typically much higher than the intra-subject approach, for all target subjects. The mean RVE for the inter-subject approach were found to be in the range of $\pm$3-75%. In the worst performing cases (Subject 1), the volume of the muscles was captured on average with an accuracy around 50%.

Overall, the DSC shows that across all combinations of subjects, the inter-subject approach was not able to capture the geometry of the muscles with a good level of accuracy. Across all combinations of subjects inputted into the segmentation pipeline, the DSC was quite low, with mean between 0.3 and 0.7, and with large interquartile range. In some combinations of subjects (e.g. Subject 5 as the target and Subject 3 or Subject 4 as the references), the DSC for some muscles was even close to 0, meaning that there was no overlapping area between the reference and automatic segmentations.

The intra-subject approach far outperformed all inter-subject approaches, considering all three error metrics. The Wilcoxon signed rank test found that 14/20 of the mean RVE within the inter-subject analysis were significantly ($p < 0.05$, greatest = 0.88, least = 2.7×10^{-5}) larger when compared with their respective intra-subject analyses (exceptional cases were target Subject 2 with reference Subjects 1, 4, and 5, target Subject 4 with reference Subject 1, and target Subject 5 with reference Subjects 2, and 4). The DSC found in the intra-subject analyses were significantly greater ($p < 0.05$, greatest = 5.1×10^{-3}, least = 2.1×10^{-5}) than their corresponding inter-subject analyses. Finally, 19/20 of the HD distributions found in the inter-subject analysis were significantly greater ($p < 0.05$, greatest = 0.068, least = 4.0×10^{-5}) than their corresponding intra-subject analyses (exception was target Subject 1 with Subject 2 as the reference).

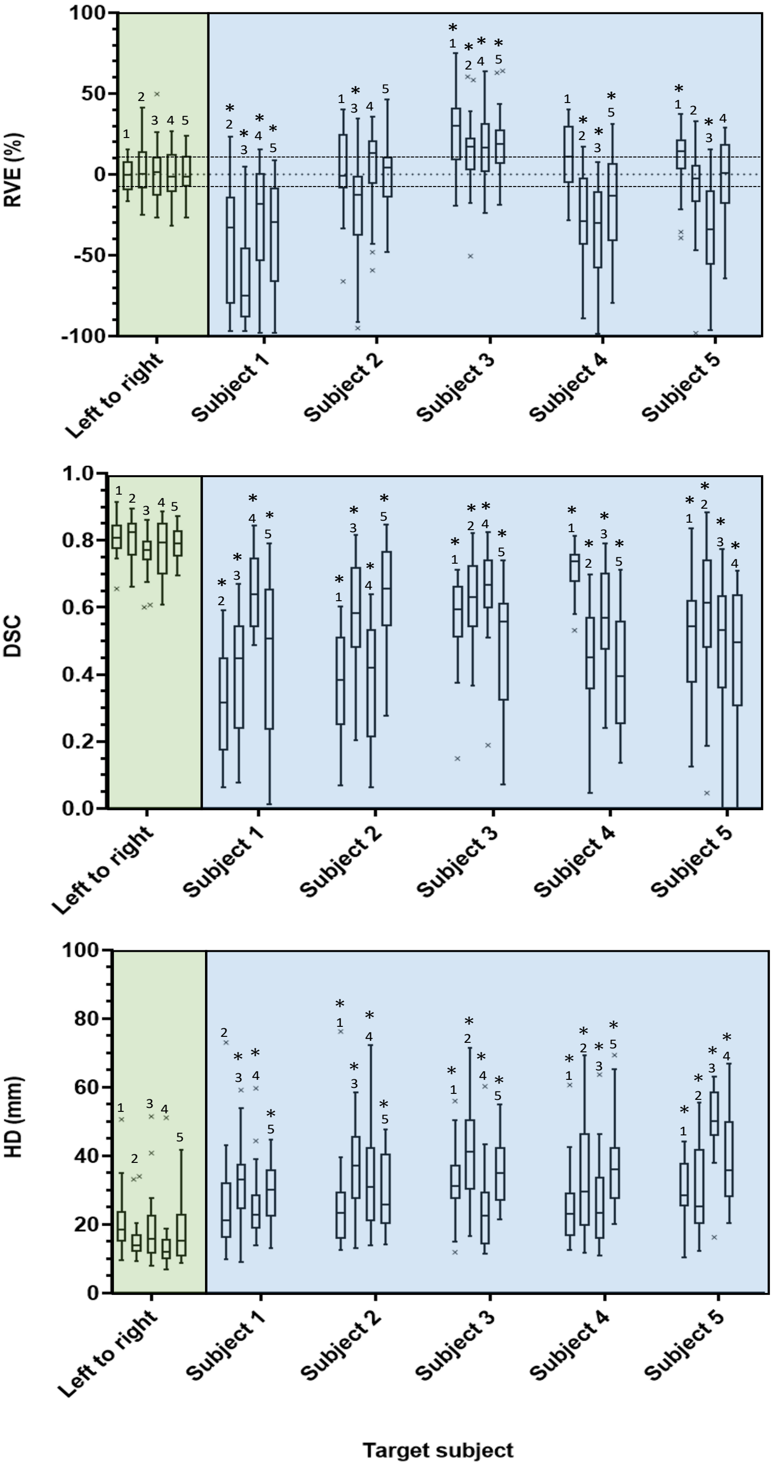


Figure 4.14: Intra (left to right, green background) and inter-subject (blue background) registration results, in terms of RVE, DSC, and HD. Each boxplot shows the error found across the 23 muscles considered. The numbers above each plot indicates the reference subject used to generate the proposed segmentations. The black dashed lines (±10%) within the RVE plot shows the acceptable level of operator dependency for that error metric. A Wilcoxon signed rank test was used between the intra-subject results for subject i and the associated 4 inter-subject results for target subject i. * above the inter-subject result shows that there was a statistically significant ($p < 0.05$) difference between the mean of the bar chart underneath, and the intra-subject result.

A visual illustration of the registration and associated segmentation results are presented in Figure 4.15, which reports the images inputted into ShIRT, the registered images, the ground truth segmentation, the automatic segmentations for the intra-subject registration (Subject 1), and the inter-subject combinations resulting in the lowest, median, and greatest average DSCs. The segmentations from the combination of subjects resulting in the worst average DSC have little to no similarity to the ground truth segmentations. This was expected as the registered images do not appear well aligned, particularly within the muscle anatomy, where there are large dissimilarities between the registered and the fixed image. The combination of subjects resulting in the median and greatest average DSC generally have a good level of agreement between the automatic and ground truth segmentations, with some muscles being captured well and others being misplaced. Again, these results are reflected within the registered images, particularly in the best performing combination of subjects, as the vastus lateralis (shown in peach, arrow in Figure 4.15) is not well registered within the registered images and is therefore not well segmented.

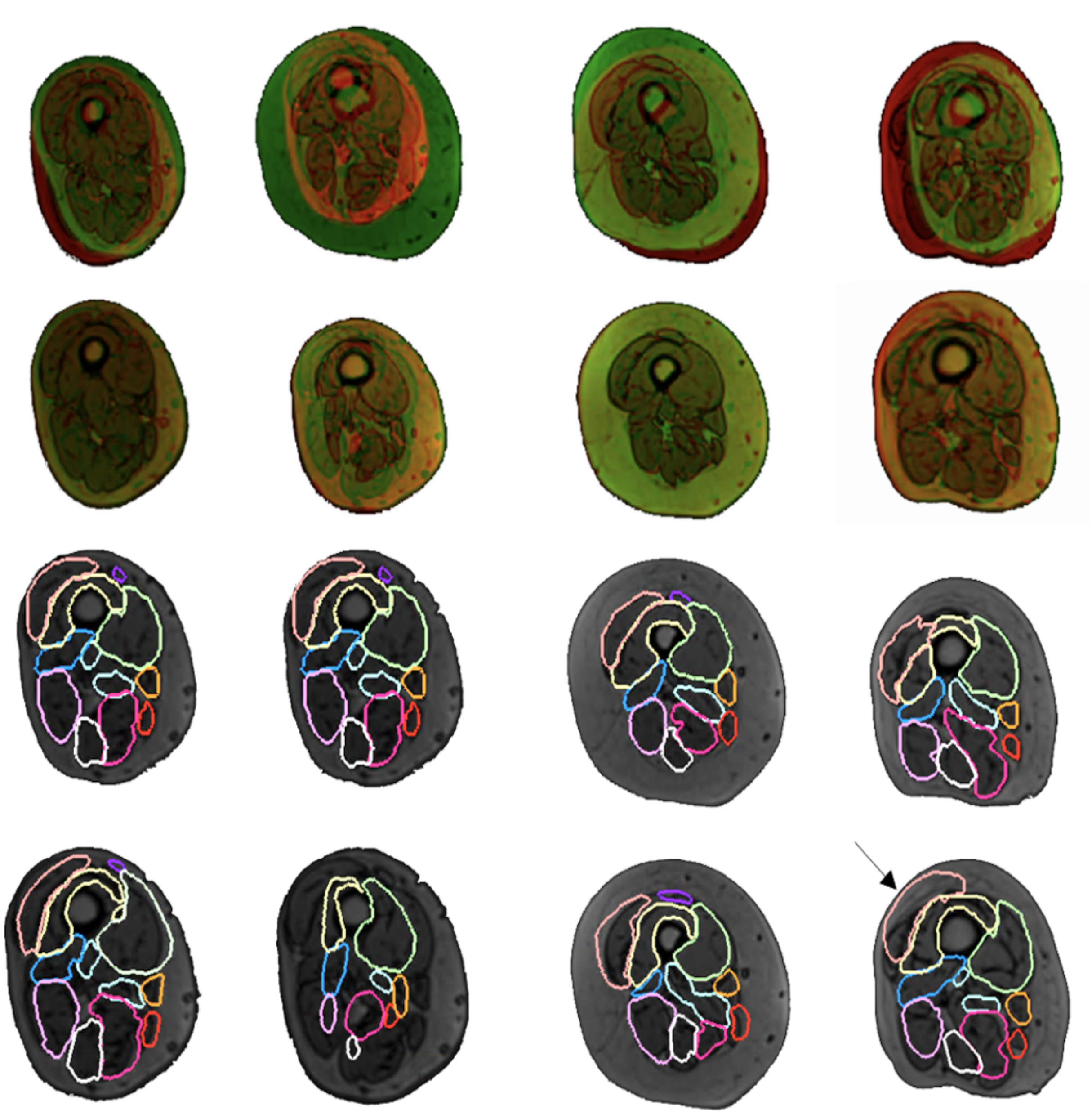


Figure 4.15: Visualisation of the registration inputs (top row), outputs (2nd row), ground truth segmentations (3rd row) and resultant segmentations (4th row), shown for the intra-subject registration (left most column), worst performing combination of subjects (target: subject 1, reference: subject 2, 2nd column), median performing combination (target: subject 3, reference: subject 5, 3rd column), and best performing combination (target: subject 5, reference: subject 1, 4th column). The black arrow highlights the vastus lateralis, mentioned in Section 4.3.5.

4.4. Discussion

The chapter aimed at building and testing an automatic muscle segmentation pipeline, using deformable image registration. Complete 3D geometries of 23 muscles were segmented from within the lower limbs of 5 subjects, firstly in an intra-subject analysis, followed by an inter-subject analysis with varying success.

4.4.1. Objective 1: Automatic pre-processing of images

The MR images were acquired in five sections: the hips, thighs, knees, calves, and feet. The images from each of these sections were first combined, forming one continuous 3D image containing the complete lower limb. This operation was performed well for the original eleven subject in the MultiSim database [66] and has since been applied to others, such as the MRI-US, PORTRAIT and STH21022 databases. Overall, the combination method works as intended and enables full 3D anatomical data to be combined. This is particularly important when examining the muscles, as artefacts often spread across multiple acquisition sections [60].

This operation was previously performed in a semi-automatic fashion, with the user being required to highlight common points within two sections, allowing the frames of reference between them to be aligned. That approach requires user input and is subject to error if the same anatomical landmark is not clearly visible within the two scans. Additionally, this semi-automatic method takes some time as there are four or five sections to be aligned, depending on if the user includes the imaging sequence containing the feet. The proposed method removed the need for the manual operation and retains the spatial location of the anatomy in a mathematically rigorous manner. Additionally, the manual process required around 10-20 minutes to perform. Using the pre-processing script reduced the processing time to between 1 and 2 minutes, and therefore, led to a significant time reduction depending on image size.

4.4.2. Objective 2: Registration and segmentation pipeline

The registration and segmentation pipeline was built and used to complete the three segmentation tasks: optimisation of registration parameters (Objective 3), intra-subject segmentation (Objective 4), and inter-subject segmentation (Objective 5). The time required to perform the registration was dependent on the NS used within the registration. Using the optimal NS of 5mm, the time required to perform the registration was between 65 and 90 minutes across the intra-subject and inter-subject registration tasks. The segmentation aspect of the pipeline required between 5 and 8 minutes, in order to apply the displacement vector fields due to the large number of points that made up each segmentation (in the order of 100000s of points). Nevertheless, this was a vast improvement compared to the time required using a manual process, which was estimated at around 10 hours [66, 87]. Therefore, the automatic segmentation method fulfilled the aim of reducing the vast time required to perform manual muscle segmentation and removing the operator dependency issues [28].

4.4.3. Objective 3: Optimisation of registration parameters

A sensitivity analysis was performed on the two registration parameters: the NS and the smoothing coefficient. The optimal value of both parameters was extracted from the analysis. The optimal NS selected was 5mm. This value resulted in segmentation with the greatest accuracy in terms of DSC, and comparable to others in terms of RVE and HD. An optimal value of the smoothing coefficient was automatically calculated within ShIRT, and this value was confirmed to be the optimal value, producing significantly greater segmentation accuracy in terms of both DSC and HD. There is a potential that the optimal values found within the segmentation tasks were optimal in some areas of the body, but not in others. Calculating an optimal value for each body segment could have merit but the outputted map would be far more difficult to interpret in this segmentation pipeline and was therefore not explored.

4.4.4. Objective 4: Intra-subject registration results

The intra-subject analysis sought to segment one limb of 5 subjects using deformable image registration, using the contralateral limb as the reference. The volume of the muscles was captured well through registration, with the average RVE reported as less than ±2%. However, the absolute values of the upper and lower quartiles of the RVE were consistently below 18%, which exceeded the ±10% that would satisfy the inclusion criteria for repeatable manual segmentation [66]. Therefore, considering this metric, some but not all muscles could be captured with an acceptable level of volume error.

The average DSC across the 5 subjects was 0.81, which is comparable to the longitudinal studies presented by Le Troter et al. (2016) [94] and Fontana et al. (2020) [93] (summarised in Section 3.4.4), who achieved DSCs of 0.9 and 0.85 respectively in the segmentation of muscle from MR images using deformable image registration in longitudinal studies. These two studies segmented 4 muscles from young healthy individuals (Le Troter et al., two cohorts: A. n = 25, age = 22 ± 1 years, B. n = 7, age = 32 ± 7, and Fontana et al.: n = 7, age = 15 ± 1 years). As outlined in Section 3.4.4, the results found using the left to right segmentation approach are close to segmenting the muscles on average with an acceptable level of accuracy, but the spread of the errors would need to be reduced significantly

There are several reasons that could explain this difference. Firstly, in this study presented in this chapter, 23 muscles were segmented, which is much larger than the 4 muscles segmented in both previous studies. Secondly, in my study the images were not captured with the optimal resolutions for muscle segmentation. The long bones were captured with lower resolution (pixel size of 1.1mm^2) than the joints. The motivation for this was that images were not captured for the purposes of muscle segmentation, but for accurate capture of the joint geometry [88, 133]. The long portions of the muscle geometries (those within the thigh and calf) were therefore not captured as accurately as they could have been, likely due to the lack of contrast at the boundaries between the muscles. Comparing this with the study by Le Troter et al. (2016) [94], the images were acquired at higher resolution (pixel size of 0.5mm^2, compared with >1mm^2 used in this study), which could be the reason that the muscles were clearer within the scans, and therefore segmented with higher accuracy using image registration. Thirdly, the variability in muscle geometry and volume is significant between left and right limbs [66], where this same variability would be less apparent between the same limb if observed at two different time points, such as the approaches used in these longitudinal studies [93, 94]. Lastly, the muscle architecture visualised within MR images is more homogenous and better defined in younger individuals than in older individuals, causing fewer inconsistent features within the images [96], which would aid the registration and segmentation accuracy.

To contextualise the deformable registration for the intra-subject muscle segmentation, we can consider two applications: 1) assessment of structural muscle characteristics; 2) for use within musculoskeletal computational modelling (i.e. multi-body dynamics models).

Considering the first application, the automatic segmentations would grant a good approximation in the assessment of structural muscle characteristics, such as the muscle volume and length (see Appendix 1). The intra-subject registration would be moderately capable of capturing the muscle length, which is also used for dynamic musculoskeletal modelling [13, 126, 134]. The average HD for this analysis across the 5 subjects was 15 mm. The HD found in this aspect of the study suggested that the muscle length would be captured with an average error of <30 mm as the attachment points at the upper or lower boundary of the muscle would each be associated with a maximum error of 15 mm. However, for further investigation of muscle characteristics such as assessing the level of fat infiltration of individual muscles, one would require the automatic and reference segmentations to be in better agreement, with a higher DSC. The reason for this is that any attempt to automatically isolate the intra-muscular fat (as presented in Section 3.3.3) from a segmentation that includes some fat tissue surrounding the muscle (as can be seen in Figure 4.15), would be skewed by the outer-muscular fat. For example, in case of the medial head of the gastrocnemius (Figure 4.13, highlighted muscle), where a portion of the fat surrounding the muscle was included in the segmentation.

For the second application, the segmentations presented in this study could be useful to reduce the time needed to acquire personalised muscle information for musculoskeletal modelling [88, 134]. Dynamic musculoskeletal models are robust to perturbations in the ranges established in this work for calculating the volume and the attachment points locations for intra-subject applications [126, 135]. In fact, this tool could be used to half the time required to perform muscle segmentation in this context, as only one limb would be required to be fully segmented to attain a segmentation of both limbs.

The limitations in the accuracy of the algorithm for intra-subject muscle segmentation could be due to different reasons. Firstly, the operator variability attributed to the ground truth, manual segmentations could have been exacerbated within the registration. The registration maps each pixel within the reference image to a pixel within the fixed image, independently of the segmentations aimed to be extracted. If the reference manual segmentations used to generate the automatic segmentations were incorrect (due to operator repeatability issues), the area that a segmentation surface would be mapped to would be misplaced. Therefore, the registration could be perfect, but when applied to a surface that is misplaced, the automatic segmentation would also be misplaced. Errors incurred due to misplaced reference segmentations would have a vastly reduced effect as the only muscles included in the segmentation study were those that were segmented with an acceptable level of operator repeatability. A second source of error could be due to the low contrast between muscles. The images used in this work were not optimised to enhance the contrast of the muscle-muscle boundaries, and so this could also limit the registration and resulting segmentation accuracy.

4.4.5. Objective 5: Inter-subject analysis

The RVE found within the inter-subject registration were large, with many average values being much greater than the ±10% that would be an acceptable level from two repeated manual segmentations [28]. There was a large variability in results depending on the combination of reference and target subjects. Also of note, the results were very different when comparing the inverse pairing of the target and reference subject (target Subject i using reference Subject j versus target Subject j using reference Subject i). Assuming a perfect registration, the displacements of the nodal mapping found within these two registrations, would be equal and opposite in direction. As the results are not equal, this is evidence that the registrations did not map the moving image to the fixed image perfectly.

The optimal combination of subjects inputted into the algorithm was Subject 5 and 2 as the target and reference respectively. This combination of subjects resulted in an average RVE of 1.6% and presented a small inter-quartile range of 9.7% (Figure 4.14). The two subjects were the most anthropometrically similar, with heights of 154 cm and 160 cm, BMI of 27.5 $\frac{kg}{m^2}$ and 30.5 $\frac{kg}{m^2}$, respectively, which is likely to be the reason for these subjects presenting the greatest accuracy considering this metric. Contrastingly, the worst performing subject combination (Subjects 1 and 2 as target and reference, respectively) varied slightly in height (Subject 1: height 164 cm, Subject 2: height 160 cm) and greatly in BMI (Subject 1: 22.8 $\frac{kg}{m^2}$, Subject 2: 30.5 $\frac{kg}{m^2}$).

The DSCs found within the inter-subject analysis were poor overall, with mean DSC of 0.35 ± 0.16. Visually from the registered images (Figure 4.15) and particularly with the worst performing case, the registration was skewed by other anatomical artefacts, such as the fat surrounding the muscle. The low DSCs found across the cohort suggested that the muscle tissue was not being located or labelled well. This conclusion was compounded with the results found with the HD as the error metric, as the average HD was consistently greater than 20 mm, often with large inter-quartile range (Figure 4.14).

The contrast of the muscle-muscle boundaries was low within the images, when compared with the high contrast between the tissue boundaries (such as the air-fat boundary, and fat-muscle boundary). The presence of these strong boundaries biased the registration, as the greater greylevel gradient shift reduces the cost function that drives the registration. The cross-sectional area, fat depth and muscle volume varied widely between the subjects and consequently the registration map required to align the tissue boundaries was highly nonlinear. The high contrast between the external boundaries, coupled with the high non-linearity of the registration, have probably led to an inaccurate muscle segmentation.

Inter-subject segmentation performed far worse than the intra-subject segmentation. This was expected as the registration was required to overcome a higher degree of variation of the anatomical differences between subjects than between opposing limbs of the same subject [66]. There could be two potential reasons for this. Firstly, the registration was not capable of capturing individual muscle geometry. Secondly, the dissimilarity of the 5 subjects' anthropometric characteristics and distribution of tissues was too great for the registration to overcome. The first reason is unlikely as the segmentations resulting from the intra-subject registration were very similar to the manual segmentations, suggesting that the algorithm is capable of capturing 3D muscle structure. Therefore, we address the second hypothesis in the following chapter, exploring an enhanced pre-processing methodology to homogenise the images of different subjects, before registration.

4.5. Conclusion

Deformable image registration can result in good segmentation accuracy of the lower limb muscles, when using the contralateral limb as a reference. This can half the required time to segment a new subject. As with the manual segmentation of one limb, accurate segmentations of the contralateral limb can be generated. On the other hand, the accuracy of inter-subject deformable image registration is not accurate enough and needs improvements for assessing the muscle properties automatically.

Chapter 5:

Enhanced inter-subject registration, multi-atlas segmentation and the generation of an augmented imaging database

This chapter is partially based on a paper published in PLoS One (2023): ‘Deformable image registration based on single or multi-atlas methods for automatic segmentation and the generation of augmented imaging datasets’ by **W. H. Henson**, C. Mazzà, E. Dall’Ara. Doi: https://doi.org/10.1371/journal.pone.0273446

5.1. Introduction

Deformable image registration can isolate the lower limb muscles when used between the contralateral limbs of one subject, as shown with the relatively high segmentation accuracy within the intra-subject segmentation task in Chapter 4. Similarly, image registration has been used multiple times in longitudinal studies, automatically segmenting the muscles within a subject at a second timepoint, using segmented imaging data of that subject at an initial timepoint as the reference [93, 94]. However, the method used in Chapter 4 is not accurate enough to generate accurate segmentations of the muscles within new subjects when using other subjects as references. In fact, the anatomical variability (fat content, muscle volumes and lengths, etc.) between subjects is far larger than that within contralateral limbs. The approach however still has potential, as deformable image registration has previously been used to segment features from different subjects with great success, such as the study by Karlsson et al. (2015), where muscle groups (lower leg, posterior thigh, anterior thigh, abdomen, arm) were segmented from full body MR images, with accuracy greater than 0.9 DSC within all muscle groups other than the abdomen [111]. With adaptations to the method presented in Chapter 4, inter-subject deformable image registration could provide a route for the automatic segmentation of muscles.

The first aim of Chapter 5, therefore, is to update the deformable image registration algorithm proposed in Chapter 4 to increase the inter-subject segmentation accuracy for individual muscles of the lower limbs. This aim was achieved by combining a novel pre-processing algorithm and multi-atlas segmentation. The novel pre-processing algorithm was designed to shift the focus of the registration towards the muscle tissue by homogenizing the layer of fat surrounding the muscle tissue. The multi-atlas approach was used to remove inaccuracies by fusing multiple labels, in a similar way that it has been used for other types of tissue segmentation [109, 110, 136].

The second aim of this chapter was to explore the possibility of using the developed deformable image registration pipeline to generate fully segmented, augmented imaging datasets. Such datasets could be used for future biomechanical model applications [137, 138] or to calibrate probabilistic deep learning methods in order to automatically segment 3D geometry of individual muscles from MR images taken from several different cohorts. Studies using deep learning methods present high segmentation accuracy of muscles, but these methods are not widely accessible due to the main limitation of current approach: the requirement of large training databases (minimum ~20 segmented 3D images, the greater this number the more robust the method) [138]. Unfortunately, generating these segmented imaging datasets might not be well suited to MR imaging, given the associated high costs and manual processing time [28, 66]. On the other hand, data augmentation is a technique widely used in association with CNNs for the purpose of supplying greater amounts of training data and helping to generalise their application to image classification and segmentation tasks [137, 138]. Within this context, image registration has previously been used to generate augmented images to facilitate the analysis of brain tumours [137]and skeletal deformities [139]. This suggests that, while not attempted before, similar approaches might be adopted for muscle segmentation.

5.2. Methods

5.2.1. Subjects

All 11 subjects from the MultiSim study were used in this chapter (for details see Section 4.2.1). The 5 subjects selected for segmentation were the same as those outlined in Section 4.2.1. The other 6 subjects were used in this chapter for the generation of augmented images (Section 5.2.5).

