Radiographic Assessment of Hip Disease in Children with Cerebral Palsy: Development of a Core Measurement Set and Analysis of an Artificial Intelligence System

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Master of Philosophy

by

Prince Josiah Sajanthan Joseph

August 2022

Abstract

Cerebral palsy is the most common physical disability during childhood. Cerebral palsy related hip disease is caused by an imbalance of muscle forces, resulting in progressive migration of the hip to complete dislocation. This can decrease function and quality of life. The prevention of hip dislocation is possible if detected early. Therefore, surveillance programmes have been set up to monitor children with cerebral palsy enabling clinicians to intervene early and improve outcomes. Currently, hip disease is assessed by analysing pelvic radiographs with various geometric measurements. This time-consuming task is undertaken frequently when monitoring a child with cerebral palsy. This thesis aimed to identify the key radiographic parameters used by clinicians (the core measurement set), and then build an artificial intelligence system to automate the calculation of this core measurement set.

A systematic review was conducted identifying a comprehensive list of previously reported measurements from studies measuring radiographic outcomes in cerebral palsy children with hip pathologies. Fifteen measurements were identified from the systematic review, of which Reimers' migration percentage was the most commonly reported. These measurements were used to perform a two-round Delphi study among orthopaedic surgeons and physiotherapists. Participants rated the importance of each measurement using a nine-point Likert scale ('not important' to critically important'). After the two rounds of the Delphi process, Reimers' migration percentage was included in the core measurement set. Following the final consensus meeting, the femoral head-shaft angle was also included.

The anteroposterior pelvic radiographs of 1650 children were then used to build an artificial intelligence system integrating the core measurement set, in collaboration with engineers from the University of Manchester. The newly developed artificial intelligence system was assessed by comparing its ability to calculate measurements and outline the pelvis and femur on a radiograph. The reliability of the dataset used to train the model was also analysed. The proposed artificial intelligence model achieved a 'good to excellent' inter-observer reliability across 450 radiographs when comparing its ability to calculate Reimers' migration percentage to five clinicians. Its ability to outline the pelvis and proximal femur was 'adequate' with the better performance observed in the pelvis than the femur. The reliability of the training dataset used to teach the artificial intelligence model was 'good' to 'very good'.

Artificial intelligence systems are feasible solutions to optimise the efficiency of hip radiograph analysis in cerebral palsy. Studies are warranted to include the core measurement set as a minimum when reporting on hip disease in cerebral palsy. Future research should investigate the feasibility of implementing a risk score to predict the likelihood of hip displacement.

Acknowledgements

I would like to express my sincerest gratitude to my primary supervisor, Professor Daniel C. Perry, who recognised my enthusiasm for academic research and formulated this MPhil opportunity. I really appreciate the trust placed in me to complete such important and exciting research. Professor Perry has supported me throughout the entirety of my degree and has always made himself available regardless of other commitments. He has helped cultivate my interest in paediatric orthopaedic surgery and medical technology and accelerated my growth as an academic.

I would like to also thank Professor Timothy Cootes, Dr Claudia Lindner and Dr Peter Thompson for all the time and effort they have spent into helping me complete this degree. Their help has been greatly appreciated.

I also acknowledge the financial grants generously provided by the Royal College of Surgeons England and the University of Liverpool, which has enabled the completion of this degree.

I am extremely grateful to my friends and family, who have always been there for me when I've needed them the most.

Finally, I would like to express my deepest gratitude to my parents, whose prayers and support have been constant throughout my education and this degree. I attribute all of my success to them.