5.2.2. Homogenizing fat tissue

Initial registration experiments of imaging sequences from two different subjects showed that the difference in the thickness of the fat surrounding the muscle tissue skewed the registration and resulted in a poor registration quality (Chapter 4). In order to homogenise this feature, the MR images of each subject were pre-processed to homogenise the distribution of fat tissue within the scans, focussing the registration on the muscle tissue. For each 2D slice in the MR dataset (example in Figure 5.1.A) within each subject, the air-skin boundary was located using a Canny edge detector [140]. The area within the skin boundary was filtered (Figure 5.1.C), in response to a threshold established from the greyscale frequency intensity plots of the images, creating a mask that contained only the muscle tissue (Figure 5.1.D). A layer of fat was wrapped around the muscle tissue (Figure 5.1.E and 5.1.F) to emphasise the outer boundary of the muscle tissue. The depth of this layer of fat was made equal to the optimal nodal spacing (NS, set to 5 mm, details in Section 4.2.4) [130].

There were two possible scenarios for the fat wrapping process: 1) the layer of fat within the image was greater than 5 mm, and 2) the layer of fat was less than 5 mm. In the first scenario, the subject's fat tissue was wrapped around the muscle tissue at a depth of 5 mm. In the second scenario, artificial fat was wrapped around the body which was built in response to the greyscale frequency intensity peak that represents the fat. The pixels within 5 mm of the muscle tissue that lay outside the body were randomly assigned values according to a uniform distribution with minimum and maximum equal to the mean ± standard deviation of the frequency intensity peak representing the fat. Through this operation, the muscle tissue remained unchanged, but the fat tissue surrounding the muscle was homogenised, meaning that all muscle characteristics (volume, shape, and fat infiltration) are all conserved. Figure 5.2 presents the original and pre-processed imaging data in the coronal plane for each of the 5 subjects, showcasing that the fat was successfully homogenised in this process.

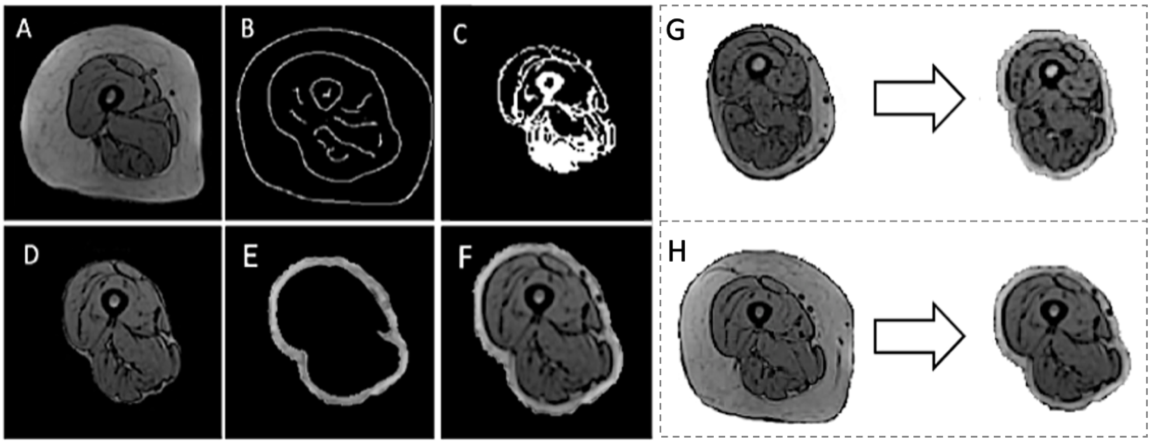


Figure 5.1: Homogenisation of the fat tissue surrounding the muscles. The process of masking the fat tissue (A, B, C, D) surrounding the muscles from the raw MR images (left) and wrapping in a homogenous layer of fat (E, F) for two images taken from different subjects (right). Two example images are presented in G and H, where the images on the left were not pre-processed and the images on the right were. The subject in example G had a fat tissue thickness less than 5 mm thick and was wrapped with artificial fat, where the subject along the bottom row had a fat tissue thickness that was sufficient.

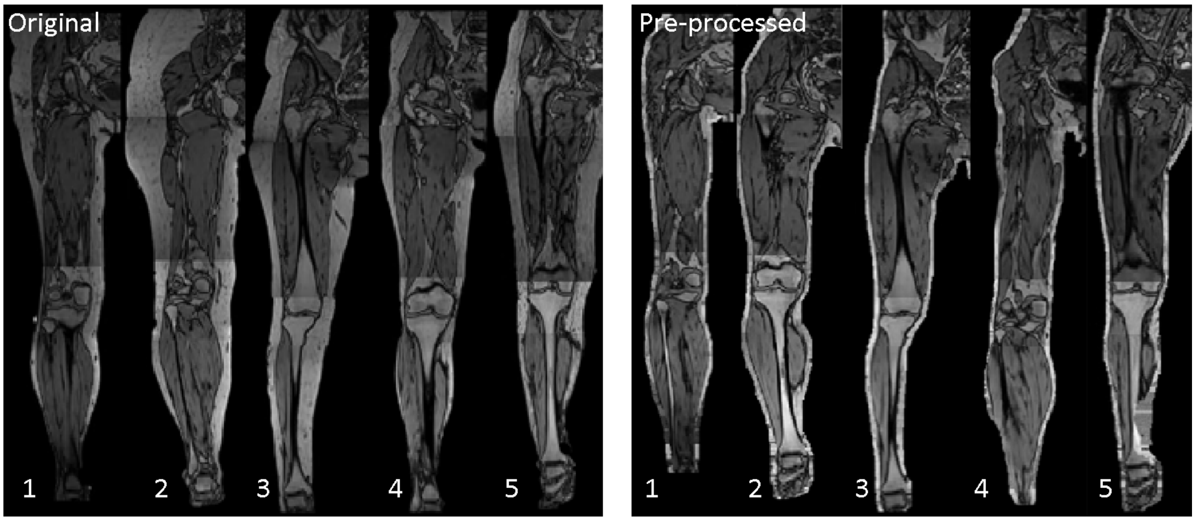


Figure 5.2: The original (left) and pre-processed (right) imaging data of the 5 subjects, shown with 2D images in the coronal plane.

5.2.3. Registration and single-atlas registration

Following pre-processing, subject imaging datasets were concatenated following the algorithm outlined in Chapter 4 Section 2.2 and were registered using the in-house deformable image registration algorithm (Sheffield Image Registration Toolkit, ShIRT) [130]. The images were registered using the optimal nodal spacing (mm) and smoothing coefficient for images such as those used, found in Chapter 4 Section 3.1. The 3D displacement vector field mapping the moving (reference) image to the fixed (target) image was computed by ShIRT and applied to the manually gathered reference segmentations, as outlined in Section 4.2.5. The registration and segmentation pipeline are summarised within the blue coloured portion of Figure 5.3.

The registration and segmentation pipeline was used to segment each of the 5 considered subjects. For each fixed subject, the other 4 subjects were used as the moving subject. Therefore, each subject was segmented 4 times using the pre-processed images. The accuracy of each set automatically generated muscle segmentation was tested independently (hence single-atlas segmentation) using the same methods presented in Section 4.2.7, i.e., the Relative Volume Error (RVE), Dice Similarity Coefficient (DSC), and Hausdorff Distance (HD) were computed for each automatically generated muscle segmentation. An additional metric called the total volume error (TVE) was also used, in order to assess how well the total muscle volume was preserved. The TVE was calculated following Equation 5.1 below, where the difference between the automatic (V_{A_i}) and reference (V_{M_i}) muscle volumes (for each muscle i = 1, 2, ..., 23) were summed and divided by the total of the reference volumes.

$$TVE = 100 \times \frac{\sum_{i=1}^{N} \left| V_{A_i} - V_{M_i} \right|}{\sum_{i=1}^{N} V_{M_i}}$$

Equation 5.1

The segmentation accuracy found in terms of the three error metrics were compared to the intra-subject results presented in the previous chapter. First, a Kolmogorov-Smirnov test was performed to test the distributions of the three error metrics found for the 20 sets of segmentations (4 sets of segmentations for each of the 5 subjects) for normality. The results indicated that the distributions of the error metrics were not normally distributed. Therefore, the non-parametric Wilcoxon signed rank test was used to compare the means of the intra-subject segmentation to that of the newly proposed inter-subject segmentation with the additional pre-processing step included. Thereafter, a second Wilcoxon signed rank test was conducted, comparing the means of the newly proposed inter-subject segmentation pipeline to those found in the previous chapter, without the additional pre-processing stage.

An additional analysis was performed, assessing the correlation between the segmentation accuracy and the volume of each automatically segmented muscle. To conduct this analysis, the Pearson's correlation coefficients between the RVE, DSC, and HD of the optimal single atlas segmentations (those with greatest mean DSC) and the volumes of the ground truth segmentations were computed.

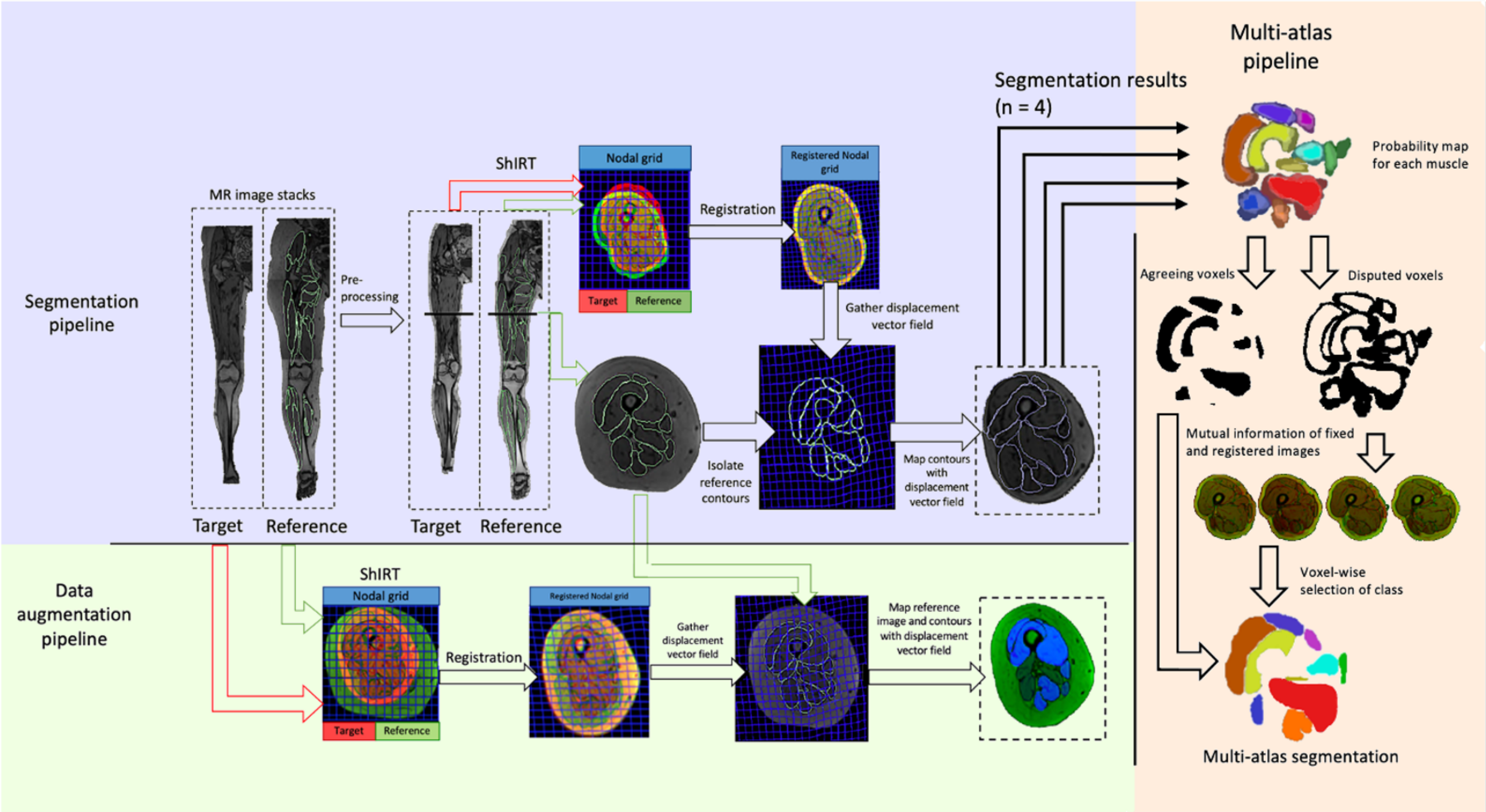

Multi-atlas pipeline
Segmentation results (n = 4)
Probability map for each muscle
MR image stacks
ShIRT
Nodal grid
Registration
Registered Nodal grid
Target
Reference
Segmentation pipeline
Pre-processing
Gather displacement vector field
Agreeing voxels
Disputed voxels
Isolate reference contours
Map contours with displacement vector field
Mutual information of fixed and registered images
Target
Reference
Target
Reference
Voxel-wise selection of class
Data augmentation pipeline
ShIRT
Nodal grid
Registration
Registered Nodal grid
Gather displacement vector field
Map reference image and contours with displacement vector field
Target
Reference
Multi-atlas segmentation

Figure 5.3: Registration, multi-atlas, and image augmentation pipelines. The image registration process, shown for one 2D slice of imaging data (location within imaging sequences highlighted with a black line). The segmentation pipeline shows the image pre-processing, registration, and map application stages. The data augmentation pipeline shows the registration of the raw images, with application of the map found through registration to the moving subject (shown in green) and its corresponding segmentations. The multi-atlas pipeline used is also shown on the right-hand side, where the four segmentations found in the single atlas approach are combined.

5.2.4. Multi-atlas segmentation

Multi-atlas registration methods are typically used to isolate the features of different registrations that agree, with the motivation of removing inaccuracies and increasing the reliability of the areas of agreement. These methods are in essence a final post-processing step wherein the agreeing features of multiple registrations are found and retained, and the areas that do not agree are removed.

Following the registration between the images of the five subjects, the resulting segmentations and registered images were used to define a multi-atlas segmentation for each muscle within each subject, in a process outlined in Figure 5.3. A probability map was defined for each voxel in each of the target images, representing the probability that a given voxel belongs to a certain muscle [106]. Considering each muscle individually within each target subject, the four automatically generated segmentations available through the single atlas approach were converted to binary images, where the voxels within the segmentation were set equal to 1 and all others to 0. The four binary images were summed, generating a probability map with each pixel taking an integer value from 0 to 4. All pixel values were then rescaled within the range (0,1), with the voxel values representing the probability that a given voxel belongs to a certain muscle. The voxels with value equal to 1 were included in the multi-atlas segmentation, and those with value equal to 0 were removed. The disputed voxels (those with probability not equal to 0 or 1), for any given muscle were assigned to a given muscle, based on which registered image best matched the target image in a certain area surrounding the disputed voxel [106]. The localised mutual information was used to find which registered image was the best match and was calculated voxel-wise as the sum of squared differences (the similarity measure used within the registration algorithm) between the registered and target images in a 25^3 voxel volume surrounding each voxel [106]. The registered image that presented the maximal agreement with the target image (that with the lowest sum of squared differences) at each of the disputed voxels was found, and the segmentation of that voxel from that registration was selected. All above processes were performed in MATLAB using matrix manipulation.

The error metrics found for the multi-atlas segmentations for each subject were compared to both the intra-subject segmentation results, and the best performing inter-subject segmentation results (selected as that with the greatest mean DSC). A Kolmogorov-Smirnov test for normality was performed to determine whether the error metrics were normally distributed. After failing this test, the Wilcoxon signed rank test was used between each of the newly proposed single atlas inter-subject registration results and the associated intra-subject segmentation results. The single atlas inter-subject segmentation with the lowest mean was then compared with the multi-atlas segmentation result, considering the three error metrics independently, using the Wilcoxon signed rank test.

5.2.5. Generation of augmented data

The deformable image registration algorithm was used to generate segmented augmented MR imaging data, summarised in green in Figure 5.3. The stacked MR imaging data from the right limb of the eleven participants were registered to each of the other subjects in the cohort, giving 110 combinations. To ensure that the registered image was distinct from both the reference and the target images, the registration was required to be imperfect. Therefore, no morphological pre-processing was applied. The displacement vector field outputted from ShIRT (Figure 5.3) was used to deform both the MR imaging sequence and the manual muscle segmentations of the reference subject, using linear interpolation to mimic the interpolation method used within ShIRT. The output of each of these processes was a fully segmented 3D image that was dissimilar to both the reference subject and the target subject (Figure 5.3).

A four-point criterion was used for checking both the images and the segmentations to ensure anatomical credibility of the augmented dataset: a) the boundaries of the long bones and the skin must be reasonably smooth and continuous; b) the positioning and orientation of the joints must be anatomically viable, with the bones fitting together realistically; c) the muscle segmentations should reflect the muscle structure; and d) the location of each of the muscles relative to one another must be realistic (e.g. the vastus lateralis must be lateral with respect to the vastus medialis). If any one of these criteria were not met, the augmented dataset was discarded. Out of the retained datasets, 15 chosen at random were retested by a different operator to confirm the specificity of the inclusion criteria.

Finally, the available muscle volumes were compared from within the augmented and original databases to measure the variability of muscle volumes between the two databases. The mean volume within each database was computed for each of the 23 muscles considered. The difference between the volume of each muscle within the database and the average was then calculated, and this value was normalised against the mean volume. The resulting values were percentages representing the distribution of available muscle volumes within each database (with and without augmentations) which after normalisation, could be compared.

5.3. Results

5.3.1. Single and multi-atlas segmentation results

A visualisation of an example registration and of the results of one segmentation are highlighted in Figure 5.4 for images taken from the hip, thigh, and shank, respectively. While the deformable image registration has accurately identified the muscle tissue in the target subject in most cases (yellow), some regions were not correctly registered (red or green). The segmentation results reflect this, where the registration appears successful overall, and the automatic segmentations are geometrically very similar to the ground truth segmentations. There are areas within the automatic segmentations that do not reflect the reference segmentations, such as the gluteus maximus in the hip section, and the tibialis muscles within the shank section. The automatic segmentations within the thigh section mostly agree with the reference segmentations. The pre-processing required 144 ± 15 seconds across the 5 subjects and the registration required between 70 and 94 minutes, slightly less than before the pre-processing. This reduction in time to perform the registration was caused by the smaller size of the images that were registered. All processes were performed using an Intel® Core™ i7-7700 CPU @ 3.60 GHz.

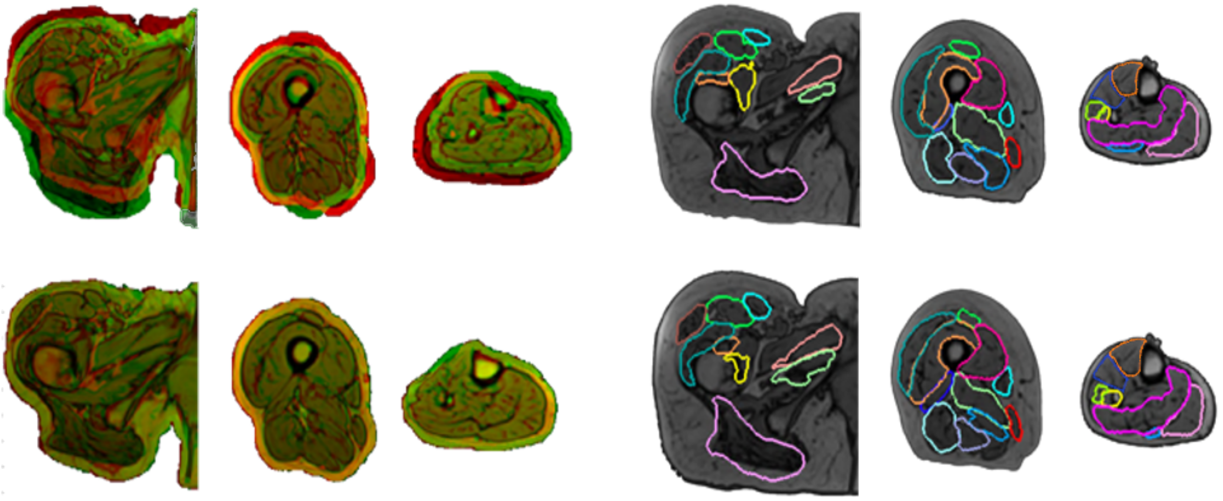


Figure 5.4: Qualitative interpretation of segmentation results. Registration and segmentation results from the combination of subjects resulting in the median average DSC (subject 4 and 2 as the target and reference, respectively). The registration inputs (top row) and outputs (bottom row) for these combinations of subjects are shown in the group of images on the left. The segmentation results on the right are shown in three image groups, where the ground truth and automatic segmentations for the target subject are shown in the top and bottom row respectively. The muscles that are not highlighted within the images, were found not to be segmented with an appropriate level of repeatability.

5.3.1.1. Volume error

The TVE for the entire muscle body was 8.2 ± 5.1 % (mean ± standard deviation) across all subject combinations. The mean RVE for the individual muscles was found to be below 12.8% for all combinations and all upper quartiles were below 40% error (Figure 5.5). The best performing combination in terms of RVE was subject 5 with 1 as the target and reference respectively, which had the smallest mean (-2.2%) and with the lowest quartiles (lower and upper quartiles of -10.5% and 6.4%, respectively). The RVE was consistent across all muscles, with no correlation found between muscle volume and RVE ($p = 0.159$); the muscles with the highest variability within this cohort (tensor fascia latae, rectus femoris, and peroneus longus) were the outliers within the distributions of RVE, probably as the registration algorithm was unable to overcome the large differences in volume. In 7/20 cases, the mean RVE was significantly lower in the left to right analysis than the inter-subject results (minimum difference between means = 2.85%, maximum p-value = 0.83). The multi-atlas analysis provided a lower inter-quartile range in terms of RVE and resulted in the mean RVE across the five subjects falling within the acceptable range of error of ±10% error [28, 66], with minimum and maximum of means = (-2.4%, 9.0%). This cannot be said for the single atlas registration results with minimum and maximum of means = (-17.8%, 14.2%). No significant differences were found between the mean RVEs found within the left to right analysis and the multi-atlas analysis (maximum difference between means = 8.6%, p-value = 0.20, target Subject 4). Furthermore, no significant differences were found between the mean RVEs found for the optimal single atlas segmentation results, selected as those with largest mean DSC (maximum difference between means 10.4%, p-value = 0.39, target Subject 3 using reference Subject 4 in single atlas segmentation).

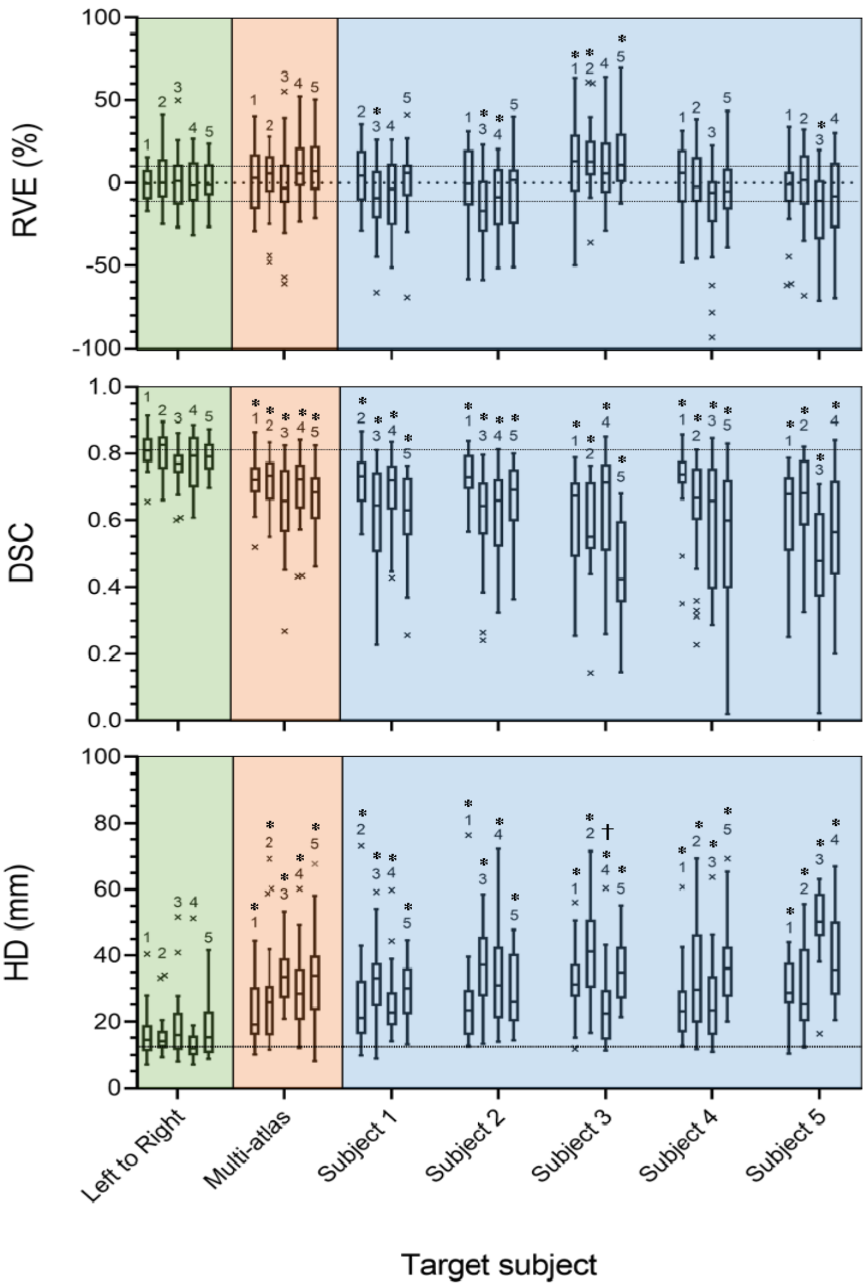


Figure 5.5: Numerical results for the intra-subject, multi-atlas, and single atlas analyses. Relative volume error (RVE, in %, top), Dice similarity coefficient (DSC, non-dimensional, centre), and Hausdorff distances (HD, in mm, bottom) found for each muscle in each subject, for the left to right/intra-subject analysis (green), multi-atlas (orange), and single atlas inter-subject approaches (blue). In both the left to right and multi-atlas analysis, the numbers above the boxplots denote the subject segmented. The numbers above each of the boxplots in the inter-subject approaches denote the reference subject used within the registration. The dashed line in RVE plot shows the acceptable level of RVE resulting from inter-operator dependence, prescribed by Montefiori et al. 2019 [66]. The grey dashed lines in the DSC and HD plots represent the mean values from the intra-subject analysis for comparison. The * above each of the plots highlights a statistically significant difference ($p < 0.05$) between the mean of the boxplot beneath, and the intra-subject (left to right) segmentation. The best performing single atlas segmentation (highest mean DSC) was compared to the multi-atlas segmentation. A † above the best performing single atlas results shows a statistically significant difference ($p < 0.05$) between the mean error of the optimal single- and multi-atlas segmentations.