TGBTG.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AA	Sharp's angle or acetabular angle
AcI	Acetabular index
ADR	Acetabular depth ratio
AI	Artificial intelligence
ANN	Artificial neural network
AUC	Area under the curve
BSCOS	British society for children's orthopaedic surgery
CEA	Centre-edge angle
CI	Confidence intervals
CMS	Core measurement set
CNN	Convolutional neural networks
COMET	Core outcome measures in effectiveness trials
COS	Core outcome set
CP	Cerebral palsy
CPIPS	Cerebral Palsy Integrated Pathway Scotland
CPUP	Uppföljningsprogram för Cerebral Pares
CT	Computerised tomography
DDH	Developmental dysplasia of the hip
DEXA	Dual-energy X-ray absorptiometry
DT	Decision tree
ESA	Epiphyseal shaft angle
FHEI	Femoral head extrusion index
FHS	Femoral head shape/congruency
FRAX	Fracture risk assessment tool
GAfREC	Governance arrangements for research ethics committees
GMFCS	Gross Motor Function Classification System
HEA	Hilgenreiner epiphyseal angle
HSA	Femoral head-shaft angle
ICC	Intraclass correlation coefficients
KNN	K-nearest neighbor
MeI	Medialization index
MHR	Mose hip ratio
MRI	Magnetic resonance imaging
NSA	Femoral neck-shaft angle

PACS	Picture archiving computer systems
PAMP	Pelvic adjusted migration percentage
PCK	Percentage of correct key points
PFA	Pelvic femoral angle
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
RF	Random forest
RMP	Reimer’s migration percentage
RNN	Recurrent neural network
SD	Standard deviation
SL	Shenton's line
STA	Sourcil Tönnis angle
SVM	Support vector machines
T1a	Trainer 1’s first annotations
T1b	Trainer 1’s second annotations
T2	Trainer 2’s annotations
THA	Total hip arthroplasty
UK	United Kingdom
USA	United States of America

Chapter 1: Introduction

1.1 Overview

Cerebral palsy (CP) is the most common childhood physical disability.(1) Individuals with CP are at a higher risk of developing hip disease, which may require preventative or reconstructive surgery.(2) Hip displacement can decrease the quality of life and function in children with CP and has been reported to be the most common cause of pain in the condition.(3-5) Long-lasting hip displacement can increase the risk of hip pain in adulthood from 10% to 30%.(6) However, studies have shown that it is possible to prevent subluxation and dislocation of the hip if displacement is detected early, resulting in improved symptoms and function as well as a reduced likelihood of developing hip pain in adulthood.(7-10) This has resulted in the introduction of national hip surveillance programmes, such as Cerebral Palsy Integrated Pathway Scotland (CPIPS), which aim to detect hip displacement early and intervene to improve the disease prognosis.(11)

Currently, in CPIPS, hip disease is assessed via radiographs that are manually interpreted and analysed by a trained expert.(11) This can be and prone to human error, especially in hip surveillance programmes. Artificial intelligence (AI) systems that can analyse hip radiographs have proven to be highly accurate at recognising different conditions and complex anatomy, making AI an ideal tool for surveillance programmes.(12-14) However, only a limited amount of research has looked into developing AI systems to analyse CP hip radiographs. In order to build software capable of automatically and accurately analysing CP hip radiographs, the key measurements that need to be recorded must be clearly listed as a 'core measurement set' (CMS). This measurement set can also serve as a minimum requirement to be recorded in clinical studies. This will allow comparisons to be drawn between studies and centres and facilitate meta-analyses to be conducted. The use of a standardised set of measurements in the United Kingdom (UK) will pave the way for uniformity in implementing a national hip surveillance program and guide the development of AI software to automate this process.

1.2 Aims and Objectives

1.2.1 Aims

Given the lack of a clinically useable AI system that can automatically interpret CP hip radiographs, this study will aim to identify the critical measurements needed to create an automated system and help develop and review an AI system capable of calculating the identified measurements.

1.2.2 Objectives

- To systematically identify all the radiographic measurements that are currently used to assess hip disease in children with CP in the literature
- To describe the most important measurements needed to assess hip disease in CP children and form a CMS by conducting a Delphi study
- To describe the reliability of an AI system capable of automatically analysing CP hip radiographs through the calculation of core measurements
- To highlight the implications of the thesis results for future research looking into hip disease in CP or the automatic analysis of hip radiographs

1.2.3 Research Questions

1. What measurements are currently used to assess hip disease in children with CP?
2. What do orthopaedic surgeons and physiotherapists believe to be the most important measurements needed to assess hip disease in children with CP?
3. Can an AI system be developed to automatically interpret hip radiographs using the key measurements in children with CP?