5.3.1.2. Dice similarity

When looking at the segmentations of the five subjects obtained using the other four as reference subjects, very variable results were observed. The greatest average DSCs were those resulting from the segmentation of Subjects 1,2, and 4, using Subjects 2, 1 and 1 as the reference, respectively. The mean DSCs found for these combinations of subjects were greater than 0.70, lower quartiles greater than 0.67, and with a wide spread of results ($0.35 < DSC < 0.88$). Subjects 3 and 5 were segmented with a consistently lower mean DSC, with the average DSC considering all reference subjects found to be 0.61 and 0.60 respectively (0.69, 0.69 and 0.67 for Subject 1, 2, and 4, respectively). Comparing the single atlas segmentation results to the left to right analysis, the mean DSC was significantly lower in the single atlas results across all 20 combinations of subjects (minimum difference between means = 0.06, maximum p-value = 2.9×10^{-3}). Similarly, all 5 of the mean DSCs found using the multi-atlas segmentation method were significantly lower than the left to right analyses (minimum difference between means = 0.08, p-value = 5×10^{-4}). No significant differences were found between the optimal single atlas results and the multi-atlas results (minimum p-value = 0.078, target Subject 2 using reference Subject 1 in single atlas segmentation). There was a weak correlation found between muscle volume and the DSC of the automatic segmentations (R^2=0.332, p-value=0.003), suggesting that the larger muscles were slightly better segmented in terms of DSC (see later section and Figure 5.7).

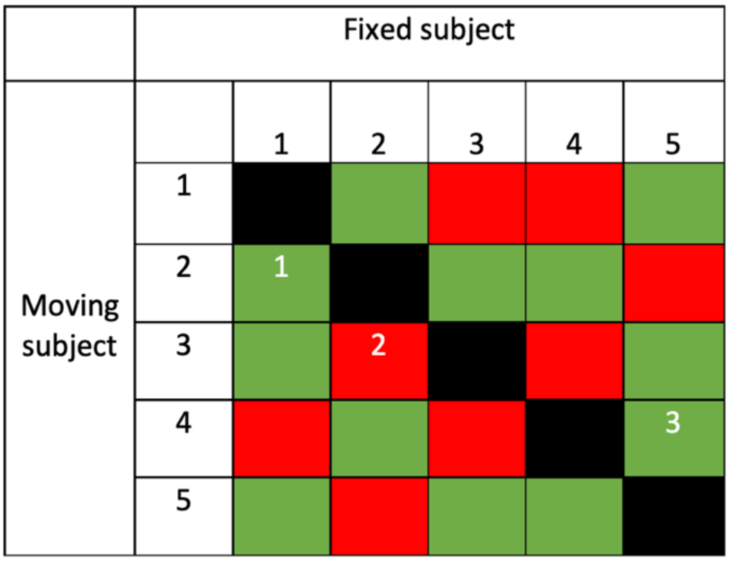
5.3.1.3. Hausdorff distance

Overall, the mean HD was typically between 15 mm and 30 mm, with the upper quartile being below 40 mm, other than the segmentations of subjects 3 and 5 using subjects 2 and 3 as references, respectively (Figure 5.5). The spread of results was large, with Interquartile ranges (IQR) being between 7 mm and 21 mm. The average HD found within the intra-subject analysis was 17.7 mm, much lower than in the other analyses. The HD distances in the multi-atlas analysis were comparable to the best cases within the single atlas results, with 4/5 means being comparable (p > 0.05). The exceptional case was the mean HD within the optimal single atlas segmentation for Subject 3 (using subject 2 as the reference) was significantly lower than the mean HD found using the multi-atlas approach (difference in mean HD = 9.8mm, p-value = 3.2×10^{-3}). The mean HDs across all 20 single atlas segmentation results was significantly greater than the associated left to right results (minimum difference in mean = 5.9mm, maximum p-value = 6.8×10^{-3}). Similarly, the mean HDs across the multi-atlas segmentation results were all significantly higher than the left to right analyses (minimum difference in mean = 6.7mm, maximum p-value = 6.8×10^{-3}). There was no correlation found between the

HD and the size of the muscle (single atlas method) for which the HD was calculated (p-value=0.089), indicating that the error was consistent across muscles of all sizes.

5.3.2. Segmentation of original and morphologically altered images

Comparing the DSC found when registering the original and morphologically altered images, there was a significant ($p < 0.05$) increase in segmentation accuracy in 12/20 of the combinations of subjects, as shown in Table 5.1. Three combinations were chosen at random with the registration inputs and outputs (with and without the additional pre-processing stage) showcased in Figure 5.6. The segmentation accuracy (in terms of DSC) of the first and third combinations selected were significantly improved ($p < 0.05$), and this is visible with the increased registration quality of the registered images in these cases in Figure 5.6. On the other hand, the combination of subjects that did not improve were registered with very similar registration quality in the original and morphologically altered images, suggesting that there was no reduction in registration quality. Additionally, in all cases, the mean DSC resulted from registration of the morphologically altered images was greater than or equal to that of the original images, suggesting that there was no reduction in segmentation accuracy.



		Fixed subject				
		1	2	3	4	5
Moving subject	1	⬛	🟩	🟥	🟥	🟩
	2	🟩 1	⬛	🟩	🟩	🟥
	3	🟩	🟥 2	⬛	🟥	🟩
	4	🟥	🟩	🟥	⬛	🟩 3
	5	🟩	🟥	🟩	🟩	⬛

Table 5.1: Statistical analysis of the effect of homogenising the fat surrounding the muscle tissue on the Dice Similarity Coefficient (DSC) of the segmentation, resulting from registration. Green squares represent a statistically significant ($p < 0.05$) increase in the segmentation accuracy (in terms of the DSC) through the pre-processing step of each combination of fixed and moving subject. The numbered squares highlight the fixed and moving subject combinations used in the registration and segmentation pipeline in Figure 5.3. Where the fixed and moving subject are identical in the table, the squares are filled in black.

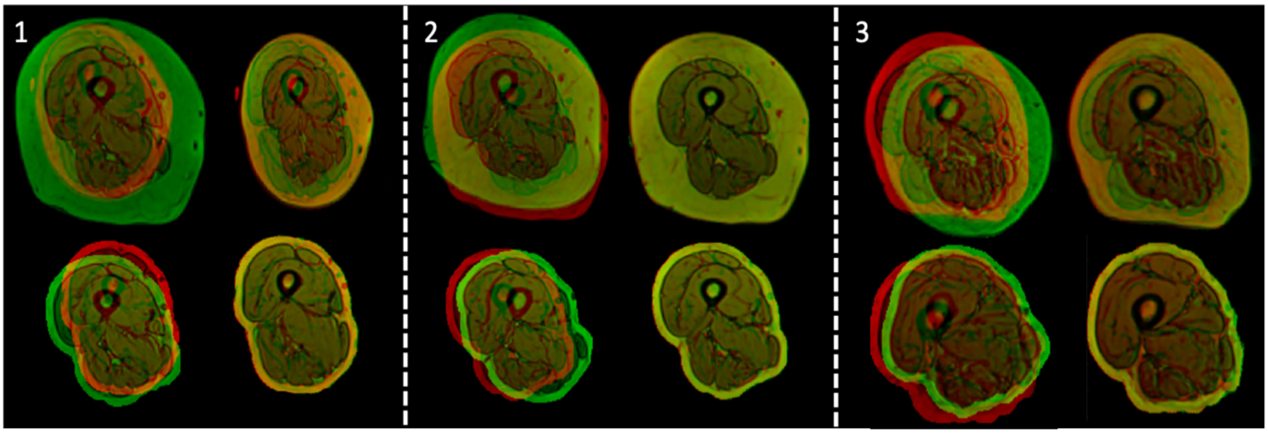


Figure 5.6: Registration inputs (left column of images in each block) and outputs (right column of images in each block), both with (bottom row of images in each block) and without (top row of images in each block) the pre-processing step. The subject combinations are numbered and highlighted in Table 5.1.

5.3.3. Comparison of results with literature

The RVE, DSC, and HD of the optimal single atlas segmentations (in terms of DSC) for each muscle of the 5 subjects are plotted against the volume of the automatically segmented muscle (mm^3) in Figure 5.7. Also plotted in Figure 5.7 are areas of the graphs that contain the errors found in three previous studies that segmented individual muscles. The data cloud of this study falls in the area of at least one previous study, particularly in the RVE and DSC plots. The HDs found in this study across all muscle volumes are greater than those found in Fontana et al. [93].

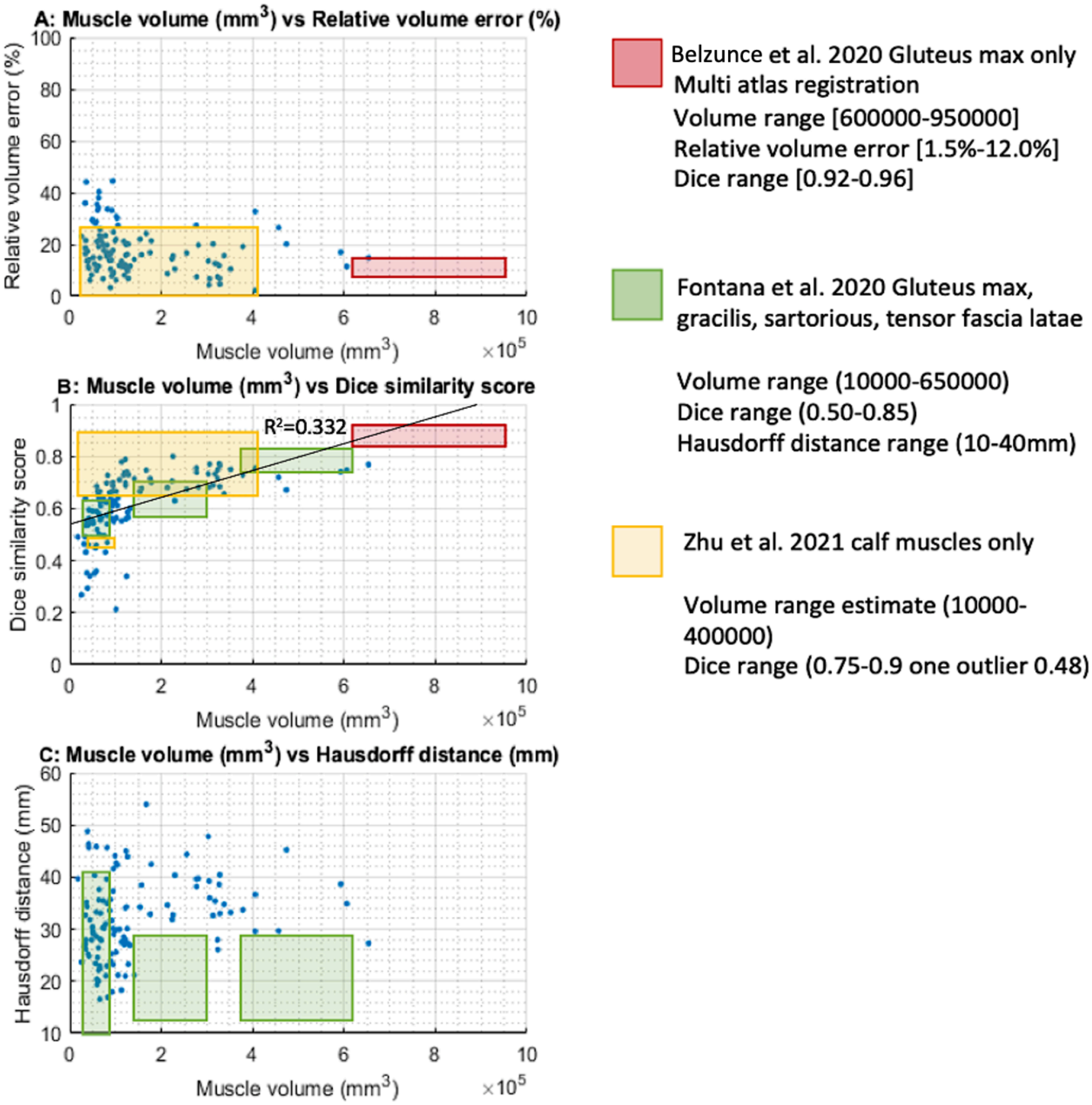


Figure 5.7: Comparison of muscle volume vs the three error metrics (A: RVE, B: DSC, C: HD) across all muscles for five subjects. Coloured areas of the graph highlight areas of muscle volume vs the three error metrics that can be found in three previous studies that automatically segmented individual muscles. The studies referred to are Belzunce et al. [141], Fontana et al. [93], and Zhu et al. [87].

5.3.4. Augmented data

After initial checking by the author, 69 of the 110 generated augmented datasets passed the inclusion criteria. 15 out of the 69 datasets were rechecked by an expert in muscle segmentation and all 15 passed, giving 100% specificity. Figure 5.8 shows some examples of the augmented images generated. Visually, the augmented images are well segmented, and are dissimilar to the reference subjects, particularly in the second row of images, where the relative fat depth of the moving subject (green) is retained, but the cross-sectional area of the thigh is equal to the fixed subject (red). The misalignment of the muscle tissue within the registered images, visible as concentrations of either red or green colours, helps to establish the difference in muscle geometry between the registered and original data. The augmented subjects generated for 1 target subject (Subject 1) are presented in Appendix 2, for more detailed visual comparison.

The anatomical variability of the muscles within the augmented database is compared to the original 11 subject database (Figure 5.9). The volumes of each of the muscles in the original and augmented databases were normalised against the corresponding average muscle volume for each muscle in the respective databases. The percentage difference with respect to the average value was then calculated for each muscle (Figure 5.9). The muscle volumes available within the augmented database were found to have a greater range of volumes, often 1.5 to 2 times greater than in the original database.

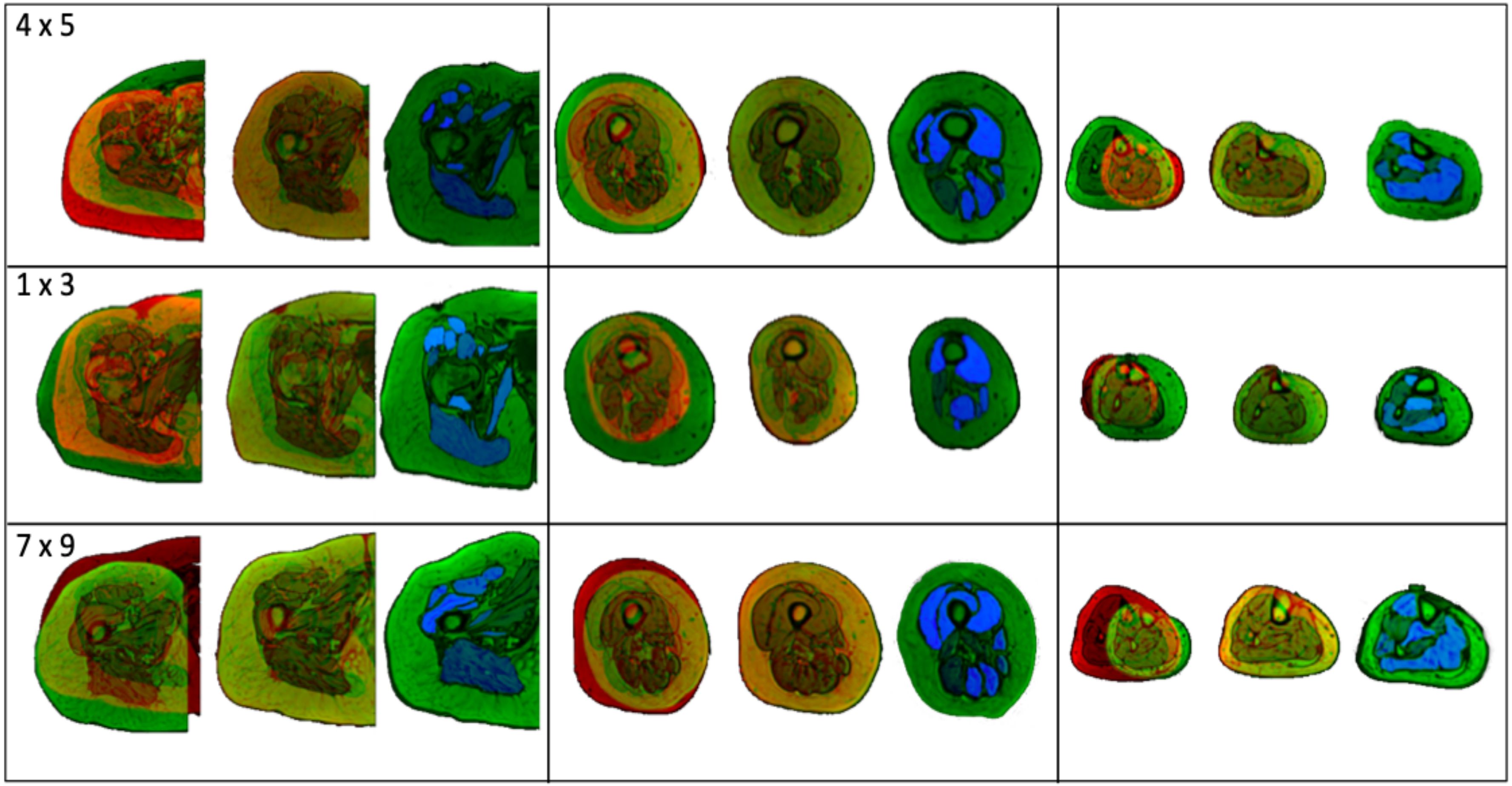


Figure 5.8: Exemplar augmented images. Inputs, outputs and resulting augmented subjects. Each row of images presents results within the hip (left), thigh (centre), and shank (right) for 3 subject combinations chosen at random (target x reference: 4 x 5 (top), 1 x 3 (middle), 7 x 9 (bottom)). Within each box, there are the inputted images into the registration (left), registered images with corresponding target image (centre) and resulting segmented, augmented images (right). The muscle labels are visible within the augmented images as the blue areas. Each muscle is assigned a distinct greyscale value and the labels are assigned alphabetically.

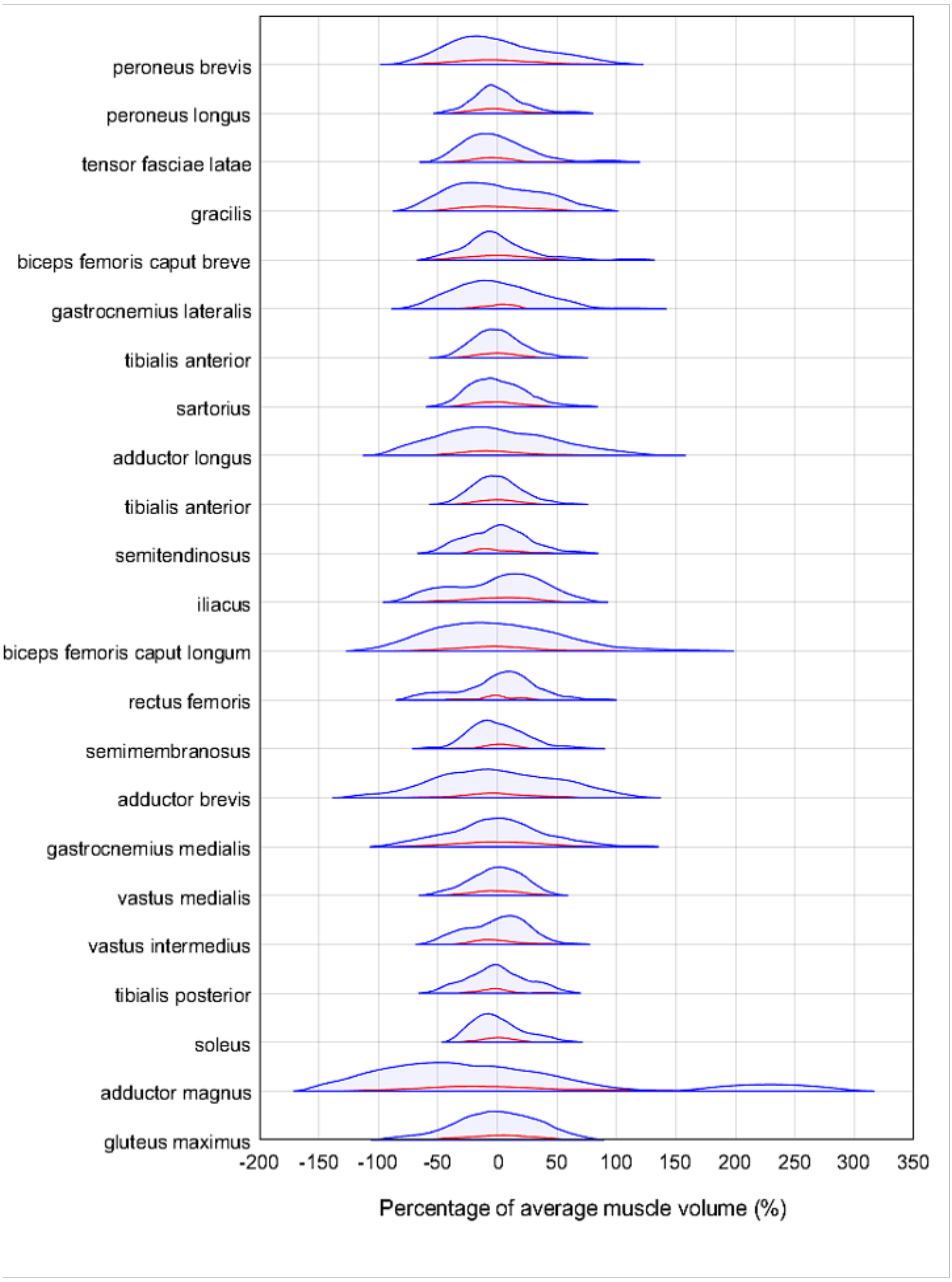


Figure 5.9: Enhancement of muscle volume variability through image augmentation. The anatomical variability of muscle volumes for each muscle, ordered from smallest to largest within the original and augmented databases shown in red and blue, respectively. The height of the distributions was not normalised, and the violin plot contains 95% of the data, with 2.5% of data cut off from each side, removing outliers.

5.4. Discussion

This chapter aimed at proposing a fully automatic tool to segment 23 major lower limb muscles simultaneously from MR imaging data from different subjects using morphological image processing, deformable image registration and a multi-atlas approach. Furthermore, the registration tool was used to generate a unique dataset including 69 fully segmented, augmented 3D images. To the best of the authors' knowledge, this study represents the first attempt to segment complete 3D muscle geometry of many individual muscles simultaneously using deformable image registration while using different subjects as a reference. Moreover, a multi-atlas approach was used for the segmentation of many individual muscles simultaneously, which is yet to be investigated in this way.

All 23 muscles were segmented from five subjects with moderate success. The registration quality was high considering the combination of subjects that resulted in the median average DSC (Figure 5.4) which suggests that in most cases, the registration performed as intended. This was confirmed by the TVE, lower than 10% on average. However, all metrics reflected a lower accuracy for the segmentation of individual muscles. Within both the inter-subject and multi-atlas analyses, the individual muscle RVE was typically larger than that of an acceptable level of inter-operator dependence (±10%) [66], with one or both of the lower and upper quartiles often exceeding ±10% in most subject combinations. The mean absolute RVE within the optimal subject combinations was 12.7%, meaning that on average, there was an over or underestimation of the muscle volume greater than that incurred by the effects of operator variability. This indicates that the method would be best suited when only interested in the volume of the overall muscle body. Capturing the total muscle volume has proven useful in studies such as Handsfield et al. [142], where regression equations were used to estimate individual muscle volume from total muscle volume and other anthropometric data such as height and BMI. The DSC results, on the other end, indicate that if the purpose of the segmentation was that of extracting internal muscle characteristics, such as the level of fat infiltration [60], then alternative approaches should be pursued regardless of the inclusion of a multi-atlas postprocessing step. Possible improvements of the method could come from a more targeted selection of the reference subject, which as shown by the reported results (Figure 5.5) can increase the accuracy of the approach both in terms of individual muscle volume and DSC. Though, it is extremely likely that a better reference than a subject's contralateral limb would be seldom found as an input of the registration algorithm. Therefore, the use of deformable image registration of images acquired with these acquisition parameters for the purpose of individual muscle segmentation could be segmented at best with an accuracy comparable to that of the left to right analysis (~0.8 DSC).

The geometry of each of the 23 muscles was captured moderately well for the optimal combination of subjects in the inter-subject analyses (those with greatest lower quartile), with mean DSC of 0.74 and IQR range of 0.71 < DSC < 0.77, and the multi-atlas approach presented very similar results. However, these quantitative measures of accuracy are significantly lower than the inter-operator dependence of the manual process, which, within the literature [13, 66, 83, 133] is consistently found to be DSCs of around 0.90 for the muscles considered in this study. While the pair of subjects led to the best results in terms of DSC were the most similar in terms of height and BMI, these anthropometric characteristics were very different in the pair having the second-best DSC (mean = 0.74, IQR of 0.69 < DSC < 0.79). This suggests that the newly proposed masking process achieved the goal of homogenising the subject imaging data and could be adapted for the removal of unwanted artefacts from within medical or indeed any other images. The worst performing combination of subjects (those with the lowest upper quartile), with mean 0.45 DSC across the 23 muscle segmentations, were those with the greatest difference in age (16 years) but similar height, weight, and BMI. One could suggest that the difference in muscle quality (and therefore difference in appearance in the images) between these two subjects could be large, in response to age-related degradation of the muscle tissue [23]. The muscle quality greatly affects the appearance of muscles within medical images and would certainly affect the quality of the registration [23].

Particularly successful approaches in the literature that used 3D deformable image registration to perform muscle segmentation were those based on longitudinal data, such as Le Troter et al. [94] and Fontana et al. [93], who attained average DSCs of 0.90 and 0.85, respectively. Similar to the latter were the DSC values here found when registering the left to right limb in the same subject, which could be used to propagate segmentations of one limb to the contralateral limb, halving the time to segment the lower limbs. Notably, the approaches found in the literature still require the manual segmentation of each subject as the baseline. Moreover, these studies segmented only a subgroup of the muscles segmented in this study. Last but not least, the images collected in this study were not optimised for muscle segmentation, as they were the widely used T1-weighted MR images with lower resolution along the long bone sections (see Section 4.2.1.1). These characteristics of the images used may limit the registration, as the muscle tissue was not as clear as it would be with better acquisition parameters.

Overall, the main limitation of the proposed method is the non-satisfactory capture of individual muscle volume. There are several potential reasons for this finding. Firstly, this could have been caused by the propagation of inaccuracies associated with the manual segmentations of the reference images through the registration. However, this aspect is likely to have had a negligible effect since the muscles with high inter-operator variability [66] were discarded at source. More likely, the issue lied in the fact that the muscle-muscle boundaries present a weak grey-level gradient, in contrast to the muscle-fat boundaries, which are shown to have a strong grey-level gradient within the MR images (figure 5.1). Since ShIRT registers grey-level gradients within the inputted images [130], the muscle-fat and muscle-bone boundaries were registered to a higher degree of accuracy than the muscle-muscle boundaries. The use of other MR imaging acquisition settings, such as the Dixon method for fat suppression, could further enhance muscle-muscle boundaries, however, these images were not collected at the time of data acquisition. The imbalance in the accuracy of the registration of the different tissues is highlighted by the greater RVE of the individual muscles, when compared to the total volume error. Moreover, ShIRT was the only registration algorithm tested. Other available registration algorithms [93, 94] could improve the accuracy of the segmentation but will have to be tested on the same dataset. For this reason, the datasets including the input MR images, the manual segmentations and the ShIRT inputs have been shared here (https://doi.org/10.15131/shef.data.21739733) for future comparison with other registration tools.

Another source of error could lie within the optimisation process of the registration parameters (NS and smoothing coefficient) [143]. While in this study these parameters were optimised for the highest overall performance in segmentation accuracy across all considered lower limb muscles, the values could be further optimised for different areas of the limb. This was not implemented in this study as a rewriting of the registration toolkit would be required. The multi-atlas approach was employed to overcome the potential limitations of the registration procedure, incorporating a probabilistic evaluation of which regions of the images belong to each muscle. This method has been used in the assessment of other tissues in the body, such as brain tissue or the prostate, with good results [70, 106, 136]. The method did not have the same impact in this case, most likely due to the sheer number of different muscles assessed, which resulted in a great number of disputed voxels within each target image; a problem which would not be associated with medical image segmentation problems with fewer classes required to be segmented. It has been noted within the literature that this voting system is best suited for a thin layer of disputed voxels surrounding the tissue of interest [106], which was not the case in the automatic segmentations outputted from the inter-subject analyses.

Despite the above limitations, the image registration protocol proposed here was clearly proved useful when adopted to generate an augmented image database of 69 subjects having a much broader range of muscle volumes and geometries than the original 11 subject database. This result came after removing 41 anatomically unrealistic datasets, which required some manual checking on the augmented datasets, suggesting that similar care should be taken if this method is replicated. These datasets made publicly available (augmented images: https://doi.org/10.15131/shef.data.20440164, augmented images segmentations: https://doi.org/10.15131/shef.data.20440203), can be used to train deep learning methods [137, 144]. Machine learning and deep learning methods are now dominant tools used in the field of medical image segmentation [13, 82, 83]. Considering the 23 muscles in the present study, the average DSC were found to be around 0.75, including only the optimal reference subject for each target subject. In comparison, deep learning methods have been used to segment the lower limb muscles with average DSC between 0.85 [82] and 0.90 [80]. These tools are typically only suitable for studies with extremely large cohorts, but this problem has been alleviated within some medical image analysis fields, such as brain tumour assessment [137] and bone segmentation [145], through data augmentation. However, this technique is yet to have been explored for muscle segmentation and the database presented here will foster efforts in this direction. To the best of the author's knowledge, in fact, this is the first study to provide a large, multi-operator assessed set of fully segmented, labelled, augmented MR imaging sequences of the lower limb. In future work, these augmented datasets will be used to calibrate CNN models, with the potential to increase segmentation accuracy and lead to a solution for more accurate automatic segmentation and characterisation of muscles *in vivo*.

5.5. Conclusion

This chapter presented a novel, fully automatic muscle segmentation method using image registration, aimed at segmenting all lower limb muscles simultaneously. The 3D deformable image registration algorithm used in this work is limited in its capacity to perform individual automatic muscle segmentation with a high accuracy. Nevertheless, this approach can be useful to provide total muscle volume and can be used as a tool to increase the number of reference datasets, enabling other methodologies (e.g., machine learning-based methods) to be explored and properly trained. Explicitly, the publicly available augmented database built in this work would enhance any future study that would aim to use deep learning approaches for the segmentation of muscles from T1-weighted MR images.

Chapter 6:

Traditional and novel deep learning-based segmentation approaches

6.1. Introduction

Deep learning is the branch of Artificial Intelligence (AI) that is used in state-of-the-art image analysis applications, so named due to the large number of connected layers used to isolate key features from within images [114, 118]. These computational architectures have been proven powerful in the segmentation of several tissues from medical images with a level of accuracy comparable to manual segmentation by human experts [80, 82, 87]. Since 2015, upon publication of the paper that first introduced UNet [114], there has been an explosion in the number of research articles using UNet or variants of it to segment medical imaging data. The current consensus in the literature is that this base network architecture is still (7 years later) the state of the art, with its contracting pathway focussed on 'what' is in the image, and the expanding pathway focussed on 'where' the key information is in an image. To demonstrate this, the Medical Segmentation Decathlon [146] is currently the best example. The challenge was set by a consortium of medical imaging groups to align the segmentation approaches used within the community and submit one generalized solution to the problem of medical image segmentation. The challenge was set to competitors to segment ten tissues/organs (brain tissue, prostate, lung, abdominal organs, heart tissues, lesions, cell nuclei, and others) within the human body, captured with different imaging modalities (CT, MR imaging, electron microscopy, or fluorescence microscopy). The best model submitted, "not new UNet" (nnUNet), was able to segment all of the 54 imaging databases included in the challenge, achieving the greatest accuracy compared to the other models submitted (DSC in the range (0.5, 0.97)) in 39 of the categories [147]. The nnUNet was named as such as it maintains the original UNet architecture, but with an additional optimization of the hyperparameters used to train the network. For this reason, logic dictates that the UNet network architecture is still the best that the research community has to offer to solve the problem of medical image segmentation.

Muscle segmentation was not among the challenges submitted in the Medical Image Segmentation Decathlon. Possible reasons for this could be that there are few openly available datasets that would be large enough to facilitate learning-based methods, or that the datasets that are currently available are inhomogeneous in their acquisition parameters [116]. Studies using deep learning techniques to perform muscle segmentation are uncommon, with very few (but a growing number of) studies available within the literature. Three noteworthy studies conducted by Ding et al. [80], Zhu et al. [87], and Ni et al. [82], reflected the current state of the art research in the application of these deep learning-based tools to perform muscle segmentation.

Ding et al. segmented seven (testing) MR imaging datasets using a standard UNet architecture trained on 23 (3:1 split in training to validation) imaging datasets [80]. The network was trained and used to segment four regions of interest from multi-acquisition (water and fat suppression) imaging datasets, two of which were individual muscles and two were muscle groups (knee extensors and flexors). The muscle groups were segmented with a mean DSC (calculated through comparison with a manual segmentation) upward of 0.9, across the seven datasets segmented. On the other hand, the individual muscles were segmented with a reduced accuracy, with DSC of 0.86 on average across the seven datasets, suggesting that the segmentation of individual muscles is more challenging than muscle groups.

Zhu et al. built on this work and investigated whether adjustments to the network architecture would yield individual muscle segmentations of improved accuracy [87]. In that study, Zhu et al. segmented 13 muscles from T1-weighted images acquired from the shank, with four test children, using the images from 16 children for training and validation (age range: 5.4 to 14.8 years old, 6 female, 14 male, 15:1 training to validation split), both with and without cerebral palsy. Six network architectures were tested in this study, including 2D, 3D, and hybrid models. The original UNet and the hybrid models presented similar accuracy (0.87 and 0.89 in DSC respectively). Neither study sought to segment all individual muscles from lower limb MR images.

Ni et al. used UNet to segment all lower limb muscles of a testing dataset of 13 young healthy adults, using 51 subjects for training the network (training and validation split not stated) [82]. In this study, an individual network was trained for each of 35 muscles, each consisting of two neural networks trained in series: the first to crop the lower limb 3D MR image in order to isolate only the muscle being segmented, and the second to segment the muscle from the cropped image. The accuracy of the segmentations was comparable to that achieved in an inter-rater analysis for 14 of the muscles, and slightly worse in the other 21, averaging 0.9 in DSC across the 35 muscles segmented [82].

While these studies showed encouraging results for automatic muscle segmentation, there are limitations that should be addressed. Segmenting all individual muscles from the entire lower limb is of great interest, as it would allow in depth analysis of the structural and functional health of each muscle individually, and future studies should aim to fulfil this objective [28]. Also, the requirement of costly computing equipment and the high computational cost incurred to train these algorithms reduce the availability of deep learning-based approaches [125]. Future algorithms should aim to make simplistic new architectures, which do not require such excessive training, and use multi-classification networks, as opposed to training an individual network for each muscle [87]. Lastly, the need for large training databases deters research groups from using these tools. Traditional image augmentation strategies, such as scaling, rotation, reflection, and non-linear deformation, have been tested for this image segmentation problem, and are not appropriate for this form of segmentation as all images are acquired in identical frames of reference. Therefore, all images in the database appear without rotation, reflection or scaling, and future image will be collected with similar or identical methods. Moreover, maximising the available data that can be used to train the network should also be pursued, such as medical imaging meta-data, which might be used to influence the learning process of neural networks. Therefore, innovative methods to overcome these limitations should be investigated.

6.1.1. Aims and objectives

The Overarching aim of this chapter was to address the limitations currently highlighted within the literature, by testing state of the art and novel deep learning-based approaches. With this aim in mind, the following objectives were defined:

1) To build and test the two main state of the art deep learning models, which have been used within the literature to segment other tissues from biomedical images: the UNet [114], and the Attention UNet [148].

2) Create and test a novel network architecture ('Spatial channel UNet') to make use of the known spatial location of each image.

3) Evaluate the effect of incorporating the augmented imaging database generated using deformable image registration (Chapter 5) in the calibration of the tested networks.

4) Retrain the best performing network for different testing subjects and compare the segmentation accuracy with that obtained with deformable image segmentation.

6.2. Methods

6.2.1. Data acquisition

6.2.1.1. Participants and image acquisition

In this study, as with the previous image registration study, the MultiSim cohort was used. Briefly, retrospective T1-weighted MR images of the lower limb from 11 post-menopausal women (mean (standard deviation): 69 (7) years old, 66.9 (7.7) kg, 159 (3) cm) were used for this study [15]. Images were collected using a Magnetom Avanto 1.5T scanner (Siemens, Erlangen Germany), with an echo time of 2.59 ms, repetition time of 7.64 ms, flip angle of 10 degrees. The study was approved by the East of England – Cambridgeshire and Hertfordshire Research Ethics Committee and the Health Research Authority (16/EE/0049). The MR images were acquired in four sequences, capturing the hip, thigh, knee, and shank. To reduce scanning time while still providing detailed geometries of the joints for use within the original study, the joints were acquired with a higher resolution (pixel size 1.05 mm^2, slice spacing 3.00 mm) than the long bone sections (pixel size 1.15 mm^2, slice spacing 5.00 mm).

6.2.1.2. Manual segmentation and label generation

The manual segmentations used to generate segmentation labels were the same as those used in the previous chapters. The T1-weighted image sequences were stacked in MATLAB forming one continuous 3D image from hip to ankle, firstly by homogenising the resolution of each of the imaging sequences taken from the different sections to be 1.00x1.00x1.00 mm^3 through tri-linear interpolation (interp3, MATLAB 2006a) as described in Chapter 4 Section 2.2.2. The fields of view of the images across the four sequences were equated by wrapping the images in blank data (greyscale value of 0), referencing the spatial metadata of the images to retain the relative subject position across the imaging sequences for each subject. Once stacked, the muscles visualised within the lower limb T1-weighted scans were semi-automatically segmented using Materialise Mimics [129].

To train the CNNs used, a segmentation label image must be created. The output of the manual segmentation process for each muscle was a stereolithography (STL) file, and each was required to be transformed into volumetric images (a two-dimensional example is shown below in Figure 6.1). This process was completed muscle-by-muscle, beginning with the STLs aligned with the respective locations in the label image, referencing the origin of the image segmented in the semi-automatic process. All pixels within the label image that intersect a triangular element of a muscle STL element were highlighted, and the area was then filled. For visualising the different muscles in the same image, the outputted binary image for each muscle was then assigned a greyscale colour (1-37), ordered alphabetically with respect to the name of the muscle. For example, the biceps femoris muscles are located in the lateral section of Figure 6.1(C) and have very low greyscale value (4 and 5, respectively), while the three vastii, located in the anterior section of Figure 6.1(C), have high greyscale value (35 – 37). The CNNs were trained with all 37 muscles, ensuring that predicted muscle tissue is more realistically distributed.

In addition to the segmented subject imaging data, manually checked augmented images produced from deformable image registration (Chapter 5) were also used to train the CNNs. As the number of subjects enrolled in the study (n=11) was significantly lower than those in other studies (range 21-64) [80, 87], the augmented images reduced the impact that this limitation would have on the accuracy of the segmentation.

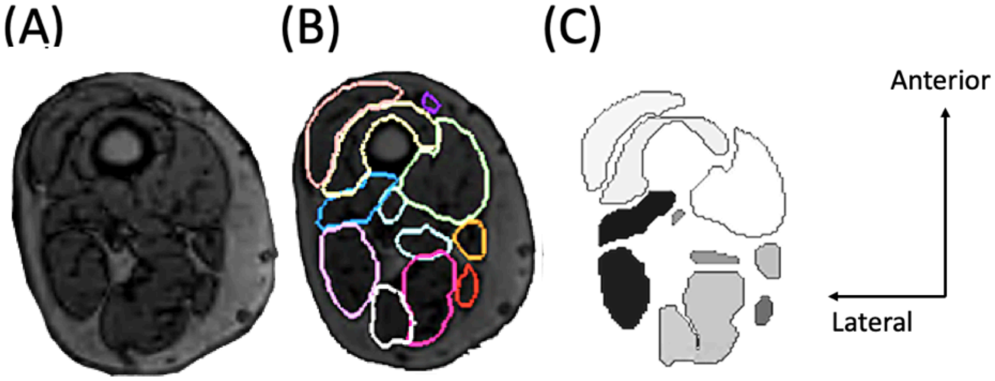


Figure 6.1: The process of generating the inputs for the neural network. The image inputted into the neural network (A), was manually segmented (B) and the segmentations were transformed into a label image (C). The greyscale colours of each muscle were ordered from 1-37 as they appear alphabetically.

6.2.2. Convolutional neural networks

Three Convolutional Neural Networks (CNNs) were used: the UNet [114], the "Attention UNet" [148], and an in-house model, named "Spatial channel UNet". The first two network architectures are well-known, state of the art architectures that have been extensively used throughout the field of medical image analysis. Attention UNet has not yet been tested in the context of muscle segmentation, hence its inclusion in this study. In the following sections, the structure and details of these network architectures are explained.

6.2.2.1. UNet

The UNet architecture presented in Figure 6.2 is a 2D multi-layered deep CNN with encoding and decoding pathways. Within the UNet architecture, sequential convolution operations are applied firstly to the input image and thereafter to feature maps: the matrices that enable feature extraction [114].

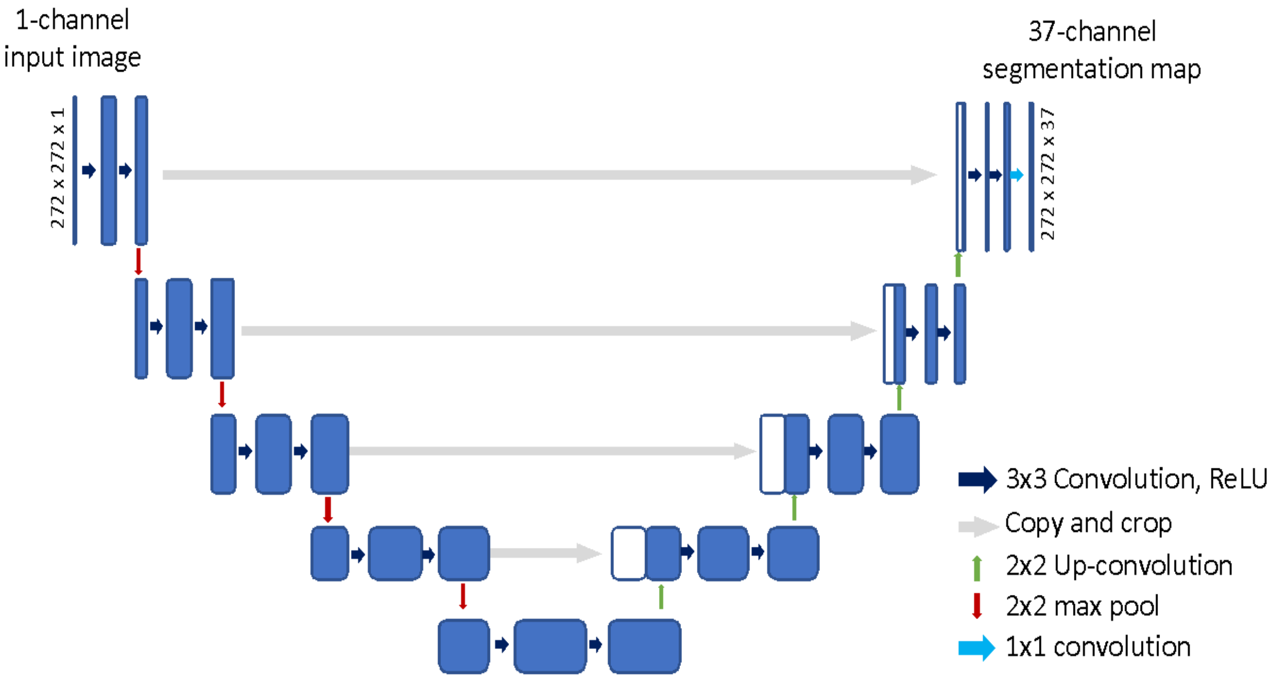


Figure 6.2: The UNet architecture [114]. Horizontal stages of the flowchart denote application of a 3x3 convolution and Rectified Linear Unit ($\mathrm{ReLU}(x < 0) = 0, \mathrm{ReLU}(x \geq 0) = x$) to a feature map. Downward and upward stages represent down-sampling and up-sampling respectively, adjusting the size of the feature maps. The copy and crop stages combine past and present feature maps to highlight previous information. The final convolution adjusts the final feature map to have the required number of output channels, in this case n=37.

The contractile, encoding path has four stages, each including two sequential convolution layers, with convolution kernels not user imposed, followed by a pooling layer, which reduces the dimensionality of the feature map by a factor of two. The number of convolution operations, and therefore, the number of features channels extracted in each convolution layer followed those imposed by the original UNet architecture [114], being: 64, 128, 256, 512, and 1024 at each respective stage of the encoding path. Max pooling (with a 2x2 kernel) was used in this study, as it has been proven to be an efficient and effective method of pooling in medical image segmentation problems, such as cell segmentation [114], or brain tissue segmentation [113]. Max pooling of a feature map passes the maximum element in each (2x2) block of the feature map to a downsampled feature map. The 2x2 kernel used in this study is applied with stride 2, reducing the size of the downsampled feature map to be half the size of the inputted feature map. The contractile, encoder path focuses on what is in the images, assessing them in a global fashion. This process is applied 4 times in series meaning many millions of combinations of convolutions are used to form millions of downsampled feature maps. For this reason, batch normalisation was used between each convolution block, normalising the weights assigned to each feature map across each mini-batch passing through each stage of the network. The inclusion of batch normalization increases the time taken to complete each epoch of training but overall reduces the number of epochs required to train a functional model.

The expanding, decoding path was built in the same manner as the encoder path, with the max pooling operations exchanged to an ‘up-convolution’ block. The up-convolutions halve the number of feature channels and double the size of the feature map through up-sampling, padding each element with zeros. Similar to the contracting path, each of the four stages of the expanding path consist of one up-sampling block, followed by two convolution blocks. Additionally, feature maps from the contractile path with the same size are copied and concatenated (in a layer named skip-connections) after the up-convolution is applied, coupling the features found within the contractile path with the expanding path. The design of the expanding path is focused on discovering where in the image the relevant features are. When coupled with the contracting path, the UNet is capable of learning what and where it is looking in the images, whether it is during the training process, or indeed in the testing phase.

The final stage of the UNet architecture occurs once the feature map is up-sampled back to the size of the inputted image, after four down-sampling and four up-sampling stages. The final stage is a convolution layer where the number of feature channels was equal to the number of classes required to be segmented, in this case equal to 37. The output was a probability map defined for each pixel in the input image, which was used to create a segmentation prediction, finding pixel-wise the class of greatest probability. The details for the training, validation, and testing protocol are outlined after each model structure is defined.

6.2.2.2. Attention UNet

The second model employed an attention module, which was built directly into the traditional UNet architecture, following the theory proposed by Ronneberger et al. [114]. The attention encoder allows the spatial location of the region of interest to be retained. The attention module adds together the feature maps from the down-sampling and up-sampling stages, through multiplication of the two and normalization. In this process, the important feature channels (those features with the greatest weights) from both the down-sampled and up-sampled are accentuated, and conversely those feature channels that are not important are de-emphasized. The attention gates shown in Figure 6.3, consist of: 1) a simple matrix addition, 2) a rectified linear unit, forcing feature channels of negative weights to be equal to 0, 3) a sigmoid function, to squeeze the weights to be in the range [0,1], and finally 4) a resampling stage, to retain the correct feature map and feature channel sizes.

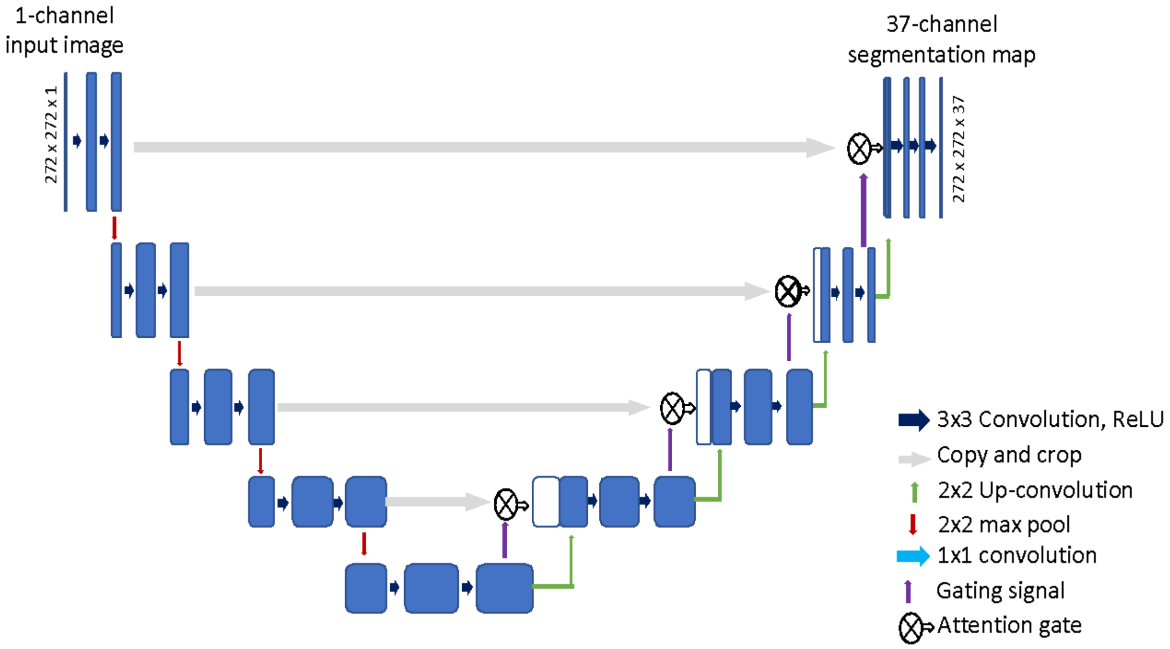


Figure 6.3: The attention UNet [148]. The signals cast between the encoding and decoding sides of the UNet are cast together in an attention gate, altering the concatenation stage of the up-sampling.

6.2.2.3. Spatial channel UNet

The third and final model assessed in this study was the in-house Spatial channel UNet. The model followed the CNN operations of the traditional UNet, with an additional Spatial channel running in parallel, designed to enrich the network's knowledge of where the imaging slice was acquired from within the body. This method builds an understanding into the network of what muscles could and should be assessed within each image. For example, thigh muscles, such as the vastii, cannot be visualised within an imaging slice taken from the calf, and the addition of the Spatial channel is designed to enhance this.