1.3 Thesis Outline

This first chapter has introduced the research topic by giving an overview, stating the aims and objectives of the research as well as the research questions, and outlining the structure of the remaining thesis.

In the second chapter, the background and context of the study are discussed in detail. An overview of hip disease in CP is presented, including the epidemiology, pathophysiology, clinical features, radiographic assessment, classification and management. Hip surveillance programmes will also be explored, focussing on CPIPS. AI systems and concepts are introduced, followed by a description of the applications of these systems in medical imaging.

In the third chapter, a systematic review is undertaken to look for articles that have used measurements to assess hip radiographs in children with CP. The different measurements reported in each study are noted, and the definitions given for each measurement are compared. The extracted data points are synthesised and analysed to identify a list of all the different measurements reported in the literature and identify the most common measurements that will likely be included in the CMS.

In the fourth chapter, a two-round Delphi study is conducted using the identified measurements from the previous chapter with orthopaedic surgeons and physiotherapists. An online survey is sent to participating orthopaedic surgeons and physiotherapists, requesting them to score the importance of the listed measurements on a nine-point Likert scale. A final consensus meeting is held to discuss results from round two and finalise the core measurement set.

In the fifth chapter, the performance of a newly proposed AI model will be assessed by analysing its ability to calculate radiographic measurements and outline the pelvis and femur. Furthermore, the inter- and intra-observer reliability of the data used to train the model will also be assessed.

The final chapter concludes the thesis by summarising the main findings of the study in relation to the research aims and questions. It will also describe the contributions of this study and outline the implications for future research.

Chapter 2: Literature Review

2.1 Overview

This chapter gives an overview of hip disease in cerebral palsy (CP) and concepts needed to understand the role of artificial intelligence (AI) in medical imaging. The four main sub-chapters discussed include a general overview of CP hip disease, hip surveillance programmes, AI and AI in medical imaging. CP, as a standalone disease, will not be covered in this review. However, any topics that are relevant to hip disease will be discussed. A basic introduction to AI will be described to help the reader understand the difference between various AI systems. AI techniques can aid radiologists in every step of the way, from requesting images and scheduling to reporting results.(15) However, for the purposes of this literature review, only the image interpretation applications of AI will be explored.

2.2 Hip Disease in Cerebral Palsy

2.2.1 Background

CP, known eponymously as Little's disease, was first reported by William Little in the 1840s.(16) Following an international workshop on the definition and classification of CP in 2004, CP has been defined as an umbrella term that describes "a group of permanent disorders of the development of movement and posture, causing activity limitation, that are attributed to nonprogressive disturbances that occurred in the developing foetal or infant brain. The motor disorders of cerebral palsy are often accompanied by disturbances of sensation, perception, cognition, communication, and behaviour, by epilepsy, and by secondary musculoskeletal problems".(17)

'Hip disease' in CP is a generic term that describes a range of hip disorders that lead to degeneration and pain of the hip joint.(4) 'Hip displacement' and 'hip migration' are also generic terms that, by definition, describe the progression of the femoral head from normal articulation to complete dislocation. 'Hip subluxation' occurs when the acetabulum still covers at least one-third of the femoral head. 'Hip dysplasia' describes the incomplete covering of the femoral head by the acetabulum so that at least one-third of the femoral head is uncovered.(18)

Hip disease is commonly seen in children with cerebral palsy and can decrease their quality of life.(5, 19) Children with hip displacement often experience pain and are known to fracture the femoral shaft

once the hip becomes dislocated.(20, 21) Adduction contractures in the hip can worsen sitting balance and make caring for children with CP especially problematic, sometimes requiring total care for daily activities with perineal care rendered nearly impossible.(22, 23) Displaced hips are unlikely to reduce spontaneously and, when left untreated, can progress to complete dislocation, which is associated with degenerative arthritis and pain.(4)

2.2.2 Epidemiology

The incidence of CP is difficult to describe accurately as it is usually diagnosed during infancy in an outpatient setting; consequently, this data is not included in the birth certificate or the hospital database making it hard to monitor.(24) However, a systematic review and meta-analysis has established an incidence of 2.11 per 1000 live births.(25) CP is more frequently seen in males than females, and gender distribution across different types of CP is insignificant.(26) Time trends from the twentieth century show that CP incidence has been relatively stable in the past.(27) However, it has been reported that improvements in neonatal care since the late 1960s have resulted in an increase in the survival of preterm infants, which has influenced the rates of CP. There was an initial increase in the rates of CP associated with the increased survival of extremely low birth weight and low gestation infants; however, since 2000, although the survival of preterm infants has remained similar, there has been a decrease in the rates of CP which has been associated with a change in practice, namely an increase in the use of antenatal steroid therapy and a decrease in the use of postnatal dexamethasone.(28)

The prevalence of hip subluxation ranges from 25% to 60%, with complete dislocation present in 10% to 15%.(29) The incidence of hip displacement decreases with age. However, the risk of hip displacement is highest between 2 to 3 years of age.(30) No significant difference has been found in the prevalence of hip displacement between males and females.(31)

A relationship between the incidence of hip displacement and the severity of CP, as described by the Gross Motor Function Classification System (GMFCS), has been established. Children at GMFCS level V have the highest incidence of hip displacement and are 2.5 to 3 times more likely to develop hip displacement than children at GMFCS level III or IV.(32) Interestingly, an association between femoral deformities and hip displacement has not been supported.(33)

2.2.3 Pathophysiology

CP, by definition, is caused by disturbances that occur during the development of the brain.(17) Due to the complex combination of factors that can be attributed to this condition, the specific underlying cause for the development of CP usually remains unknown. It is helpful to view the aetiology of CP as casual pathways made up of risk factors.(34, 35) Multiple risk factors can contribute to the causal pathway, but the most important risk factors include preterm birth, intrauterine growth restriction, perinatal infection, and multiple births.(36, 37)

Injury to the developing brain may result in an inability to inhibit nerve impulses, causing hypertonia.(38) Hip displacement primarily occurs due to asymmetric spasticity in the hip adductors, hip flexors and medial hamstrings, positioning the hip in flexion, adduction, and internal rotation. This combination of abnormal positioning and continued muscle contraction results in the lateralization and proximal migration of the femoral head.(39)

As the femoral head continues to migrate in the posterolateral direction, the acetabulum gradually expands, completely dislocating the femoral head from the hip joint. This increased pressure also erodes the lateral lip of the acetabulum and deforms the femoral head.(39, 40)

Without protection from the acetabulum, the femoral head is further deformed due to the immense pressures imposed by surrounding soft tissue. Consequently, the femoral head degenerates, resulting in osteoarthritis and pain.(4)

2.2.4 Clinical Features

In CP, children with hip disease may present with hip pain, progressive adduction contractures or leg length discrepancy caused by unilateral hip dislocation. Additionally, in children with anterior hip dislocations, the femoral head can be palpated in the groin. A windswept hip may also be seen in rare cases where one hip has an abduction abnormality and the contralateral hip has an adduction abnormality resulting in a ‘windswept’ appearance.(41, 42)

2.2.4.1 Physical Examination

A physical examination can be carried out to investigate any suspicions of hip disease. The following steps summarise the different elements of the exam: (43)

1. Thomas test is used to assess hip flexion contractures.
2. Hip adduction contractures and spasticity are assessed via the measurement of hip abduction. This is assessed by passively abducting the hip whilst the hip and knee are fully extended.
3. Internal and external rotation is assessed with the hip and knee flexed.
4. Popliteal angle is measured to assess hamstring length.
5. Range of motion is assessed in the knees and ankles.
6. Leg length is measured.
7. Duncan Ely test is performed to assess the rectus femoris length.
8. Hip extension is measured.
9. Scoliosis and pelvic obliquity are evaluated (especially in non-ambulatory patients).
10. Difficulties in sitting, standing and walking are assessed.

A physical examination alone is inadequate to diagnose and assess the severity of hip disease. Therefore, a radiographic evaluation is carried out in addition to this.