The architecture of the Spatial channel UNet, seen in Figure 6.4, maintained the original UNet, with a fully connected linear layer running in parallel [148]. The model required a measure of the position along the lower limb each image was acquired from. The percentage along the lower limb (p) that each image was acquired was used to provide this information, referencing the image number and the total number of images within each sequence. This spatial reference was tied to each image and both were inputted into the network model. The images ran through the UNet portion of the model and the spatial reference ran through the fully connected linear layer. The fully connected layer had 100 input neurons representing each percentage along the lower limb, and 37 output neurons representing the number of classes (muscles) being segmented from the images, with all input and output neurons being connected. The percentage along the lower limb (p) was converted into a 100x1 binary matrix, with the p^{th} element set equal to 1 and all other elements equal to 0. The network was trained to strengthen connections between a given position along the lower limb and the muscles that could be contained in images acquired from those positions, and conversely, weaken the connections for those muscles that could not. The output of the Spatial channel was a probability matrix, each muscle being within a certain image, which was then multiplied with the result of the final convolutional layer in the UNet structure, influencing the segmentation prediction.

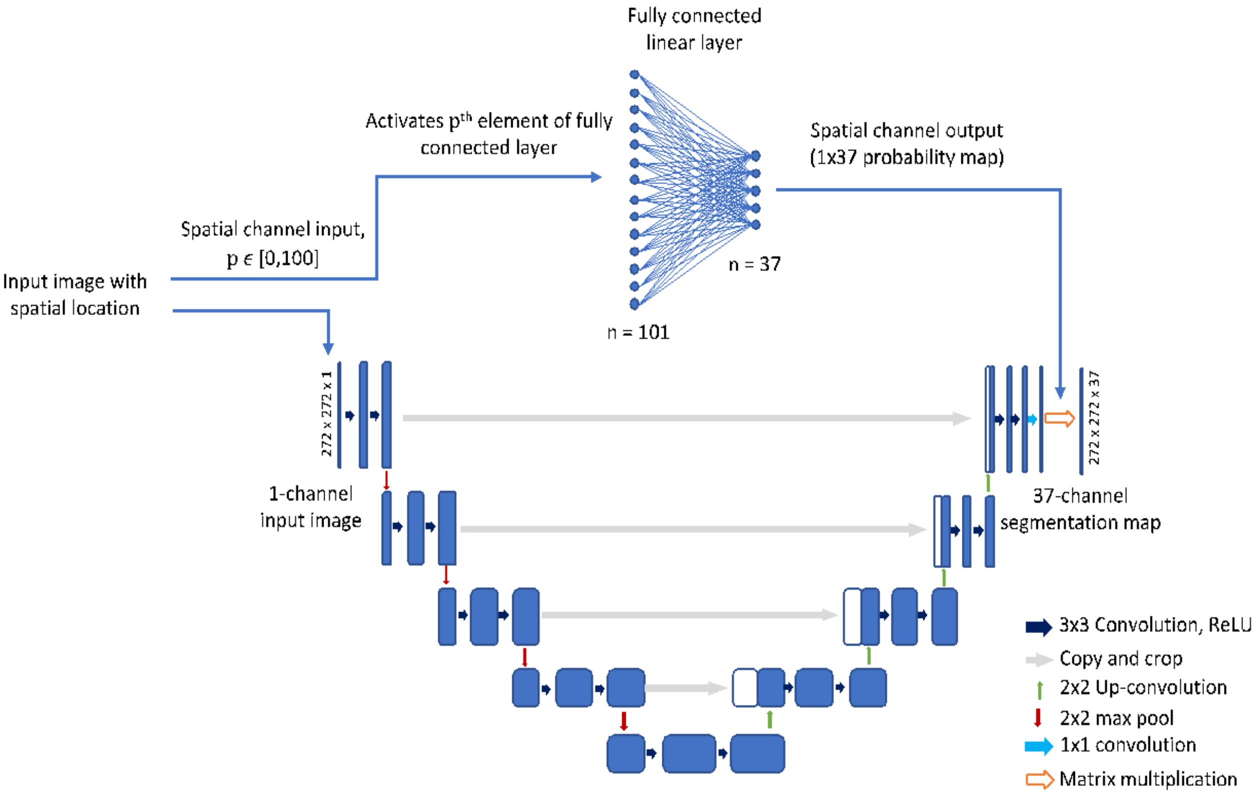

Fully connected linear layer
Activates p^th element of fully connected layer
Spatial channel output (1x37 probability map)
Spatial channel input, p ϵ [0,100]
Input image with spatial location
n = 37
n = 101
272 x 272 x 1
272 x 272 x 37
1-channel input image
37-channel segmentation map
3x3 Convolution, ReLU
Copy and crop
2x2 Up-convolution
2x2 max pool
1x1 convolution
Matrix multiplication

Figure 6.4: The Spatial channel UNet. The spatial location of the input image is calculated. Spatial information and the image are input into the network simultaneously, where they are split instantly. The image data goes through the standard UNet architecture (see Figure 6.2), while the spatial channel takes an integer, $p \in [0,100]$, and activates the pth node of the input to the fully connected linear layer. Each input node is connected to each output node, of which there are 37 (equal to the number of muscles), allowing the locations of the muscles to be learned along the longitudinal axis.

6.2.3. Training method

All three models were trained using identical methods, on a NVIDIA® GeForce RTX™ 3060 Ti Graphics Processing Unit (GPU) with 8GB of memory. A stochastic gradient descent method was used, separating the training database into batches (of size 8 due to memory card size), employing the Adam optimizing algorithm to iteratively update network weights during back-propagation. The learning rate was set at 0.001 and decreased by a factor of 0.5 after 30 epochs of training. In total, networks were trained for 90 epochs. All models were trained with the multi-class cross entropy loss function [149]. Other hyper-parameters were tuned empirically for the optimal network performance, whilst ensuring the GPU memory and capacity were not exceeded.

Throughout all network training an 80:20 split between training and validation data was used. The training data was used to fit the models to the lower limb MR images, and the validation data to verify that the algorithm was learning as intended in response to the training method selected. Due to the small number of subjects in the study cohort, a 'leave one out' testing system was used.

6.2.4. Experiments

Three experimental procedures were followed. The first found which of the three networks was able to automatically segment the muscles with the highest accuracy. The second assessed the effect of including the augmented images generated through deformable image registration [81]. The third and final experiment used the optimal network architecture and training database found through the first two experiments, and retrained that network five times independently, leaving each of the five subjects segmented in the previous work out as the testing dataset. The third experiment was used to enable a direct comparison between automatic segmentation methods.

6.2.4.1. Comparisons among models

The three outlined models were all trained following a 'leave one out' approach, testing one subject that was chosen at random. The labelled imaging data of ten subjects were used to train (eight subjects used for training, two for validation) each of the networks. The trained network was then used to segment the testing subject, allowing the segmentation accuracy to be assessed. The same three error metrics used in the previous chapters were used to measure the performance of each model (see Section 4.2.2.3). Though 37 classes, or muscles were segmented by each network, only the results for the 23 muscles identified in Section 3.2.3.3. were analysed, removing unreliable training data from the interpretation of the results.

6.2.4.2. Incorporation of augmented data

All three models were retrained, after including the augmented images generated through image registration into the training database. Fifty-two labelled datasets were used in the training database, and 16 were used in the validation database. Those augmented images created from the registration of images involving the one testing subject, whether it be as the fixed or target subject within the registration (see Section 5.2.5 & 5.3.4), were removed from the training database.

6.2.4.3. Varying testing dataset

The third and final experiment used only the best performing model, which was retrained for five further subjects, being those five subjects segmented in the previous studies (Chapters 4 and 5). The training database was adapted before the training phase for each of the tests, such that the augmented images related to the testing dataset were removed. The distribution of datasets within the training and validation data for each of the experiments is presented in Table 6.1. This experiment was performed for two reasons: 1) to ensure that the choice of testing dataset did not affect the results, and 2) to allow direct comparison between segmentation methods across a number of subjects.

Datasets	Initial test	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Total	66	65	64	64	67	64
Training	52	52	51	51	53	51
Validation	16	15	15	15	16	15

Table 6.1: Number of labelled datasets (each dataset contains approximately 1000 images) within the training and validation databases for the initial testing subject, and the other 5 subjects for whom the best performing network was retrained.

6.2.5. Evaluation metrics

The evaluation metrics used to quantify the accuracy of the segmentations outputted from each of the networks were the Relative Volume Error (RVE, %), the Dice Similarity Coefficient (DSC), and Hausdorff Distance (HD, mm). These complementary error metrics (calculations outlined in Section 4.2.2.3.) were found through comparison of automatic and manual segmentations.

Each model was checked for convergence through analysis of the training and validation loss curves.

6.2.6. Statistics

Firstly, to compare the initial results (without augmented images in the training database) from each of the models, the distributions of RVE, DSC, and HD found for the 23 muscles segmented in the testing subject by each model were tested for normality using a Kolmogorov-Smirnov test, concluding that they were not normally distributed. Thereafter, the results from the three different models were tested for statistically significant differences using a Wilcoxon signed-rank test.

After retraining the networks with the augmented images, two statistical tests were used. First, comparing the results with and without the augmented images included in the training for each network structure independently, to analyse the effect of incorporating the augmented imaging dataset into the training phase. To do so, a Wilcoxon signed-rank test was used, after a Kolmogorov-Smirnov test concluded that the error metrics were not normally distributed. Secondly, the segmentation accuracies found for the three different models after being retrained with the augmented images were compared. Similarly, Wilcoxon signed-rank tests were used to compare the RVE, DSC, and HD found using the three different networks.

Finally, the error metrics found after retraining the best performing network for the five retested subjects were assessed for normality again using a Kolmogorov-Smirnov test. After failing this test, the RVE, DSC, and HD found for the five retested subjects were statistically tested against those found for the initial testing subject using a Kruskal-Wallis non-parametric Analysis of Variance (ANOVA).

6.3. Results

6.3.1. Initial model evaluation

6.3.1.1. Training convergence

All models had converged at the selected number of training epochs and each of the models were found to be a good fit to the training and validation data. Figure 6.5 shows that for all three models, the training loss had converged at or before the 90th epoch. Moreover, the validation loss was unchanged from the 30th epoch, until the 90th meaning that the parameters found for the models at the 90th epoch were valid to use for the testing dataset. The RVE, DSC, and HD calculated for each segmented muscle (23 muscles per subject) in the testing subject are reported in Figure 6.6.

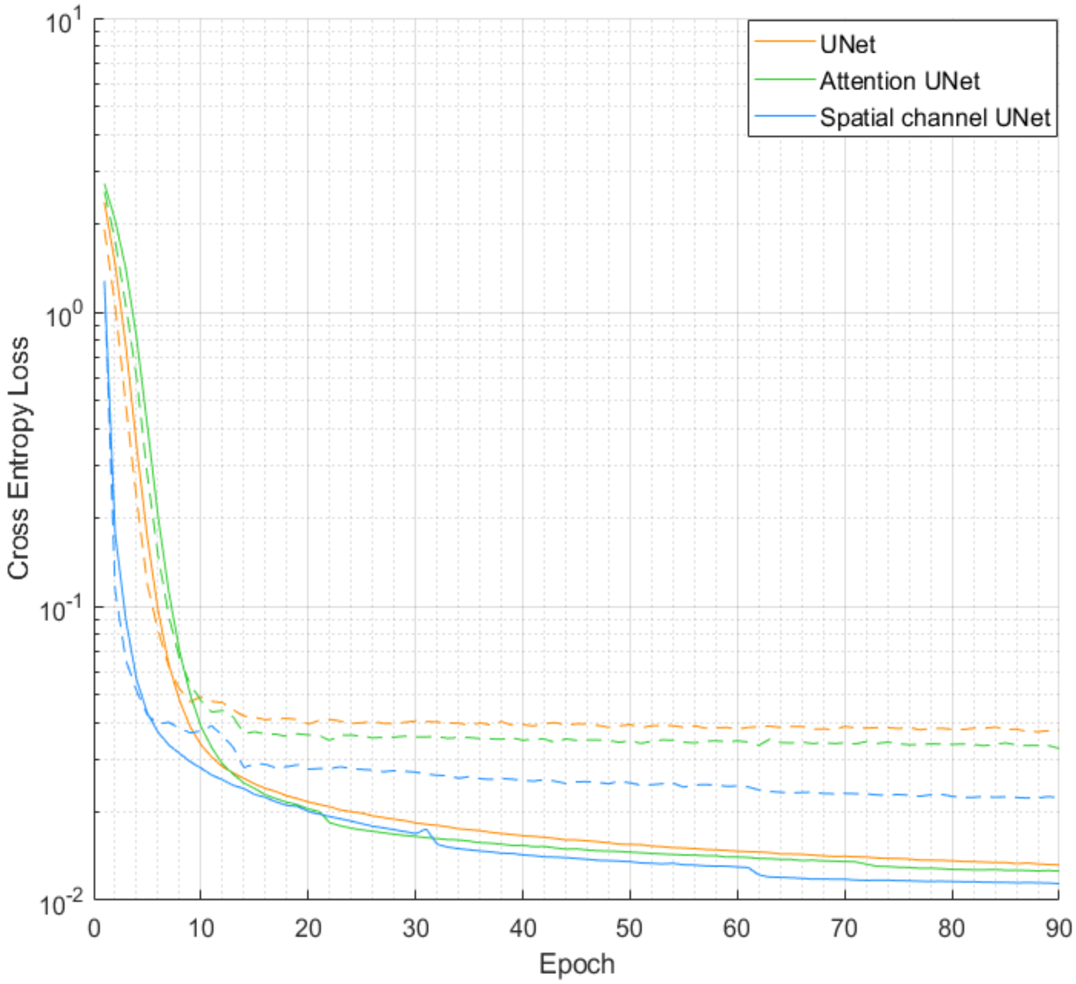


Figure 6.5: Training (solid) and validation (dashed) loss curves calculated throughout the training phase for the UNet (orange), Attention UNet (green), and Spatial channel UNet (blue). The cross entropy loss shown was calculated for each batch training or validation data and averaged across each epoch.

6.3.1.2. Number of parameters and training time

The details regarding number of parameters, time for each iteration, and overall training time are all presented in Table 6.2. The UNet and Spatial Channel UNet had very similar numbers of parameters, while the Attention UNet had a much larger number. The training time reflected the number of parameters with the training time of the UNet and Spatial channel UNet being comparable, but the Attention UNet required more time on average to train. The time to segment the testing dataset was less than 30 seconds for all models.

	UNet	Attention UNet	Spatial channel UNet
Number of parameters	1745920	2444288	1749620
Average batches per second (it/s)	6.38	4.34	6.19
Total training time (hrs)	4.14	6.31	4.38

Table 6.2: Number of parameters, average number of batches trained per second, and total training time for the three models tested.

6.3.1.3. Initial segmentation accuracy

For the 23 muscles segmented, the RVE was comparable among each of the three networks, with no statistically significant difference found between the results (e.g. $p > 0.544$, comparison of UNet and Attention UNet), as shown in Figure 6.6. Explicitly, the mean (± standard deviation) RVEs found for each of the three networks were -2.1% ± 19.3% (UNet), -3.1% ± 21.4% (Attention UNet), and -3.5% ± 18.2% (Spatial channel UNet), with similar interquartile ranges.

The DSCs, presented in Figure 6.6, for the Attention UNet (0.73 ± 0.06) and Spatial channel UNet (0.74 ± 0.06) were not significantly different ($p = 0.735$), whereas the DCS found for the UNet (0.80 ± 0.03) was significantly higher than the other models ($p < 0.0099$).

The HDs found for the 23 muscles are presented in Figure 6.6 for the three networks. For HD, no significant differences (e.g. $p > 0.225$, comparison of UNet and Spatial channel UNet) were found between the means of the UNet (29.4 mm ± 15.1 mm), Attention UNet (30.9 mm ± 19.0 mm), and Spatial channel UNet (28.1 mm ± 15.2 mm).

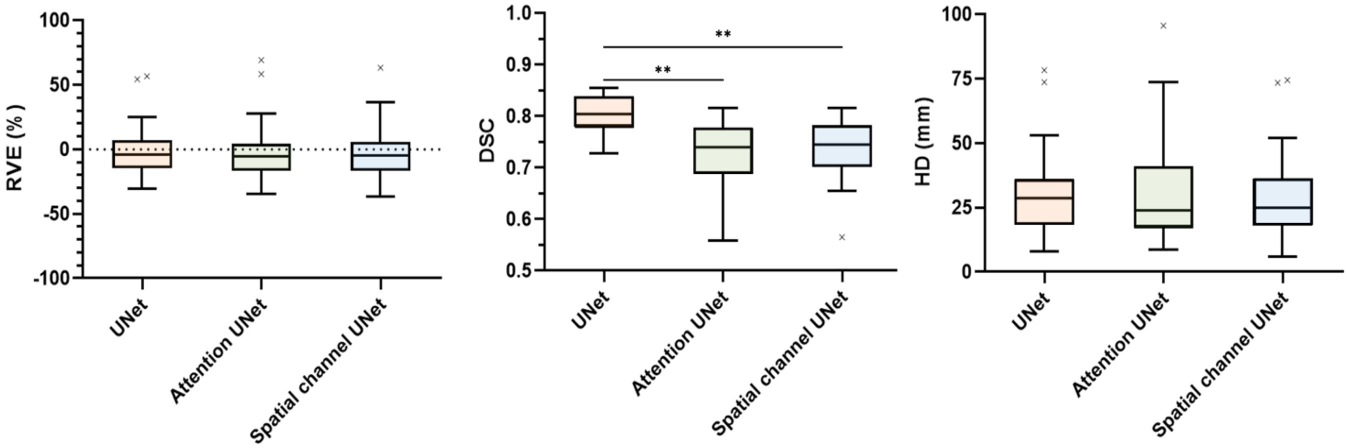


Figure 6.6: The RVE (%), DSC, and HD (mm) error metrics calculated across the three models tested. Lines above boxplots highlight significantly different results (* refers to p<0.05, ** refers to p<0.01).

6.3.2. Augmented data inclusion

Across all three error metrics, the 23 muscles of the test subject were consistently segmented more accurately after retraining the models with the augmented dataset included in the training database (Figure 6.7). The UNet improved the least after retraining, with no significant difference for RVE (p = 0.857) across the 23 muscles, but the upper and lower quartiles after inclusion of the augmented data fell within the acceptable level of operator variability (±10% [66]). Similarly, the Attention UNet showed no significant improvement in RVE between augmented and not augmented training datasets (p = 0.074). For Spatial channel UNet there was a small significant reduction of RVE after training with the augmented dataset (from -3.5% to -0.5% RVE, difference in mean 3%, p = 0.024).

Considering DSC, all three models were more accurate with the inclusion of the augmented data, with significant improvements for both the Attention UNet (DSC from 0.73 ± 0.06 to 0.79 ± 0.06, an 8.2% increase, p = 0.0074) and the Spatial channel UNet (mean DSC from 0.74 ± 0.06 to 0.81 ± 0.05, a 9.5% increase, p = 3.5×10^{-5}). The muscles segmented with the lowest DSC, (gracilis, adductor brevis and peroneus longus), were drastically improved, particularly in the Attention (from 0.56 to 0.72 for gracilis, from 0.66 to 0.80 for adductor brevis, from 0.66 to 0.77 for peroneus longus) and Spatial channel (from 0.57 to 0.72 for gracilis, from 0.65 to 0.86 for adductor brevis, from 0.67 to 0.76 for peroneus longus) UNet models.

The HD found with the inclusion of the augmented dataset for the original (HD from 29.5 mm $\pm$ 15.1 mm to 22.5 mm $\pm$ 11.2 mm, a 27.1% decrease, p = 7.5×10^{-5}) and Spatial channel (HD from 28.1 mm $\pm$ 15.2 mm to 22.8 mm $\pm$ 13.7 mm, an 18.9% decrease, p =

0.026) UNets were significantly lower, with the two being comparable in response to the updated training database.

Overall, the traditional UNet trained with the augmented images segmented the test subject with the greatest accuracy, with the consistently low RVE, high DSCs, and equivalent HDs compared to the other two models.

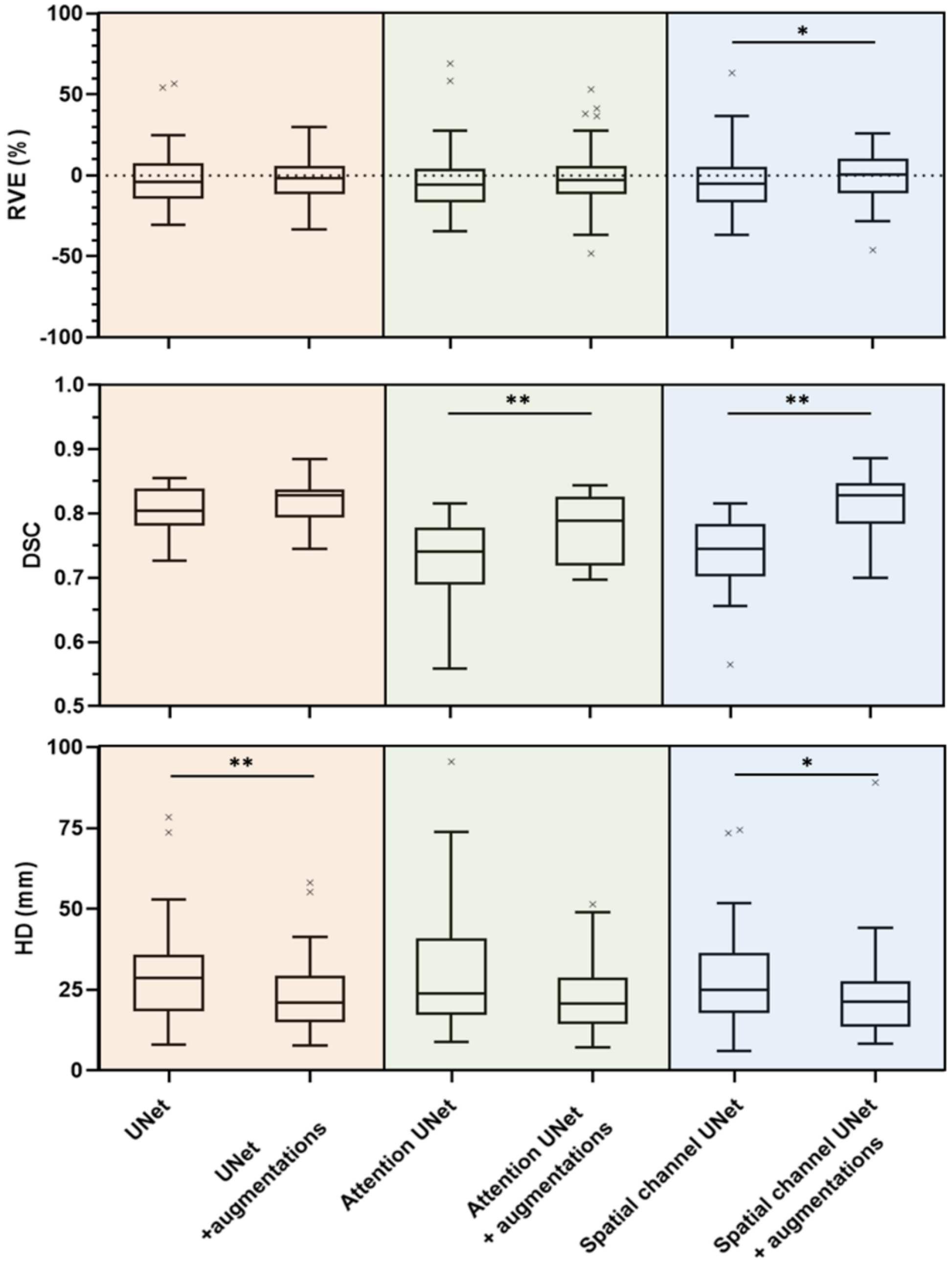


Figure 6.7: The RVE (%), DSC, and HD (mm) for the 23 muscles of the test subject calculated using each of the three models tested, both with and without augmented images included in the training phase. The lines linking the boxplots highlight significantly improved results after inclusion of the augmented images (* refers to p<0.05, ** refers to p<0.01).

6.3.3. Selection of the best performing model

The RVE, DSC, and HD found through automatic segmentation of the testing subject, using the UNet, Attention UNet, and Spatial channel UNet models after training with the augmented images are presented in Figure 6.8. The RVE across all three models were comparable (UNet: 2.2% $\pm$ 14.6%, Attention UNet: -0.8% $\pm$ 19.9%, Spatial channel UNet: -0.5% $\pm$ 15.8%, mean $\pm$ standard deviation), with the smallest p-value = 0.11, found through comparison of the UNet and Spatial channel UNet. There were significant differences found in the DSC (UNet: 0.81 $\pm$ 0.04, Attention UNet: 0.78 $\pm$ 0.05, Spatial channel UNet: 0.81 $\pm$ 0.05). The mean of the UNet model was found to be significantly greater than that of the Attention UNet model (p = 5.9$\times 10^{-4}$). The same was true when comparing DSC between the Spatial channel UNet and the Attention UNet (p = 0.011). Finally, the HD found across all three models were comparable, and no significant differences were observed (UNet: 22.5 mm $\pm$ 11.2 mm, Attention UNet: 23.1 mm $\pm$ 11.1 mm, Spatial channel UNet: 22.8 mm $\pm$ 13.7 mm, minimum p = 0.36 found comparing UNet and Spatial channel UNet). As the UNet produced a comparable segmentation accuracy to the others in terms of RVE, greater than the others for DSC (with a greater lower quartile than the Spatial channel UNet, Figure 6.8), and comparable HD, this model was retrained for other testing subjects.