2.2.5 Radiographic Assessment

The degree of hip migration can be assessed on an anteroposterior radiograph. A radiographic assessment allows the disease severity to be accurately quantified through the measurement of and between anatomical landmarks. The rate of hip migration can be reliably calculated by comparing consecutive radiographs, given that the positioning of the hip and pelvis are consistent. This can determine an individual's risk of developing hip subluxation or dislocation.(44)

Various radiographic measurements have been used to assess hip disease in CP, usually in combination with each other. Different measurements assess different aspects of the hip. Some common measurements include: Reimers' migration percentage (RMP) which assesses the acetabulum in relation to the femur and is used to assess the severity of hip displacement(45); femoral neck-shaft angle (NSA) which assesses the femur and is usually measured alongside femoral anteversion when planning to perform femoral osteotomies for the correction of coxa valga and femoral anteversion(46-48); acetabular index (AcI) which assesses the acetabulum and is used to quantify the degree of acetabular coverage of the femoral head, allowing surgeons to assess the severity of hip dysplasia and decide whether to reconstruct the acetabulum(49); and femoral head-shaft angle (HSA) which assess the proximal femur in cases where the femoral head is in valgus when compared to the femoral neck.(50, 51) Given the vast number of measurements available and the lack of a standard set of measurements, the reporting of CP hip radiographs varies greatly in research and clinical practice. Conversely, RMP

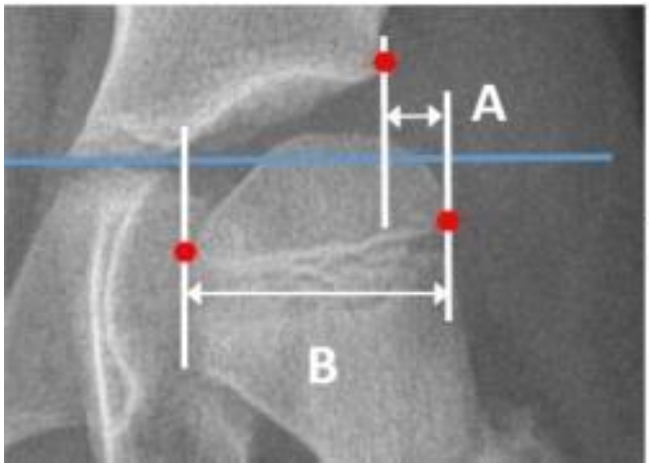
has proven to be particularly reliable and has been regarded as the ‘gold standard’ measurement to take when hips are assessed in children with CP.(45, 52-54)

2.2.5.1 Reimers’ migration percentage

In 1980, Reimers developed a measurement to assess the severity of hip displacement in children with CP. He classified hips as subluxated if the RMP was at least 33% or dislocated if RMP was at least 90%.(45) Although this criterion has been used in most studies, several recent studies have redefined subluxation and dislocation threshold values as 30% and 100%, respectively.(31, 55-59)

Reimers describes RMP as “the fraction (expressed as a percentage) of the visible part of the femoral head which on an AP radiograph has migrated beyond Perkin’s line/acetabular ridge.”(45) Perkin’s line is a vertical line perpendicular to Hilgenreiner’s line drawn along the most lateral aspect of the acetabular roof.(60) Hilgenreiner’s line is a horizontal line drawn through the most inferior aspect of the tri-radiate cartilages.(61) RMP is calculated by using the formula: A/B X 100=RMP, where A is the section of the femoral head that lies lateral to Perkin’s line and B is the total width of the femoral head (Figure 2.1).(62, 63)

Figure 2.1 Reimers’ migration percentage. Hilgenreiner's line is in blue. Perkin's line is the middle vertical line.



$$\mathrm{RMP} = 100\% \times A/B$$

2.2.5.2 Additional Imaging

Computerised tomography (CT) scans with three-dimensional reconstruction can be used to further assess femoral head deformities and acetabular deficiencies prior to surgery.(64) They are especially useful in anterior hip dislocations where RMP might be within normal ranges due to the absence of femoral head lateralisation.(65) CT scans are also used to assess proximal femoral geometry and

femoral anteversion unless coxa valga is present, in which case ultrasound is a cheap and more accurate alternative.(41, 66)

2.2.6 Classification

This condition can be classified based on motor abnormality, topography, or function. The motor abnormality classification divides the condition into spastic, dyskinetic and ataxic CP.(67) The topographic classification divides CP into quadriplegia, hemiplegia, diplegia, monoplegia, and triplegia.(68) However, motor abnormality and topographic classifications have been reported to be inconsistent and unreliable. This is due to disagreements on definitions and descriptions of terminology in each classification.(19, 69, 70)

Functional classification, more specifically the GMFCS, is a more reliable system to categorise CP and has been validated.(71, 72) The GMFCS is composed of five levels: I, II, III, IV and V (Table 2.1). Each level corresponds to a distinction in functional limitation, use of hand-held mobility devices or wheelchair, and quality of movement. Each of these distinctions are then further expanded upon by age group, allowing for a more accurate classification.(73)

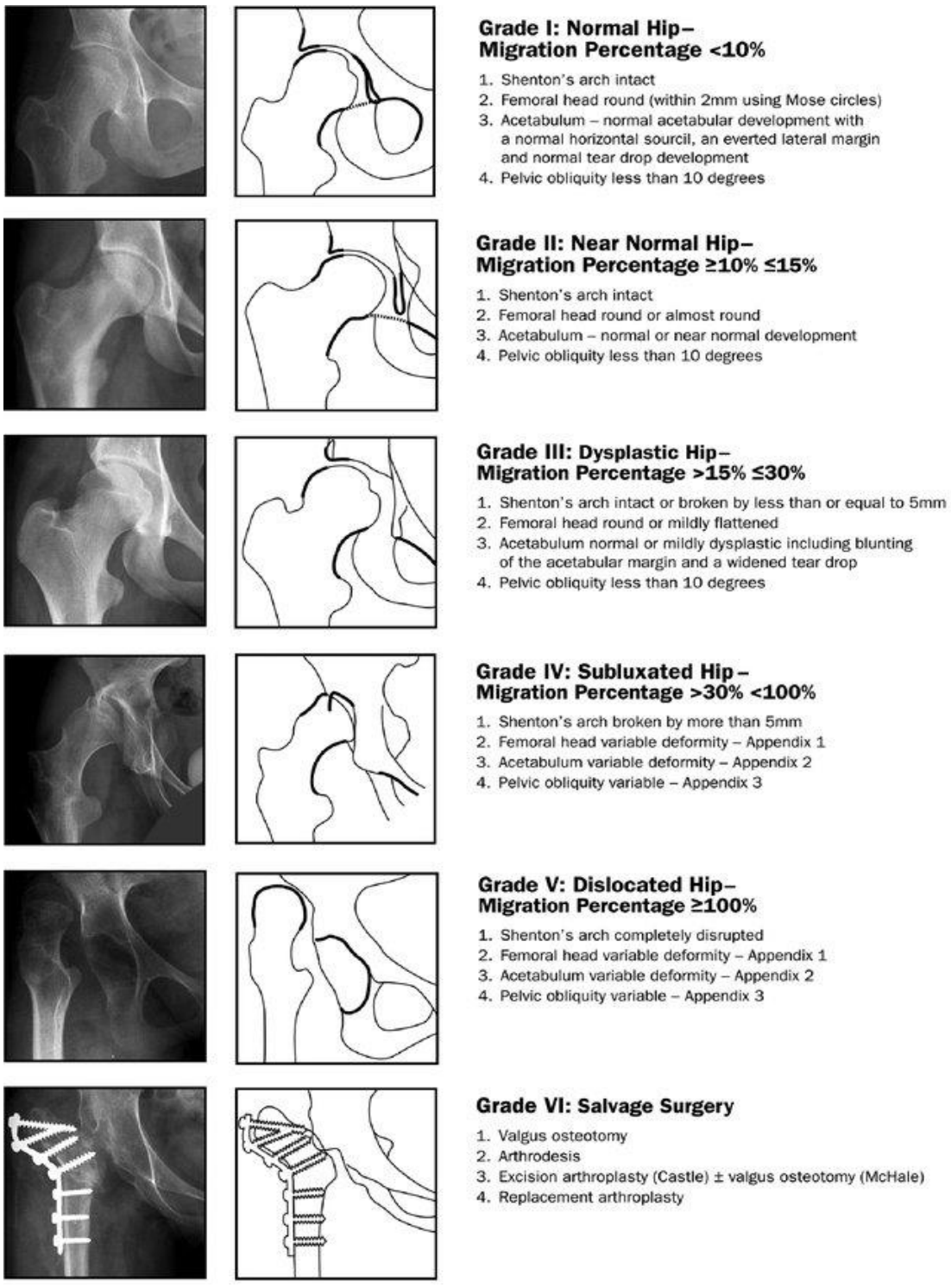
Table 2.1 GMFCS general headings for each level (73)