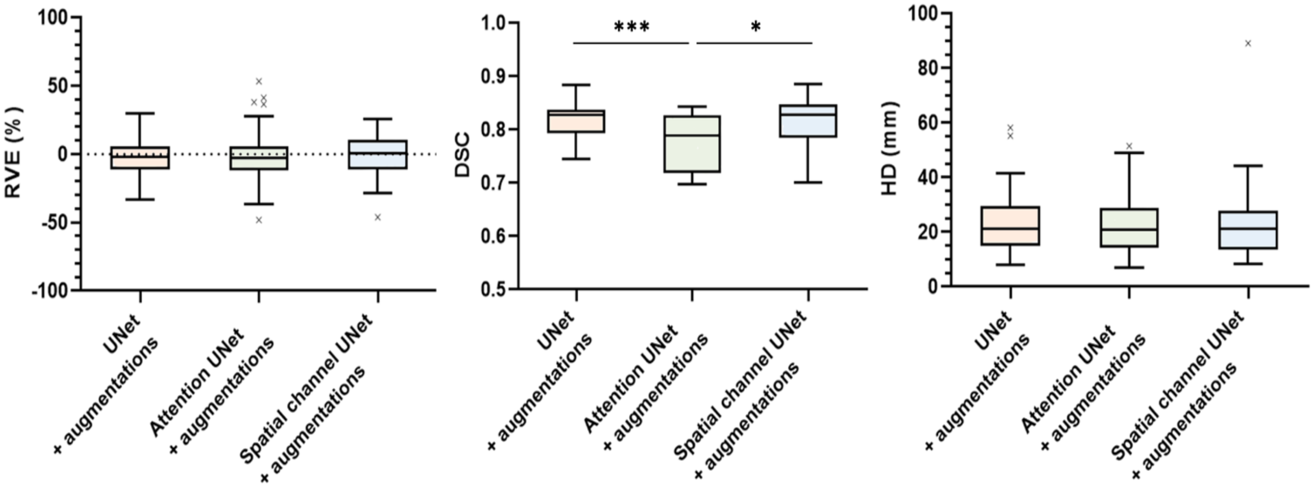


Figure 6.8: The RVE (%), DSC, and HD (mm), found when segmenting the one testing subject using the UNet, Attention UNet, and Spatial channel UNet after training with the augmented images included. The mean RVE, DSC, and HD found for each of the three different segmentation models, were all statistically analysed. * above a line connecting two bar charts signifies p < 0.05, *** signifies p <0.001.

6.3.4. Retraining for different testing subjects

The UNet with the augmented images was retrained, exchanging the testing subject to each of the 5 subjects that were segmented in the previous studies and the results are shown in Figure 6.9.

Overall, the results for each of the five different subjects were similar to the results from the initial testing subject. No statistically significant differences were found between the RVE (testing subject: 2.2% ± 14.6%, Subject 1: -8.6% ± 18.7%, Subject 2: -0.9% ± 4.2%, Subject 3: -1.5% ± 12.1%, Subject 4: 2.9% ± 9.0%, Subject 5: -0.6% ± 10.5%, mean ± standard deviation, minimum p = 0.15 between testing subject and Subject 1). Subjects 2 and 4 were segmented with a significantly higher DSC (0.02 difference, p = 0.044 for test subject vs Subject 2; 0.02 difference, p = 0.026 for test subject vs Subject 4) compared to the initial testing subject, but Subject 3 was segmented with a significantly lower DSC (-0.05 difference, p-value= 0.0039). Only Subject 5 was segmented with a significantly lower HD than the initial test (-10.0 mm difference, p = 0.033), the other four retested subjects were consistent with the initial test (p > 0.32).

All six subjects were segmented with a mean RVE within the ±10% that would be acceptable with respect to an operator dependency study [66] (between -8.7% to 3.7%). As shown in Table 6.3, the average RVE across all 23 muscles was all within the acceptable operator dependency level. Only Subject 1 was found to be significantly different to any other within the six tested subjects. The mean DSCs across all six tested subjects were in the range (0.76, 0.84), with Subject 3 being segmented with the lowest DSC and the other five subjects being segmented with a DSC > 0.81, and this difference in mean DSC was significantly lower than 4/5 of the other subjects. All muscles, other than the gracilis, were segmented with a mean DSC above 0.75, when averaged across the six tested subjects (Table 6.3). The HD again was very consistent across the cohort, with most subjects being segmented with a mean HD within the range (20.6, 26.7 mm), with only Subject 5 being segmented with a significantly lower average HD (12.9 mm) than the other five subjects.

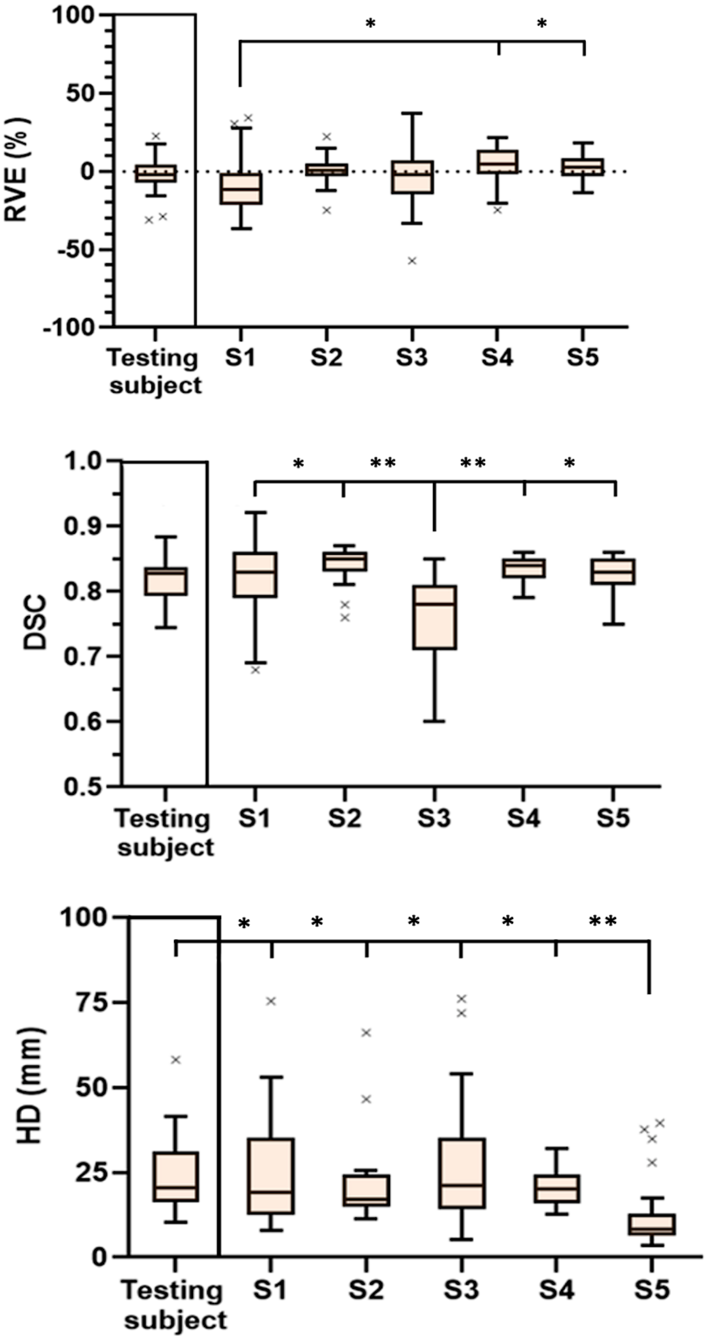


Figure 6.9: RVE (%), DSC, and HD (mm) calculated from the segmentations of 23 muscles from 6 different subjects: the initial testing subject, and the five subjects segmented in the previous chapters. The lines connecting the initial testing subject to the others highlight statistically significant differences between the distributions of results (* refers to p<0.05, ** refers to p<0.01). The longer line highlights the subject that was significantly different to the others.

Muscle	Relative volume error (%)	Dice similarity coefficient	Hausdorff distance (mm)
adductor brevis	-5.5 ± 13.6	0.79 ± 0.06	16.9 ± 7.1
adductor longus	-2.1 ± 6.4	0.82 ± 0.04	15.8 ± 6.1
adductor magnus	-1.5 ± 4.3	0.84 ± 0.02	21.7 ± 7.5
biceps femoris caput breve	1.3 ± 17.2	0.81 ± 0.03	13.1 ± 5.3
biceps femoris caput longum	-5.7 ± 5.0	0.83 ± 0.03	18.8 ± 12.0
gastrocnemius lateralis	-5.1 ± 14.7	0.82 ± 0.04	24.2 ± 13.8
gastrocnemius medialis	0.3 ± 4.4	0.84 ± 0.02	31.7 ± 21.2
gluteus maximus	-0.1 ± 5.8	0.85 ± 0.02	29.1 ± 23.6
gracilis	-7.3 ± 16.3	0.74 ± 0.07	27.9 ± 6.1
iliacus	-4.6 ± 5.0	0.82 ± 0.03	17.7 ± 4.9
peroneus brevis	-2.8 ± 11.9	0.79 ± 0.03	30.8 ± 7.7
peroneus longus	-0.3 ± 14.4	0.80 ± 0.04	16.2 ± 6.6
rectus femoris	-2.5 ± 15.5	0.84 ± 0.03	12.4 ± 6.4
sartorius	-3.3 ± 16.2	0.79 ± 0.03	28.1 ± 23.5
semimembranosus	-3.6 ± 7.1	0.84 ± 0.04	22.0 ± 19.8
semitendinosus	2.7 ± 10.6	0.84 ± 0.02	23.2 ± 10.2
soleus	4.7 ± 6.6	0.84 ± 0.06	17.8 ± 4.8
tensor fasciae latae	9.8 ± 16.4	0.77 ± 0.05	24.5 ± 14.5
tibialis anterior	-8.8 ± 15.2	0.80 ± 0.05	23.3 ± 10.9
tibialis posterior	1.1 ± 16.3	0.78 ± 0.04	20.5 ± 13.3
vastus intermedius	-0.5 ± 5.1	0.82 ± 0.04	22.0 ± 7.0
vastus lateralis	-2.5 ± 11.5	0.82 ± 0.03	29.1 ± 15.4
vastus medialis	-6.0 ± 4.6	0.84 ± 0.02	29.9 ± 15.3

Table 6.3: RVE (%), DSC, and HD (mm) found for the 23 muscles included in the automatic segmentation. Values within the table are the mean errors (± standard deviation) averaged across the 6 tested subjects.

The segmentations generated by each of the retrained networks are presented in Figure 6.10. Overall, the segmentations predicted by the model were visually very similar to the ground truth segmentations. Both the location and geometry of all 23 muscles assessed were well predicted. The gluteus maximus and adductor magnus (highlighted within Figure 6.10) were extremely well predicted, reflected by the high DSC found for this muscle, being above 0.85 across all subjects other than the adductor magnus within subject 3 (DSC = 0.81). However, there were some shortcomings in the predictions, particularly within the shank of Subject 3 (see light blue arrow pointing to tibialis anterior mislabelling, Figure 6.10) and thigh of Subject 1 (see purple arrow pointing to the vastus medialis mislabelling, Figure 6.10), where muscles were clearly mislabelled. The 2D cross sections appeared to show that the predictions (automatic segmentation) and references (manual segmentation) were in good agreement. The three error metrics suggested that there were some areas of disagreement. These disagreements were relatively heterogeneous between the different muscles considered, as shown in Table 6.3.

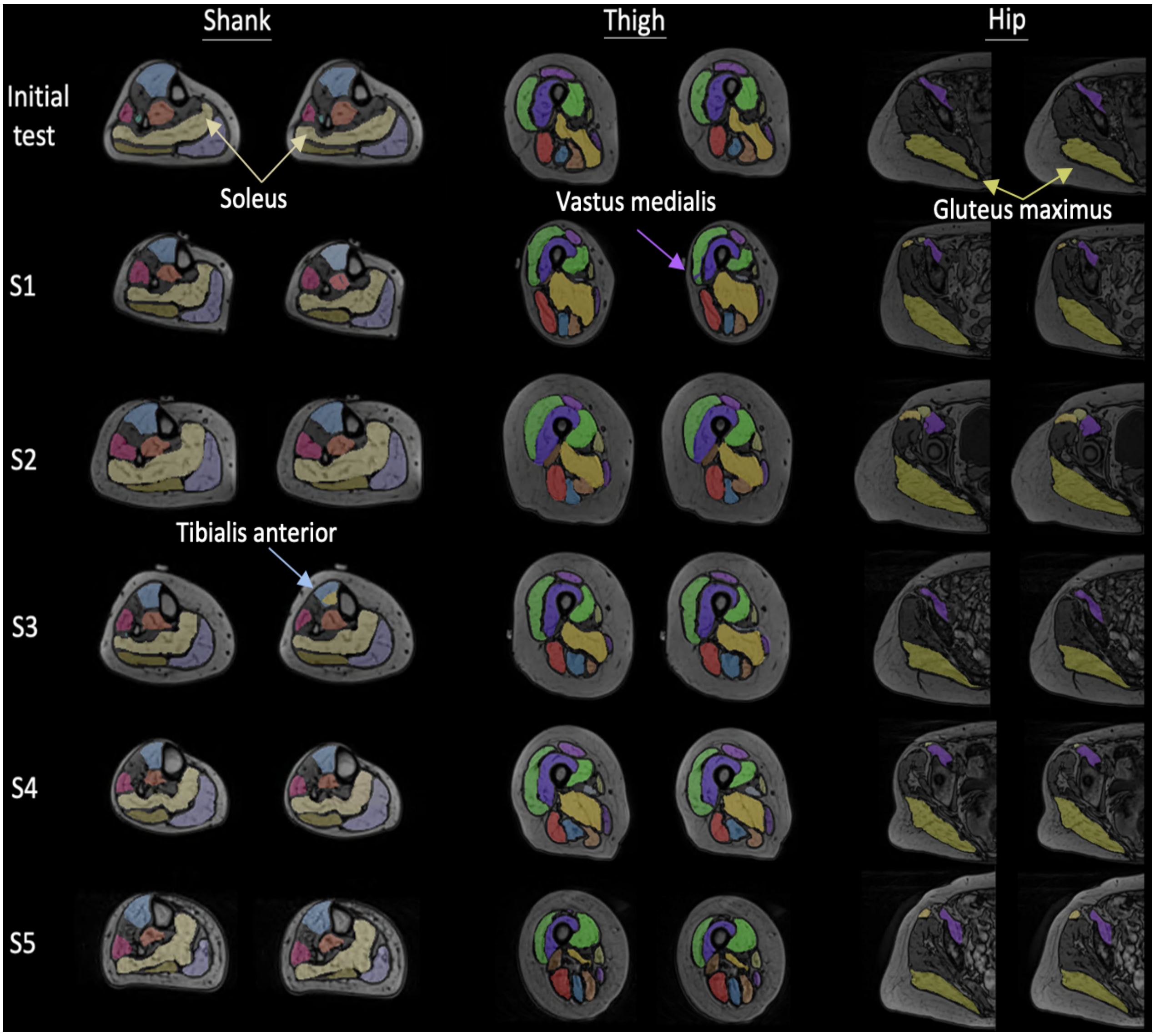


Figure 6.10: Visual representation of ground truth and automatic segmentations outputted from the best performing model, the UNet. Three slices taken from halfway along the shank, thigh, and hip sections are presented, with the manual and automatic segmentations shown on the left and right, respectively. Each row of images corresponds to each of the six segmented subjects. The highlighted muscles are referred to in Section 6.3.4 above, pointing to the soleus (light yellow), gluteus maximus (yellow), vastus medialis (purple), and tibialis anterior (light blue).

6.3.5. Direct comparison between segmentation methods

Five subjects were segmented using all four automatic approaches used within the thesis: left to right segmentation, inter-subject single atlas segmentation and multi-atlas segmentation, and the traditional UNet trained with augmented images. Therefore, a direct comparison is drawn between the Relative Volume Error (RVE, %), Dice Similarity Coefficient (DSC) [131], and Hausdorff Distance (HD, mm) [132] found using the four different methods for the five subjects (Figure 6.11). The segmentation of the right limb through registration of manually segmented imaging data of the left limb gave the most accurate segmentation in 4 out of 5 subjects considering both RVE and HD. In the circumstances where the left to right registration was not the best performing segmentation method in terms of RVE and HD, the UNet was. In all but 1 subject, the UNet was able to segment the 23 muscles with the highest DSC, with the left to right registration being the best performing segmentation method in that case. The optimal single atlas and multi-atlas were the 3rd or last placed segmentation methods across all three-error metrics in all but Subject 1.

Finally, the time requirements for each of the segmentations differed widely. The left to right and single atlas inter-subject segmentation methods required between 1 and 1.5 hours to segment one limb from the subjects, the multi-atlas approach required between 4 and 5 hours (required 4 single atlas registration iterations), and the UNet required around 16 hours to be trained but performed the segmentation in less than 2 minutes.

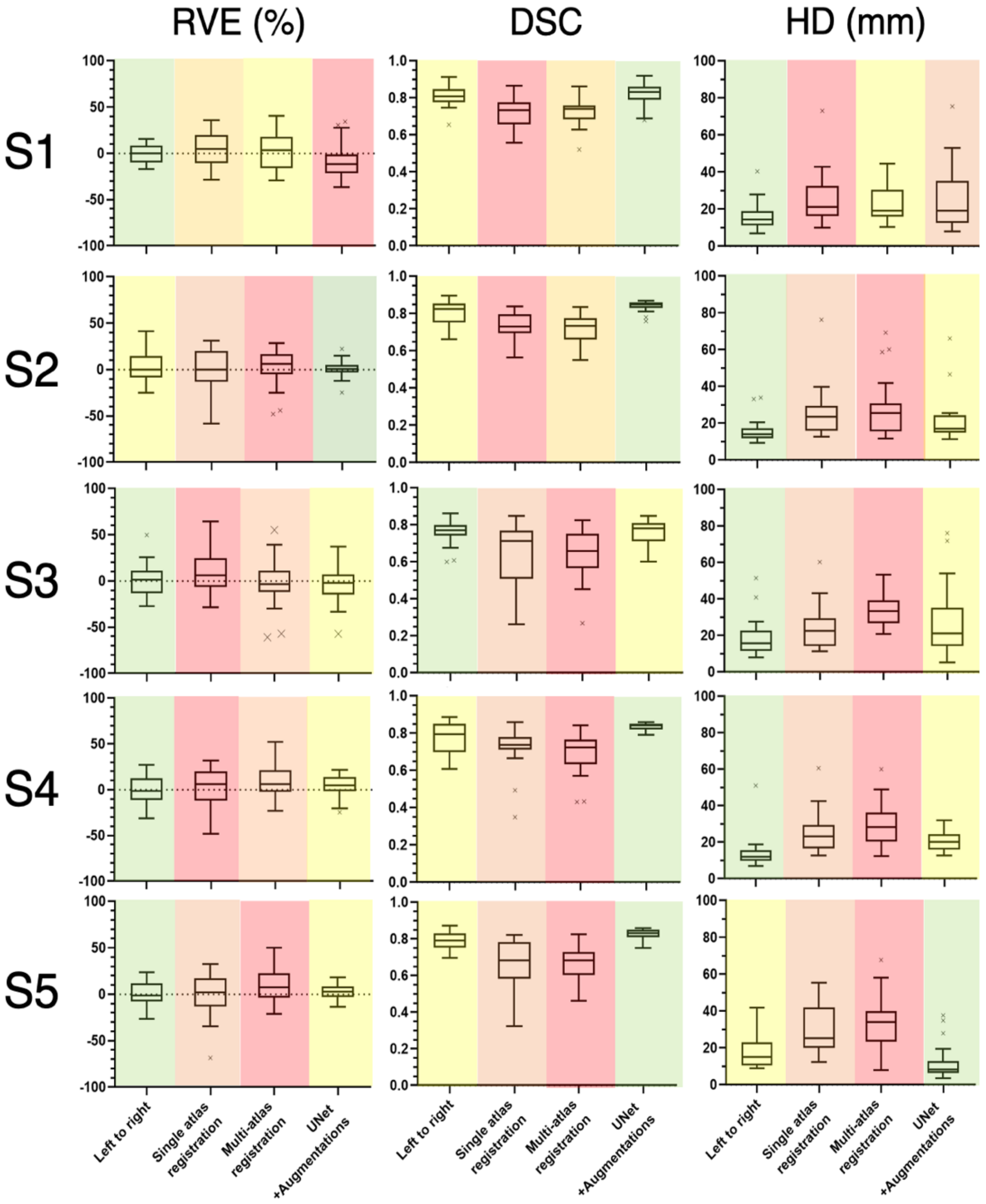


Figure 6.11: Comparison of the four segmentation methods used within this thesis. Each plot presents the segmentation accuracy of each method for a given subject (see row labels) and considering a certain error metric (see column label). The background colour surrounding each boxplot denoted the methods rank when compared to the other methods (the best method is coloured in green, 2nd in yellow, 3rd in orange, and 4th in red). The rank was determined as that with the greatest (DSC), or lowest (RVE, HD) mean. If the mean was within 1% in RVE, 0.01 in DSC, or 1mm in HD, then the method with the smallest inter-quartile range was given the higher rank.

6.4. Discussion

The chapter aimed to investigate the capacity of state-of-the-art CNNs to segment 23 major lower limb muscles from T1-weighted MR images and propose novel methods to overcome current computational limitations and data requirements noted within the literature [80, 82]. Furthermore, the benefit of using the previously generated augmented imaging database was analysed. Though similar studies do exist in the literature [80, 82, 87], these state-of-the-art models have not been tested to segment all major lower limb muscles simultaneously in a cohort with T1-weighted images. A model capable of automatically segmenting all major muscles from lower limb T1-weighted MR images whilst matching a human operator level of accuracy would be of great interest to the research community, as new subjects could be processed without the need of retraining new or existing models.

Across three error metrics: RVE, DSC, and HD, the 23 muscles were segmented with a moderate to high level of accuracy by all three models tested. All models were shown to have converged through assessment of the training and validation loss curves, suggesting that the models were fully trained and that the training and backpropagation algorithms used were valid. Both the traditional UNet [114] and Spatial channel UNet were able to segment the muscles well within the subject initially chosen for testing. Upon inclusion of the augmented imaging database into the training phase, these networks could segment the muscles with low average RVE (UNet = -2.1%, Spatial channel UNet = 1.0%) and a moderately high DSC (UNet = 0.81, Spatial channel UNet = 0.81). The Attention UNet [148] was also retrained with the augmented imaging database and though the RVE was comparable to the other models, this model showed a lower average DSC of 0.79. The Attention module has been used very effectively in the segmentation of other tissues but has not been used widely in multi-classification segmentation tasks [148, 150]. Therefore, the Attention UNet may have been negatively affected by the large number of classes (or muscles) in this task, given that this network was outperformed by the traditional UNet model. Further changes could be made to this network, such as incorporating a multi-headed attention module [151], but this would incur vast numbers of additional parameters, greatly increasing training time.

The use of augmented imaging data, particularly within a relatively small study cohort (n = 11) benefitted all three of the models tested, with significantly improved segmentation accuracy in at least one of the three error metrics. The DSC of the segmentations predicted by the Attention UNet model were improved but still did not match the level achieved by the other two models. The Spatial channel UNet benefitted the most with the inclusion of the augmented database in the training algorithm but achieved a comparable level of accuracy to the traditional UNet. The spatial channel method was designed to allow the network to reduce the number of muscles that could be present within each image slice. However, the disparity in the DSC found between the UNet and Spatial channel UNet when trained without the augmented database suggested that the spatial channel running in parallel with the UNet limited the ability of the network to segment the muscles. The high RVE and HD suggested that the extremities of the muscles were not well captured. For example, the testing subject may have a muscle present within a certain imaging slice at a given percentage along the lower limb, whereas this muscle might not have been present in the other subjects at the same location. The Spatial channel would therefore prevent the inclusion of this muscle in the prediction for that certain image, which would lead to an incorrect segmentation. When retrained with additional augmented images, which were shown to increase the variability of muscle structure when comparing the augmented and original databases [81] (Section 5.3.4), the effect of this limitation was reduced. The Spatial channel UNet and traditional UNet produced segmentations of similar accuracy, suggesting that the spatial channel did not enhance the learning process for the models trained in this work, which solidified the choice of model used for further testing. The increase in segmentation accuracy found using the Spatial channel UNet when retrained with the augmented images, suggested that with a more diverse training database (possibly with further augmentation), the segmentation accuracy could increase even further.

Traditional UNets have been used in the past to perform multi-class segmentation of the muscles from MR imaging data [80, 82, 87]. The study by Ding et al. [80] segmented two muscle groups and two individual muscles from water and fat suppressed images, inputting two images into the network. The two individual muscles (gracilis and sartorius) were segmented with a DSC of 0.86 on average, 6.1% greater than that in this study. However, in this study, 23 individual muscles were segmented with an average DSC of 0.82 across six subjects, which is comparable. This average segmentation accuracy was robust, with low standard deviation (± 0.08), and was approaching the criteria outlined in Section 3.4.4. The disparity between the results presented by Ding et al. and those found in this study could be due to their use of two MR imaging types as a multi-channel input, where in this study only T1-weighted images were used, bringing less information per training iteration to the network. Zhu et al. [87] used a similar approach, training many different network models to segment muscles from only the calf. Their group presented a method to segment the calf muscles with an average DSC of 0.89 using a hybrid 2D and 3D model. Comparatively, the muscles within the calf were segmented with an average DSC of 0.82 in the current study across the six tested subjects. They also tested the traditional UNet, which was able to segment the muscles with an average DSC of 0.87. Their study cohort consisted of young subjects both with and without cerebral palsy, in contrast to older post-menopausal subjects in this study. Not only did Zhu et al. [87] have a greater number of subjects (n = 20), enhancing the training of the networks, but the younger subjects used were likely to have a more homogenous muscle tissue appearance in MR images due to the effects of age-related degradation of skeletal muscle tissues [53] (see Figure 6.12 for comparison). The inhomogeneities of muscle tissues would typically lead to higher operator variability in the manual segmentation process, especially in cohorts with musculoskeletal disorders or older individuals [66]. Therefore, it was expected that the automatic segmentations presented in this study were less accurate than those reported by Zhu et al. [87]. Finally, another comparable study was that of Ni et al. [82], who segmented all lower limb muscles using a UNet architecture but exchanged the traditional 2D inputs for full 3D images. Ni et al. trained 35 individual networks, for a study cohort of 64 young healthy athletes (for difference in muscle structure see Figure 6.12), incurring extreme computational expense and training time. However, Ni et al. proposed a segmentation accuracy of around 0.9 in DSC, significantly higher than that found in this study or by Zhu et al. [87]. This again could be contributed to the higher homogeneity and definition of muscle appearance in the MRI in the young athletic cohort.