GMFCS Level	General Description
I	Walks without limitations
II	Walks with limitations
III	Walks Using a Hand-Held Mobility Device
IV	Self-Mobility with Limitations; May Use Powered Mobility
V	Transported in a Manual Wheelchair

Multiple studies classify hips as described by Reimers, where an RMP of 33% or more indicates a subluxated hip, and an RMP of 90% or more indicates a dislocated hip.(31, 45, 55-57) In 2009, Robin et al. expanded Reimers' classification of hip subluxation and dislocation into a more elaborate arrangement encompassing a broader spectrum of hip disease in CP.(45, 58) The Melbourne Cerebral Palsy Hip Classification System is a reliable six-grade ordinal scale that classifies hips in cerebral palsy based on morphology.(58, 74) Each of the six grades corresponds to a basic description of the hip defined by qualitative and quantitative measures (Figure 2.2). The qualitative components of this system include: (1) integrity of Shenton's arch; (2) shape of the femoral head; (3) shape of the acetabulum; and

(4) pelvic obliquity. Quantitatively, hips are assessed using RMP. (58) The Melbourne Cerebral Palsy Hip Classification System was initially developed for skeletally mature children. However, it has now also been validated in children between 2-7 years with open triradiate cartilage.(75) Although the Melbourne Cerebral Palsy Hip Classification System is more extensive and offers more detailed comparisons across different studies, the current popularity and adequacy of Reimers' classification makes it a more reliable choice when classifying hips in CP.(31)

Figure 2.2 The Melbourne Cerebral Palsy Hip Classification System (58)



2.2.7 Management

The management of hip disease in CP can be divided broadly into two main categories: surgical and non-surgical. Treatment for hip disease in CP is usually indicated when patients experience pain, restricted mobility or other symptoms of poor hip instability; however, treatment may also be undertaken prophylactically in order to avoid complete dislocation and complications related to this including osteoarthritis and avascular necrosis of the femoral head.(76) The decision to operate should not be based on radiographic signs of hip disease or osteoarthritis. It should instead be based on the presence of pain or the limitations caused by contractures.(77) Therefore, the primary outcomes of treating hip disease in CP are to alleviate symptoms, improve functionality and prevent dislocation as opposed to restoring the hip to anatomical normality.

2.2.7.1 Non-Surgical Management

Non-surgical management involves monitoring, physical therapy, abduction bracing, and sometimes botulinum toxin A injections.(64) Weight-bearing exercises and postural management programmes have been shown to reduce the severity of hip disease and the need for treatment. However, no studies have reported a complete halt in the progression of CP hip disease by using non-surgical techniques.(78, 79) Furthermore, abduction bracing and spasticity-reducing treatments, such as intrathecal baclofen or botulinum toxin A, are also unsuccessful at reducing the risk and severity of hip disease.(76, 80) However, spasticity-reducing treatments may relieve painful spasms.(81, 82) Although non-surgical management can delay and reduce the severity of hip disease, it is not a definitive treatment. Due to the limited benefits of non-surgical management, surgery is a likely direction taken during the lifetime of a child with CP hip disease.