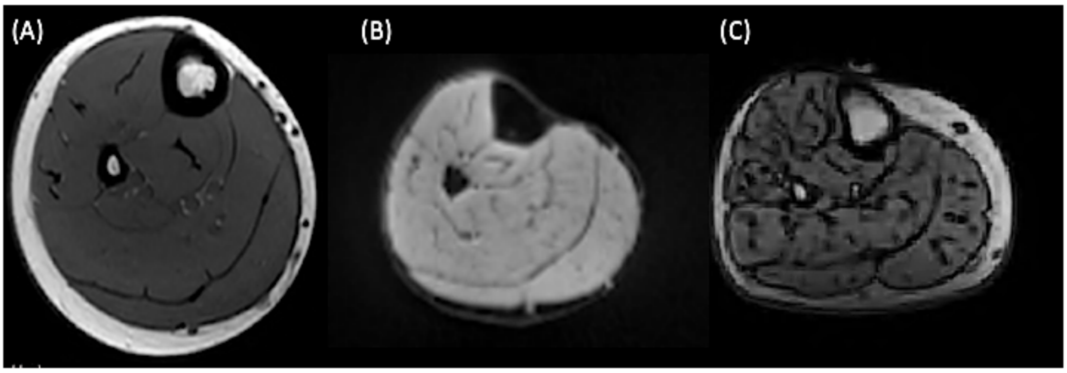


Figure 6.12: Example MR images taken from Zhu et al. [87] (A), Ni et al. [82] (B), and an example used in this study. Both those used in Zhu et al. [87], and this study were T1-weigthed images, where Ni et al. [82] used a fat suppression acquisition sequence.

After retraining the UNet model for the five subjects segmented using other methods, the model was mostly robust to the changes in the testing subject. Considering each of the three error metrics (Figure 6.9), there was one subject that was segmented with significantly lower accuracy than some of the others. The subject that was segmented with lower accuracy than the others was not consistent across the three error metrics, as these metrics represent errors found in different features within the segmentations. Overall, the UNet model was robust, suggesting that if tested for new subjects (not included in the study cohort), segmentations of similar accuracy might be produced automatically. The comparison between the different segmentation methods found that the left to right registration was the most accurate segmentation method considering the RVE and HD, as these metrics are volumetric, and surface based in their comparison between different segmentations. These two metrics are 3D in nature, well suited to the 3D registration method used in the left to right segmentation method, but not to the 2D UNet method. Moreover, the left to right registration was the best of the three deformable image registration-based segmentation methods. The variability in the muscle structure between opposing limbs is noted in the literature, but it is far smaller than the variability found between different subjects [66], leading to the left to right registration outperforming the other two registration-based approaches. The best reference that can be used to segment a certain target limb, is the opposing limb, and as these results highlight, even when using a multi-atlas approach. On the other hand, the UNet with a training database bolstered by augmented images generated through deformable image registration provided the best results in terms of DSC as these methods give predictions of the class of pixels in a pixelwise manner [106], which is well suited to DSC as an error metric.

The results presented in this chapter did have some associated limitations. Firstly, as is the case with many deep learning applications, the amount of data limited the training phase of all three tested networks. This is clear as the augmented images increased segmentation accuracy across all three tested networks. Additionally, the training phase required a significant amount of computational time, which could be boosted with the use of High-Performance Computing with multiple GPU cores. These were not used as one of the aims of the chapter was to foster these methods in situations where these systems may not available. However, the Spatial channel UNet required a very similar amount of time to the traditional UNet, suggesting that other methods that enrich the learning process with meta data could be adopted with little effect on computational time. Moreover, the 2D approaches used throughout this chapter could limit the learning process as the muscle structures segmented are 3D in reality. Three-dimensional applications of UNet-style structures have been tested in medical image analysis [123], but these approaches typically affect the time and computational requirements on the learning process. Specifically with large images such as these (~250x250x1000 pixels), a 3D application of the UNet would severely increase the computational expense incurred by the learning process. Additionally, the low number of full 3D labelled images (n = 11) could cause more bias in 3D UNet [123] and lead to reduced overall performance. Finally, the fact that 12 of the muscles were trained to be segmented from the imaging data but were not considered in the analysis of the results could have affected the segmentation accuracy outputted. Although this was a necessary choice as those muscles failed the reproducibility criteria, it could lead to areas of the muscles that were included being overfit.

Future studies should take three things into account. First, they should expect to find a lower segmentation accuracy when using deep learning-based methods to segment subjects with MSK disorders or older individuals, as the muscles are likely to appear less homogenous within images. Second, the number of subjects enrolled in a cohort plays an important role, the greater the number of subjects, the higher the segmentation accuracy. Third, the MR image types should be multiple focussing on fat suppression, as studies using images of these types provided more accurate predictions than those of traditional T1-weighted images (i.e., those used in this study). However, using more sophisticated and less common acquisition settings could reduce the utility of the segmentation method, so caution is advised in this respect.

6.5. Conclusion

This chapter showcased the use of 2 state-of-the-art deep CNNs in the application of muscle segmentation, as well as a novel CNN that built in an understanding of the spatial location of each image. Though novel deep learning-based methods built to segment the muscles from MR images are still recommended, it was shown that knowledge of the spatial location of the images was not beneficial to the novel network's training in this context. Also assessed was the benefit of training each of the networks with augmented images generated through image registration. It was shown that the accuracy of the predicted segmentations was drastically improved for 2 out of the 3 models, but only slightly with the third one. Augmented data is known to increase the robustness of segmentation models, and for spatially invariant images such as these, generating them through image registration is certainly a valid method to boost the size of a training dataset. This work highlights that the research community is close to presenting an automatic segmentation method that is equally accurate to the gold standard manual process, but it is currently more difficult for these automatic methods to segment inhomogeneous muscle structures.

Chapter 7:

General discussion and conclusions

7.1. Conclusion

The research conducted and presented in this PhD thesis fulfilled the aims and objectives outlined in Chapter 1. In particular, Chapters 4-6 are all closely tied to one another, each with the results of the previous chapters motivating the techniques used to solve increasingly difficult challenges. Chapter 3 first presented an in-depth critical appraisal of the current gold standard approach. Manual segmentation was used to segment the MultiSim database (that used throughout the thesis), and the repeatability issues highlighted when this technique was used were presented. Many of the lower limb muscles were immediately removed from further study, as their boundaries were not able to be reliably located within the Magnetic Resonance (MR) images used. The second research outcome of Chapter 3 was a novel application of manual muscle segmentation: investigating the fat infiltration between healthy (NW), obese (OB), and Dynapenic Abdominal Obese (DAO) subjects. Comparisons of the level of fat infiltration showed that the DAO had a significantly higher level of fat infiltration that the NW cohort and similar level to the OB cohort. Further, the OB cohort were able to exert more force than the NW and DAO cohorts, DAO and NW were comparable.

Chapter 4 centred around building an automatic muscle segmentation pipeline employing deformable registration to segment all lower limb muscles within one subject, using others as references, something that had not been explored previously. The intra-subject (left to right) segmentation was found to be relatively accurate with mean DSC between 0.75 and 0.85. On the other hand, the mean DSC found within the initial inter-subject segmentation was between 0.30 and 0.70. Therefore, the inter-subject segmentation was found to be significantly less accurate than the intra-subject (left to right) segmentation, and it was hypothesised that the other tissues visible within the scan (explicitly, the fat tissue surrounding the muscles) were skewing the registration, causing an inadequate registration.

Chapter 5 aimed at optimising the registration and segmentation pipeline built in Chapter 4, to maximise the segmentation accuracy of the inter-subject segmentation approach. The registration pipeline was improved by pre-processing the images so that the fat tissue surrounding the muscle tissue was homogenised. There was a significant increase in segmentation accuracy in 12/20 combinations of subjects after the images were pre-processed, with the mean DSC of the optimal subject combinations being between 0.68 and 0.73 DSC. Though there was a significant increase in segmentation accuracy, the segmentation accuracy was still found to be significantly lower than the intra-subject segmentation. Therefore, a multi-atlas approach was used to remove inaccuracies and increase confidence in the areas of agreement between different inter-subject segmentation iterations. Though the multi-atlas approach did help the segmentation accuracy, with the segmentation accuracy being between 0.66 and 0.75, it was found still to be lower than the intra-subject analysis. As the deformable image registration-based approaches used did not appear to have the capacity to segment the muscles to a satisfactory level, the remaining part of the work focused on the development, application, and testing of a deep learning approach. Nevertheless, considering the low number of available subjects, an approach to augment the training dataset was developed, using deformable registration. By registering the imaging data of each subject to all others, the database was increased from 11 subjects to 80 subjects (11 original subjects plus 69 augmented subjects).

Chapter 6 focused on the use of CNNs to automatically segment all lower limb muscles simultaneously from the MultiSim cohort. Three separate analyses were investigated. Firstly, the benefit of the use of augmented imaging datasets was evaluated for three models (UNet, Attention UNet, and Spatial channel UNet). All three networks were significantly improved, highlighting that the augmented imaging database achieved the goal of alleviating the unavoidable limitation of the small number of segmented subjects. Secondly, the segmentation accuracy of the three models was compared for an initial testing subject, finding that the UNet was the optimal network model. Finally, the robustness of the UNet was assessed, by retraining it for the five subjects segmented with other models, allowing direct comparison between the segmentation methods. The UNet was robust to the changes in the testing subject, with mean DSC all between 0.80 and 0.85 across the six testing subjects. Comparing between different segmentation methods, the left to right registration was the optimal segmentation method, but the UNet trained with the augmented images was a clear second best, outperforming both the single and multi-atlas registration-based methods. The outcome of the chapter was that deep learning-based methods, aided by deformable image registration can provide a good automatic segmentation model for the muscles from T1-weighted MR images.

7.2. Limitations

There were inevitable and unavoidable errors incurred in the manual process, due to operator dependency issues [28]. These errors would certainly have had an effect across all tested segmentation methods. Within the deformable image registration-based approaches, the automatic segmentations were generated through deformation of the manual segmentations. Therefore, any errors associated with the manual segmentation would have been apparent within the automatically generated segmentations. Furthermore, the error metrics are all calculated by direct comparison with the manual segmentations that they sought to replicate. Any errors associated with the manual segmentations that were compared to would also affect the computed error metrics. These inaccuracies associated with the manual segmentation method would also have affected the training of the CNNs. Training CNNs with data generated by humans with human error included is not a novel problem, and typically causes the results to be limited to replicate the level of human error and not outperform it. Though, the effects of these inaccuracies were minimised by removing the muscles that failed an inclusion criterion from further study (Chapter 3 Section 2). These errors could not be avoided as all analyses required some reference muscle segmentations that should only be gathered using the gold standard approach.

The imaging data used throughout the thesis was also likely to have an effect on the accuracy found between the different approaches. Firstly, T1-weighted images are not designed to highlight subtle differences between soft tissue, but rather highlight areas of white or grey matter [152]. Images of this acquisition setting do highlight the different tissues visible in the lower limb (trabecular and cortical bone, muscle, and fat), but other acquisition settings such as fat-suppression may highlight the muscles and remove other tissues. Though, scans of this acquisition setting are extremely common and standard clinical practise, so their use throughout this thesis were justified as an automatic segmentation method proven with T1-weighted images would have a more wide-reaching impact. Secondly, the resolution of the images may have affected the accuracy of the automatic segmentation methods tested, as higher resolution images would have more detailed features that would facilitate the isolation of muscle boundaries more readily. On the other hand, with higher resolution images, the computational expense of all methods used would have been exaggerated, with the number of pixels required to be processed being greater. The selection of resolution of the images particularly within the long area of the lower limb must be chosen with caution as higher resolution images require more time to be gathered, meaning the subject would be in the MR imaging chamber longer.

Likely the limitation of greatest impact was the small number of subjects available (n = 11). A small cohort such as this did not impact the registration-based methods but are known to affect the training of neural networks [116]. The impact of this limitations for the neural network-based segmentation methods was alleviated using augmented imaging datasets (Section 6.2.2.). Moreover, there are more variables for future researchers to take into account, namely: the characteristics of the subjects under investigation (e.g. number of subjects, age, musculoskeletal disorder, sex), and the images they are able to acquire. The characteristics of the subjects under investigation impacts the potential feasibility for the methods explored to produce accurate segmentations as the presence of a musculoskeletal disorder (sarcopenia due to aging, Motor Neuron Disease, muscular dystrophy, for further examples see Section 2.2.2) affects the homogeneity of the muscles visible within medical images [53, 60].

7.3. Impact

The work conducted throughout the PhD has contributed to the publication of two papers (one as first author [81] and one as a co-author [66]). Additionally, codes and databases have been written and published on GitHub and Figshare: **1)** automatic pre-processing of MR images acquired from different sections (https://github.com/whhenson1), **2)** multi-atlas segmentation (https://doi.org/10.15131/shef.data.21763982), **3)** augmented images and associated segmentations, located at https://doi.org/10.15131/shef.data.20440164, and https://doi.org/10.15131/shef.data.20440203, respectively), and **4)** the traditional UNet, Attention UNet, and Spatial channel UNet (https://github.com/whhenson1). Two further papers are currently in preparation for publication: “Traditional and novel Deep learning-based muscle segmentation from MR images enhanced with deformable image registration-generated augmented imaging datasets” (first author), and “Comparison of fat infiltration and muscle functional capacity in healthy, obese, and dynapenic abdominal obese adults” (shared first authorship with Dr. Lisa M Dowling).

The data and codes written in the process of producing this thesis have been made publicly available with the aim of focussing future researchers toward using and adapting existing tools to analyse data, rather than on exploring traditional (image registration) or probabilistic (deep learning) approaches. There are already countless retrospective datasets [122] that could be segmented using tools (such as those explored in this thesis) that could aid our understanding of the mechanisms under which muscle disorders operate. The impact of the thesis, therefore, is that the investigations into the benefits and limitations of automatic segmentation tools have been explored, which the author hopes will aid future researchers to apply such tools to more clinical research questions.

Moreover, the automatic pre-processing algorithm, segmentation methods, and post-processing techniques designed and tested within this thesis serve as groundwork for clinical translation of such methods to segment the muscles of real patients. The methods presented and tested in this thesis could be used in their current state to generate predictions of individual muscle characteristics of new patients within the clinic, which would need only minor adjustment by trained experts (such as radiologists). After retraining, the deep learning-based approaches, with the availability of a greater number of imaging datasets captured for more subjects, the methods presented in this thesis could be used to produce segmentations of even greater accuracy and enable processing of subjects and patients. The methods presented could be used to allow quantitative guidance of the appropriate steps to counter muscle disorders with targeted physiotherapy, inform clinicians whether a treatment is benefitting a patient, and catalyse our understanding of muscle disorders and the specific areas that they affect.

The thesis aimed at providing the community with an automatic muscle segmentation tool and this has been achieved, using a variety of different methods. The motivation for this was to facilitate future quantitative studies into the effects and characteristics of muscle disorders. The potential of this tool is twofold, with the first being in providing clinicians with a quantitative measure of the damage to individual muscles allowing a better understanding of risk of an individual being diagnosed with a muscle disorder. The second was to enable quantitative methods to assess the progression or intervention of muscle disorders [90, 153]. The individual automatic muscle segmentation methods explored in this thesis could, with further testing on larger cohorts, provide a tool capable of supplying a quantitative analysis of the state of an individual's muscle health to the community.

7.4. Future work

Future work should be focussed on three areas:

1) Further testing. The deep learning-based approaches built in Chapter 6 should be tested further through collecting more lower limb MR imaging data of older females and manually segmenting all muscles. Supplying these extra datasets into the training phase and pairing the new subjects with the eleven current subjects to generate more augmented imaging data could boost the segmentation accuracy found using all three of the networks tested. The more subjects that can be added to the database, the greater the segmentation accuracy is likely to be. The recently collected STH21022 database of older women (n = 27, normal weight, obese and dynapenic abdominal obese) could be incorporated into the database, as these images were acquired with identical scanning parameters (see Section 3.3.2.1.). Caution is advised as the operator repeatability of the manual segmentation process must be characterised for each muscle before proceeding to retrain the networks with additional subjects.

2) Continuing the investigation of muscle fat infiltration or muscle damage into subjects with muscle disorders. The work outlined in Chapter 3 Section 3, regarding the level of fat infiltration into the quadriceps and hamstrings can be extended using the techniques used for automatic segmentation of individual muscle to quantify the level of fat infiltration into specific muscles. The combination of these two methods could lead to a better understanding of the progressive characteristics of muscle disorders and could also provide quantitative insight into the effects of new and existing intervention strategies.

3) A robust segmentation tool capable of generating segmentations of any new subject. The MR images that are collected not only from within clinics but also research groups often are collected with varying acquisition settings and parameters. An automatic segmentation tool should be robust to changes such as these. This thesis lays the groundwork for a general answer to the question of which methods should be used to automatically segment individual muscles. The methods tested and analysed within this thesis should be carried forward with larger and more diverse datasets. For example, retraining the CNNs tested within chapter 6 (with the enhancement of data augmentation) with large databases consisting of many subjects from different cohorts (young and old,

with and without a muscle disorder, different acquisition settings) could lead to a robust automatic segmentation tool for any new subject that requires processing.

WORD COUNT:
44591

8.1. References

1. Woolf, A.D. and B. Pfleger, *Burden of major musculoskeletal conditions.* Bull World Health Organ, 2003. **81**(9): p. 646-56.
2. Palmer, K.T. and N. Goodson, *Ageing, musculoskeletal health and work.* Best Pract Res Clin Rheumatol, 2015. **29**(3): p. 391-404.
3. Malik, K.M., R. Beckerly, and F. Imani, *Musculoskeletal Disorders a Universal Source of Pain and Disability Misunderstood and Mismanaged: A Critical Analysis Based on the U.S. Model of Care.* Anesth Pain Med, 2018. **8**(6): p. e85532.
4. Rantanen, T., *Muscle strength, disability and mortality.* Scand J Med Sci Sports, 2003. **13**(1): p. 3-8.
5. Simpson, J.M. and S. Salkin, *Are elderly people at risk of falling taught how to get up again?* Age Ageing, 1993. **22**(4): p. 294-6.
6. Novak, I., et al., *Early, Accurate Diagnosis and Early Intervention in Cerebral Palsy: Advances in Diagnosis and Treatment.* JAMA Pediatr, 2017. **171**(9): p. 897-907.
7. Edmunds, K.J., et al., *Quantitative Computed Tomography and Image Analysis for Advanced Muscle Assessment.* Eur J Transl Myol, 2016. **26**(2): p. 6015.
8. Doi, K., *Diagnostic imaging over the last 50 years: research and development in medical imaging science and technology.* Phys Med Biol, 2006. **51**(13): p. R5-27.
9. Ayache, J.S.D.N., *Medical image analysis: progress over two decades and the challenges ahead.* IEEE Trans Pattern Anal Mach Intell, 1995. **22**(1).
10. Casiraghi, E., et al., *MIAQuant, a novel system for automatic segmentation, measurement, and localization comparison of different biomarkers from serialized histological slices.* Eur J Histochem, 2017. **61**(4): p. 2838.
11. Jalalifar, S.A., et al., *Impact of Tumour Segmentation Accuracy on Efficacy of Quantitative MRI Biomarkers of Radiotherapy Outcome in Brain Metastasis.* Cancers (Basel), 2022. **14**(20).
12. Bhalodiya, J.M., S.N. Lim Choi Keung, and T.N. Arvanitis, *Magnetic resonance image-based brain tumour segmentation methods: A systematic review.* Digit Health, 2022. **8**: p. 20552076221074122.
13. Arnold, E.M., et al., *A model of the lower limb for analysis of human movement.* Ann Biomed Eng, 2010. **38**(2): p. 269-79.
14. Byrne, N., et al., *A systematic review of image segmentation methodology, used in the additive manufacture of patient-specific 3D printed models of the cardiovascular system.* JRSM Cardiovasc Dis, 2016. **5**: p. 2048004016645467.
15. Hollon, M. and A. Faloye, *Utilization of the Transgastric View of the Left Atrial Appendange for Procedural Guidance during Left Atrial Appendage Clip Via Video-assisted Thoracoscopic Surgery.* J Cardiovasc Echogr, 2020. **30**(4): p. 211-213.
16. Munteanu, S.E., et al., *Characterisation of first metatarsophalangeal joint osteoarthritis using magnetic resonance imaging.* Clin Rheumatol, 2021. **40**(12): p. 5067-5076.
17. Dodin, P., et al., *Automatic human knee cartilage segmentation from 3D magnetic resonance images.* IEEE Trans Biomed Eng, 2010. **57**(11).
18. Shim, H., et al., *Knee cartilage: efficient and reproducible segmentation on high-spatial-resolution MR images with the semiautomated graph-cut algorithm method.* Radiology, 2009. **251**(2): p. 548-56.
19. Belvedere, C., et al., *New comprehensive procedure for custom-made total ankle replacements: Medical imaging, joint modeling, prosthesis design, and 3D printing.* J Orthop Res, 2019. **37**(3): p. 760-768.

20. Ahmed, S.M. and R.J. Mstafa, *A Comprehensive Survey on Bone Segmentation Techniques in Knee Osteoarthritis Research: From Conventional Methods to Deep Learning.* Diagnostics (Basel), 2022. **12**(3).

21. Wani, I.M. and S. Arora, *Computer-aided diagnosis systems for osteoporosis detection: a comprehensive survey.* Med Biol Eng Comput, 2020. **58**(9): p. 1873-1917.

22. Scott, D.L. and G.H. Kingsley, *Use of imaging to assess patients with muscle disease.* Curr Opin Rheumatol, 2004. **16**(6): p. 678-83.

23. Fielding, R.A., et al., *Sarcopenia: an undiagnosed condition in older adults. Current consensus definition: prevalence, etiology, and consequences. International working group on sarcopenia.* J Am Med Dir Assoc, 2011. **12**(4): p. 249-56.

24. Ignasiak, D., et al., *The effect of muscle ageing and sarcopenia on spinal segmental loads.* Eur Spine J, 2018. **27**(10): p. 2650-2659.

25. Cruz-Jentoft, A.J., *Diagnosing sarcopenia: turn your eyes back on patients.* Age Ageing, 2021. **50**(6): p. 1904-1905.

26. Cruz-Jentoft, A.J., et al., *Sarcopenia: European consensus on definition and diagnosis: Report of the European Working Group on Sarcopenia in Older People.* Age Ageing, 2010. **39**(4): p. 412-23.

27. Cruz-Jentoft, A.J., et al., *Sarcopenia: revised European consensus on definition and diagnosis.* Age Ageing, 2019. **48**(1): p. 16-31.

28. Pons, C., et al., *Quantifying skeletal muscle volume and shape in humans using MRI: A systematic review of validity and reliability.* PLoS One, 2018. **13**(11): p. e0207847.

29. Casas-Herrero, A., et al., *Functional capacity, muscle fat infiltration, power output, and cognitive impairment in institutionalized frail oldest old.* Rejuvenation Res, 2013. **16**(5): p. 396-403.

30. Emanuelsson, E.B., et al., *MRI characterization of skeletal muscle size and fatty infiltration in long-term trained and untrained individuals.* Physiol Rep, 2022. **10**(14): p. e15398.

31. Naimo, M.A., et al., *Skeletal Muscle Quality: A Biomarker for Assessing Physical Performance Capabilities in Young Populations.* Front Physiol, 2021. **12**: p. 706699.

32. Pillen, S., I.M. Arts, and M.J. Zwarts, *Muscle ultrasound in neuromuscular disorders.* Muscle Nerve, 2008. **37**(6): p. 679-93.

33. Freeman, M.A. and V. Pinskerova, *The movement of the knee studied by magnetic resonance imaging.* Clin Orthop Relat Res, 2003(410): p. 35-43.

34. B, H.B.B., *Physiology, Smooth Muscle.* Treasure Island: StatPearls, 2022.

35. Severs, N.J., *The cardiac muscle cell.* Bioessays, 2000. **22**(2): p. 188-99.

36. Charles, J., et al., *From fibre to function: are we accurately representing muscle architecture and performance?* Biol Rev Camb Philos Soc, 2022. **97**(4): p. 1640-1676.

37. Blazevich, A.J. and N.C. Sharp, *Understanding muscle architectural adaptation: macro- and micro-level research.* Cells Tissues Organs, 2005. **181**(1): p. 1-10.

38. University, O.S., *10.2 Skeletal Muscle.* Anatomy and physiology 1.1 OE.

39. Edman, K.A., *The velocity of unloaded shortening and its relation to sarcomere length and isometric force in vertebrate muscle fibres.* J Physiol, 1979. **291**: p. 143-59.

40. Gordon, A.M., A.F. Huxley, and F.J. Julian, *The variation in isometric tension with sarcomere length in vertebrate muscle fibres.* J Physiol, 1966. **184**(1): p. 170-92.

41. Burkholder, T.J. and R.L. Lieber, *Sarcomere length operating range of vertebrate muscles during movement.* J Exp Biol, 2001. **204**(Pt 9): p. 1529-36.

42. Bianchi, E., et al., *Innovative Strategies in Tendon Tissue Engineering.* Pharmaceutics, 2021. **13**(1).
43. Taga, G., *A model of the neuro-musculo-skeletal system for human locomotion. I. Emergence of basic gait.* Biol Cybern, 1995. **73**(2): p. 97-111.
44. Delp, S.L., et al., *An interactive graphics-based model of the lower extremity to study orthopaedic surgical procedures.* IEEE Trans Biomed Eng, 1990. **37**(8): p. 757-67.
45. Wan, J.J., et al., *Muscle fatigue: general understanding and treatment.* Exp Mol Med, 2017. **49**(10): p. e384.
46. Karlsson, J., *Localized muscular fatigue: role of muscle metabolism and substrate depletion.* Exerc Sport Sci Rev, 1979. **7**: p. 1-42.
47. Ryder, S., et al., *The burden, epidemiology, costs and treatment for Duchenne muscular dystrophy: an evidence review.* Orphanet J Rare Dis, 2017. **12**(1): p. 79.
48. Norwood, F.L., et al., *Prevalence of genetic muscle disease in Northern England: in-depth analysis of a muscle clinic population.* Brain, 2009. **132**(Pt 11): p. 3175-86.
49. Dowling, J.J., et al., *Treating pediatric neuromuscular disorders: The future is now.* Am J Med Genet A, 2018. **176**(4): p. 804-841.
50. Yin, L., et al., *Skeletal muscle atrophy: From mechanisms to treatments.* Pharmacol Res, 2021. **172**: p. 105807.
51. Ramdas, S. and L. Servais, *New treatments in spinal muscular atrophy: an overview of currently available data.* Expert Opin Pharmacother, 2020. **21**(3): p. 307-315.
52. Wallengren, O., et al., *Comparison of the 2010 and 2019 diagnostic criteria for sarcopenia by the European Working Group on Sarcopenia in Older People (EWGSOP) in two cohorts of Swedish older adults.* BMC Geriatr, 2021. **21**(1): p. 600.
53. Cruz-Jentoft, A.J., M.C. Gonzalez, and C.M. Prado, *Sarcopenia not equal low muscle mass.* Eur Geriatr Med, 2023.
54. Sayer, A.A. and A. Cruz-Jentoft, *Sarcopenia definition, diagnosis and treatment: consensus is growing.* Age Ageing, 2022. **51**(10).
55. Library, J.H.U.M., *Muscular dystrophy.*
56. Engel, W.K., *The essentiality of histo- and cytochemical studies of skeletal muscle in the investigation of neuromuscular disease. 1962.* Neurology, 1998. **51**(3): p. 655 and 17 pages following.
57. De Jonghe, B., et al., *Acquired neuromuscular disorders in critically ill patients: a systematic review. Groupe de Reflexion et d'Etude sur les Neuromyopathies En Reanimation.* Intensive Care Med, 1998. **24**(12): p. 1242-50.
58. Martin, P.T. and H.H. Freeze, *Glycobiology of neuromuscular disorders.* Glycobiology, 2003. **13**(8): p. 67R-75R.
59. Rahman, S. and M.G. Hanna, *Diagnosis and therapy in neuromuscular disorders: diagnosis and new treatments in mitochondrial diseases.* J Neurol Neurosurg Psychiatry, 2009. **80**(9): p. 943-53.
60. Mercuri, E., et al., *Muscle MRI in inherited neuromuscular disorders: past, present, and future.* J Magn Reson Imaging, 2007. **25**(2): p. 433-40.
61. Morrow, J.M., et al., *MRI biomarker assessment of neuromuscular disease progression: a prospective observational cohort study.* Lancet Neurol, 2016. **15**(1): p. 65-77.
62. Shohji, T., *[9. Basics of Radiation Exposure in X-ray CT].* Nihon Hoshasen Gijutsu Gakkai Zasshi, 2021. **77**(10): p. 1223-1230.
63. Kalender, W.A., *X-ray computed tomography.* Phys Med Biol, 2006. **51**(13): p. R29-43.

64. Ta Yeong Wu, N.G., Chee Yang Teh & J Xiao Wen Hay *Theory and Fundamentals of Ultrasound.* Part of the SpringerBriefs in Molecular Science book series, 2012. **Advances in Ultrasound Technology for Environmental Remediation pp 5–12**.
65. Albayda, J. and N. van Alfen, *Diagnostic Value of Muscle Ultrasound for Myopathies and Myositis.* Curr Rheumatol Rep, 2020. **22**(11): p. 82.
66. Montefiori, E., et al., *MRI-based anatomical characterisation of lower-limb muscles in older women.* PLoS One, 2020. **15**(12): p. e0242973.
67. Bloembergen, N., E.M. Purcell, and R.V. Pound, *Nuclear magnetic relaxation.* Nature, 1947. **160**(4066): p. 475.
68. Bloch, F., *The Principle of Nuclear Induction.* Science, 1953. **118**(3068): p. 425-30.
69. Fawzi, A., A. Achuthan, and B. Belaton, *Brain Image Segmentation in Recent Years: A Narrative Review.* Brain Sci, 2021. **11**(8).
70. Klein, S., et al., *Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information.* Med Phys, 2008. **35**(4): p. 1407-17.
71. Burakiewicz, J., et al., *Quantifying fat replacement of muscle by quantitative MRI in muscular dystrophy.* J Neurol, 2017. **264**(10): p. 2053-2067.
72. Wren, T.A., et al., *Three-point technique of fat quantification of muscle tissue as a marker of disease progression in Duchenne muscular dystrophy: preliminary study.* AJR Am J Roentgenol, 2008. **190**(1): p. W8-12.
73. Leung, D.G., *Advancements in magnetic resonance imaging-based biomarkers for muscular dystrophy.* Muscle Nerve, 2019. **60**(4): p. 347-360.
74. Godi, C., et al., *Longitudinal MRI quantification of muscle degeneration in Duchenne muscular dystrophy.* Ann Clin Transl Neurol, 2016. **3**(8): p. 607-22.
75. Layec, G., et al., *The validity of anthropometric leg muscle volume estimation across a wide spectrum: from able-bodied adults to individuals with a spinal cord injury.* J Appl Physiol (1985), 2014. **116**(9): p. 1142-7.
76. Marcon, M., et al., *Cross-sectional area measurements versus volumetric assessment of the quadriceps femoris muscle in patients with anterior cruciate ligament reconstructions.* Eur Radiol, 2015. **25**(2): p. 290-8.
77. Fukunaga, T., et al., *Muscle volume is a major determinant of joint torque in humans.* Acta Physiol Scand, 2001. **172**(4): p. 249-55.
78. Holzbaur, K.R., et al., *Moment-generating capacity of upper limb muscles in healthy adults.* J Biomech, 2007. **40**(11): p. 2442-9.
79. Trappe, S.W., et al., *Calf muscle strength in humans.* Int J Sports Med, 2001. **22**(3): p. 186-91.
80. Ding, J., et al., *Deep learning-based thigh muscle segmentation for reproducible fat fraction quantification using fat-water decomposition MRI.* Insights Imaging, 2020. **11**(1): p. 128.
81. Henson, W.H., C. Mazza, and E. Dall'Ara, *Deformable image registration based on single or multi-atlas methods for automatic muscle segmentation and the generation of augmented imaging datasets.* PLoS One, 2023. **18**(3): p. e0273446.
82. Ni, R., et al., *Automatic segmentation of all lower limb muscles from high-resolution magnetic resonance imaging using a cascaded three-dimensional deep convolutional neural network.* J Med Imaging (Bellingham), 2019. **6**(4): p. 044009.
83. Ogier, A., et al., *Individual muscle segmentation in MR images: A 3D propagation through 2D non-linear registration approaches.* Annu Int Conf IEEE Eng Med Biol Soc, 2017. **2017**: p. 317-320.
84. Skorupska, E., et al., *Reliability of MR-Based Volumetric 3-D Analysis of Pelvic Muscles among Subjects with Low Back with Leg Pain and Healthy Volunteers.* PLoS One, 2016. **11**(7): p. e0159587.

85. Springer, I., et al., *Intra- and interobserver variability of magnetic resonance imaging for quantitative assessment of abductor and external rotator muscle changes after total hip arthroplasty.* Eur J Radiol, 2012. **81**(5): p. 928-33.
86. Lareau-Trudel, E., et al., *Muscle Quantitative MR Imaging and Clustering Analysis in Patients with Facioscapulohumeral Muscular Dystrophy Type 1.* PLoS One, 2015. **10**(7): p. e0132717.
87. Zhu, J., et al., *Deep learning methods for automatic segmentation of lower leg muscles and bones from MRI scans of children with and without cerebral palsy.* NMR Biomed, 2021. **34**(12): p. e4609.
88. Montefiori, E., et al., *An image-based kinematic model of the tibiotalar and subtalar joints and its application to gait analysis in children with Juvenile Idiopathic Arthritis.* J Biomech, 2019. **85**: p. 27-36.
89. Redl, C., M. Gfoehler, and M.G. Pandy, *Sensitivity of muscle force estimates to variations in muscle-tendon properties.* Hum Mov Sci, 2007. **26**(2): p. 306-19.
90. Gadermayr, M., et al., *A comprehensive study on automated muscle segmentation for assessing fat infiltration in neuromuscular diseases.* Magn Reson Imaging, 2018. **48**: p. 20-26.
91. Heskamp, L., et al., *Whole-muscle fat analysis identifies distal muscle end as disease initiation site in facioscapulohumeral muscular dystrophy.* Commun Med (Lond), 2022. **2**(1): p. 155.
92. Bas, J., et al., *Fat fraction distribution in lower limb muscles of patients with CMT1A: A quantitative MRI study.* Neurology, 2020. **94**(14): p. e1480-e1487.
93. L Fontana, A.M., E Scalco, et al., *Multi-Steps Registration Protocol for Multimodal MR Images of Hip Skeletal Muscles in a Longitudinal Study.* Appl. Sci., 2020. **10**(7823).
94. Le Troter, A., et al., *Volume measurements of individual muscles in human quadriceps femoris using atlas-based segmentation approaches.* MAGMA, 2016. **29**(2): p. 245-57.
95. Wokke, B.H., et al., *Comparison of dixon and T1-weighted MR methods to assess the degree of fat infiltration in duchenne muscular dystrophy patients.* J Magn Reson Imaging, 2013. **38**(3): p. 619-24.
96. Lees, M.J., et al., *Muscle quality as a complementary prognostic tool in conjunction with sarcopenia assessment in younger and older individuals.* Eur J Appl Physiol, 2019. **119**(5): p. 1171-1181.
97. Ma, J., *Dixon techniques for water and fat imaging.* J Magn Reson Imaging, 2008. **28**(3): p. 543-58.
98. Sasaki, K. and R.R. Neptune, *Individual muscle contributions to the axial knee joint contact force during normal walking.* J Biomech, 2010. **43**(14): p. 2780-4.
99. Slicer, D., https://www.slicer.org.
100. BIODEX, https://www.biodex.com.
101. Ladang, A., et al., *Biochemical Markers of Musculoskeletal Health and Aging to be Assessed in Clinical Trials of Drugs Aiming at the Treatment of Sarcopenia: Consensus Paper from an Expert Group Meeting Organized by the European Society for Clinical and Economic Aspects of Osteoporosis, Osteoarthritis and Musculoskeletal Diseases (ESCEO) and the Centre Academique de Recherche et d'Experimentation en Sante (CARES SPRL), Under the Auspices of the World Health Organization Collaborating Center for the Epidemiology of Musculoskeletal Conditions and Aging.* Calcif Tissue Int, 2023. **112**(2): p. 197-217.
102. Broderick, B.J., et al., *Technique for the computation of lower leg muscle bulk from magnetic resonance images.* Med Eng Phys, 2010. **32**(8): p. 926-33.

103. Engstrom, C.M., et al., *Segmentation of the quadratus lumborum muscle using statistical shape modeling.* J Magn Reson Imaging, 2011. **33**(6): p. 1422-9.

104. Andrews, S. and G. Hamarneh, *The Generalized Log-Ratio Transformation: Learning Shape and Adjacency Priors for Simultaneous Thigh Muscle Segmentation.* IEEE Trans Med Imaging, 2015. **34**(9): p. 1773-87.

105. Kalisman, A. and M. Kalisman, *Image processing--principles and its future applications in reconstructive and aesthetic plastic surgery.* Clin Plast Surg, 1986. **13**(3): p. 513-27.

106. Asman, A.J. and B.A. Landman, *Non-local STAPLE: an intensity-driven multi-atlas rater model.* Med Image Comput Comput Assist Interv, 2012. **15**(Pt 3): p. 426-34.

107. Wang, H. and P.A. Yushkevich, *Multi-atlas segmentation with joint label fusion and corrective learning-an open source implementation.* Front Neuroinform, 2013. **7**: p. 27.

108. Yokota, F., et al., *Automated muscle segmentation from CT images of the hip and thigh using a hierarchical multi-atlas method.* Int J Comput Assist Radiol Surg, 2018. **13**(7): p. 977-986.

109. Sun, L., L. Zhang, and D.Q. Zhang, *Multi-Atlas Based Methods in Brain MR Image Segmentation.* Chin Med Sci J, 2019. **34**(2): p. 110-119.

110. Dong, P., et al., *Multi-Atlas Segmentation of Anatomical Brain Structures Using Hierarchical Hypergraph Learning.* IEEE Trans Neural Netw Learn Syst, 2020. **31**(8): p. 3061-3072.

111. Karlsson, A., et al., *Automatic and quantitative assessment of regional muscle volume by multi-atlas segmentation using whole-body water-fat MRI.* J Magn Reson Imaging, 2015. **41**(6): p. 1558-69.

112. Pereira, S., et al., *Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images.* IEEE Trans Med Imaging, 2016. **35**(5): p. 1240-1251.

113. Bhandari, A., J. Koppen, and M. Agzarian, *Convolutional neural networks for brain tumour segmentation.* Insights Imaging, 2020. **11**(1): p. 77.

114. O. Ronneberger, P.F., T. Brox *U-Net: Convolutional Networks for Biomedical Image Segmentation.* Springer, Cham., 2015. **vol 9351**(In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.).

115. Z. Alom, T.M.T., C Yakopcic et al., *A State-of-the-Art Survey on Deep Learning Theory and Architectures.* Electronics, 2019. **8**: p. 292.

116. Hesamian, M.H., et al., *Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges.* J Digit Imaging, 2019. **32**(4): p. 582-596.

117. S Ghosh, P.B., S. T. Acton, S. S. Blemker; N. Ray, *Automated 3D muscle segmentation from MRI data using convolutional neural network.* 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China,, 2017. **pp. 4437-4441**

118. A. Krizhevsky, I.S., G. E. Hinton, *ImageNet classification with deep convolutional neural networks.* Communications of the ACM 2017. **60**(6): p. 84-90.

119. Krithika Alias AnbuDevi, M. and K. Suganthi, *Review of Semantic Segmentation of Medical Images Using Modified Architectures of UNET.* Diagnostics (Basel), 2022. **12**(12).

120. Guo, Z., et al., *Deep Learning-based Image Segmentation on Multimodal Medical Imaging.* IEEE Trans Radiat Plasma Med Sci, 2019. **3**(2): p. 162-169.

121. Arnold, A.S., et al., *Contributions of muscles to terminal-swing knee motions vary with walking speed.* J Biomech, 2007. **40**(16): p. 3660-71.

122. Schlaeger, S., et al., *Thigh muscle segmentation of chemical shift encoding-based water-fat magnetic resonance images: The reference database MyoSegmenTUM.* PLoS One, 2018. **13**(6): p. e0198200.
123. O. Çiçek, A.A., S. S. Lienkamp, T. Brox, O. Ronneberger, *3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation.* arXiv, 2016.
124. Faulkner, J.A., et al., *The aging of elite male athletes: age-related changes in performance and skeletal muscle structure and function.* Clin J Sport Med, 2008. **18**(6): p. 501-7.
125. M. I. Razzak, S.N., A. Zaib *Deep Learning for Medical Image Processing: Overview, Challenges and the Future.* Part of the Lecture Notes in Computational Vision and Biomechanics book series (LNCVB,volume 26), 2017.
126. Carbone, V., et al., *TLEM 2.0 - a comprehensive musculoskeletal geometry dataset for subject-specific modeling of lower extremity.* J Biomech, 2015. **48**(5): p. 734-41.
127. Holzbaur, K.R., et al., *Upper limb muscle volumes in adult subjects.* J Biomech, 2007. **40**(4): p. 742-9.
128. J. Juntu, J.S., D. Van Dyck, J. Gielen *Bias Field Correction for MRI Images.* Part of the Advances in Soft Computing book series, 2005. **volume 30**(AINSC).
129. Materliase, *Mimics.*
130. Barber, D.C. and D.R. Hose, *Automatic segmentation of medical images using image registration: diagnostic and simulation applications.* J Med Eng Technol, 2005. **29**(2): p. 53-63.
131. LR, D., *Measures of the amount of ecological asociation between species.* Ecology, 1945. **26**(3): p. 297-302.
132. Rockafellar, R.T.W., Roger J-B, *Variational Analysis.* Springer-Verlag, ISBN, 2005: p. 117.
133. Modenese, L., et al., *Investigation of the dependence of joint contact forces on musculotendon parameters using a codified workflow for image-based modelling.* J Biomech, 2018. **73**: p. 108-118.
134. Hayford, C.F., et al., *Predicting longitudinal changes in joint contact forces in a juvenile population: scaled generic versus subject-specific musculoskeletal models.* Comput Methods Biomech Biomed Engin, 2020. **23**(13): p. 1014-1025.
135. Valente, G., et al., *Are subject-specific musculoskeletal models robust to the uncertainties in parameter identification?* PLoS One, 2014. **9**(11): p. e112625.
136. Gholipour, A., et al., *Multi-atlas multi-shape segmentation of fetal brain MRI for volumetric and morphometric analysis of ventriculomegaly.* Neuroimage, 2012. **60**(3): p. 1819-31.
137. Nalepa, J., M. Marcinkiewicz, and M. Kawulok, *Data Augmentation for Brain-Tumor Segmentation: A Review.* Front Comput Neurosci, 2019. **13**: p. 83.
138. Shen Z, X.Z., Olut S, Niethammer M., *Anatomical Data Augmentation via Fluid-based Image Registration.* arXiv, 2020.
139. Li, C., J.C. Gore, and C. Davatzikos, *Multiplicative intrinsic component optimization (MICO) for MRI bias field estimation and tissue segmentation.* Magn Reson Imaging, 2014. **32**(7): p. 913-23.
140. Canny, J., *A computational approach to edge detection.* IEEE Trans Pattern Anal Mach Intell, 1986. **8**(6): p. 679-98.
141. Belzunce, M.A., et al., *Automated multi-atlas segmentation of gluteus maximus from Dixon and T1-weighted magnetic resonance images.* MAGMA, 2020. **33**(5): p. 677-688.
142. Handsfield, G.G., et al., *Relationships of 35 lower limb muscles to height and body mass quantified using MRI.* J Biomech, 2014. **47**(3): p. 631-8.

143. Dall'Ara, E., D. Barber, and M. Viceconti, *About the inevitable compromise between spatial resolution and accuracy of strain measurement for bone tissue: a 3D zero-strain study.* J Biomech, 2014. **47**(12): p. 2956-63.

144. Shorten, C., T.M. Khoshgoftaar, and B. Furht, *Text Data Augmentation for Deep Learning.* J Big Data, 2021. **8**(1): p. 101.

145. Noguchi, S., et al., *Bone segmentation on whole-body CT using convolutional neural network with novel data augmentation techniques.* Comput Biol Med, 2020. **121**: p. 103767.

146. Antonelli, M., et al., *The Medical Segmentation Decathlon.* Nat Commun, 2022. **13**(1): p. 4128.

147. Isensee, F., et al., *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.* Nat Methods, 2021. **18**(2): p. 203-211.

148. O Oktay, J.S., L Le Folgoc, et al., *Attention U-Net: Learning Where to Look for the Pancreas.* arXiv, 2018.

149. PyTorch, *CrossEntropyLoss.*

150. Zhenhuan Z., C.F., Leyi X, Xilong Q, *DEA-UNet: a dense-edge-attention UNet architecture for medical image segmentation.* Journal of Electronic Imaging, 2022. **31**(4).

151. Xutai M., J.P., J. Cross, L. Puzon, Jiatao G., *Monotonic Multihead Attention.* arXiv, 2019.

152. Gerwinn, H., et al., *The (in)consistency of changes in brain macrostructure in male paedophiles: A combined T1-weighted and diffusion tensor imaging study.* J Psychiatr Res, 2015. **68**: p. 246-53.

153. Jenkins, T.M., et al., *Longitudinal multi-modal muscle-based biomarker assessment in motor neuron disease.* J Neurol, 2020. **267**(1): p. 244-256.

9.1. Appendix 1 – Automatic calculation of muscle length

This appendix is based on a paper published in P (2020): 'MRI-based anatomical characterisation of lower-limb muscles in older women' by E. Montefiori, B. Kalkman, **W. H. Henson**, M. Paggiosi, E. McCloskey, C. Mazzà. Doi: https://doi.org/10.1371/journal.pone.0242973

9.1.1. Methods

Individual muscle segmentations were used to calculate the anatomical muscle length, which was calculated as the length of the centreline from the 3D muscle segmentation. To generate the centreline, points representing the centre of mass (see Figure 9.1) of each segmentation were calculated at regular cross sections (spacing of 10 mm) in the sagittal plane. A smooth curve was fitted to the points using a moving average filter, with the span of the filter being selected individually for each muscle. The muscle length were then denoted as the arc length of the fitted smoothed curve constituting the centreline of the 3D segmentations. All above computations were performed in MATLAB R2019b (The Mathworks Inc., Natick, MA, USA).

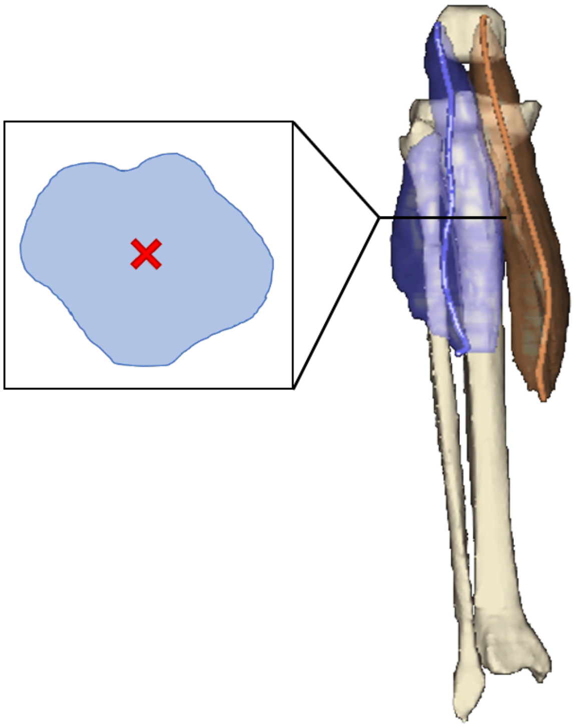


Figure 9.1: The process of calculating the length of the muscle automatically. The centre of mass (red cross) was calculated at regular section of each muscle segmentation. These centres of masses were connected with a smooth line to calculate the muscle length automatically.

9.2. Appendix 2 – Example augmented images

Figure 9.2 below showcases one segmented imaging slice from an original dataset and the augmented images that were generated by mapping the other ten subjects to Subject 1 (target). Visually, the augmented images are visually distinct from the target subject.

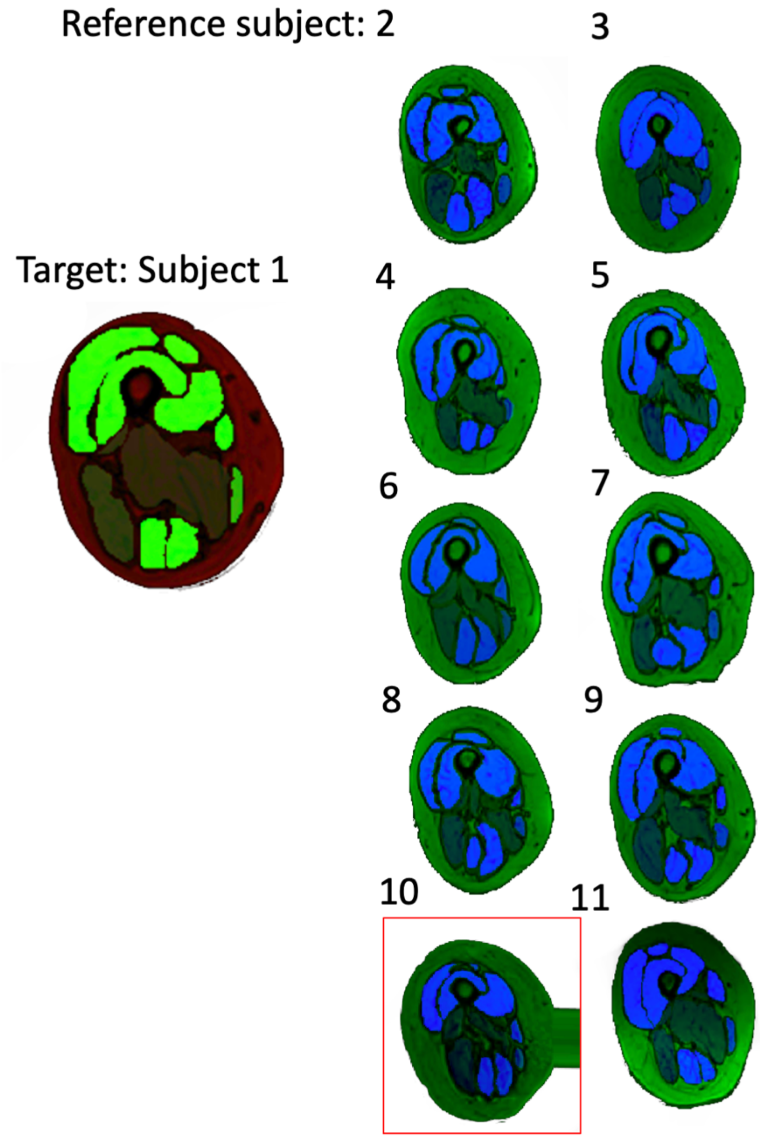


Figure 9.2: Display of augmented datasets for one target subject. The image on the left shows a cross-section of the target subject (Subject 1) with the manual segmentations for that image shown in green. The 10 images on the right are cross-sections of the augmented datasets, generated when keeping subject 1 as the target for the registration, whilst using the other 10 subjects as the reference dataset. Segmentations are reported in blue. One augmented dataset marked with a red square did not pass the inclusion criteria, due to the discontinuity in the boundary of the body.