2.2.7.2 Surgical Management

Surgical management can be divided further into preventative, reconstructive, and salvage surgery. The use of these surgical methods are dependent upon the severity and prognosis of the patient's condition.(81)

Much like non-surgical management, preventative surgery is performed to prevent or delay the progression of a displaced hip to severe subluxation or complete dislocation. It is usually indicated in children with a progressive RMP between 25%-60% with limited hip abduction <30°.(64) Preventative surgery consists of soft-tissue procedures, including adductor releases and an iliopsoas release. An adductor release usually involves myotenotomy of the adductor longus, adductor brevis and gracilis.(29,

76) The iliopsoas release can be performed differently depending on the child's ambulatory status. In non-ambulatory children, the iliopsoas tendon is divided near the lesser trochanter. However, only the psoas tendon is divided in ambulatory children, preserving the iliacus fibres.(64) Additionally, hamstring procedures, such as semitendinosus lengthening, can be performed in non-ambulatory children with limited knee extension.(41, 64) Ambulatory children may also undergo selective dorsal rhizotomy to reduce muscle spasticity.(83) Regardless of the preventative surgical procedure performed, there is still a risk of re-displacement during the growth phase.(76)

Reconstructive surgical techniques are used when the hip is so severely subluxated that it cannot be reversed using preventative surgery alone, represented by an RMP of >60%.(64) Additionally, reconstructive surgery is most effective in children ≥4 years without degenerative changes in the acetabulum or femur. Children <4 years of age have reportedly lost >90% of the correction of the NSA, an angle used to describe the geometry of the proximal femur and hip. Older children tend to have degenerative disease in the hip joint, making it harder to remodel.(84, 85) Reconstructive surgery generally involves osteotomies of the femur and acetabulum.(29) The most common hip reconstruction surgery in children with CP is the proximal femoral varus derotation osteotomy, an osteotomy focusing on the femur. Femoral procedures are indicated in cases where structural deformities of the femur are present, such as coxa valga and increased femoral anteversion. Femoral osteotomies also lengthen the spastic muscle as the bone is shortened.(41) Acetabular procedures are indicated in children with hip dysplasia. The Dega acetabuloplasty is a particularly effective acetabular surgery and is undertaken in all dislocated hips and some subluxated hips.(86-88)

Salvage procedures are indicated when reconstructive surgery either fails or is no longer a viable option due to delayed presentation or degenerative hip disease.(89) Examples of salvage procedures include proximal femoral excision, valgus osteotomy, proximal femoral excision and valgus osteotomy, prosthetic interposition arthroplasty, total joint arthroplasty and arthrodesis.(90-95) The most common salvage procedure is the Castle procedure, a variation of the proximal femoral resection.(81, 90) Valgus osteotomies are used to increase hip abduction in children who have difficulties maintaining hygiene around the perineal area, whilst arthroplasty is performed in ambulatory children with an arthritic hip.(41) Research on salvage surgery in children with CP and hip disease is inadequate, and the type of salvage surgery performed should be decided after evaluating the patient.(96)

2.3 Hip Surveillance Programmes

2.3.1.1 Background

The discovery of the preventability of hip dislocation in children with cerebral palsy was a significant turning point in the prognosis of these children.(97) This principle is the primary purpose behind the development of hip surveillance programmes in children with CP. In 1994, the first hip surveillance programme, named 'Uppföljningsprogram för Cerebral Pares' (CPUP), was started in southern Sweden to detect hip displacement early in the disease and intervene before the hip progressed to dislocation.(56) This prevention programme successfully and significantly decreased the incidence of hip dislocation over ten years.(56)

Following the implementation of CPUP in Sweden, similar surveillance programmes were adopted by other areas, including Norway, Denmark, Iceland, British Columbia, Australia and Scotland.(98-102) Although each scheme is run slightly differently, they all evaluate age and GMFCS to determine the frequency of check-ups and use RMP to assess the severity of hip disease.

2.3.1.2 Benefits of Hip Surveillance

Hip surveillance programmes can significantly lower the incidence of hip dislocation with early detection and intervention.(103, 104) They allow children to have preventative procedures performed at an earlier age and a more viable stage of disease, preventing the occurrence of hip dislocation and the need for reconstructive or salvage procedures.(105) Children who do not participate in surveillance programmes are more likely to develop hip dislocation and undergo surgery at an older age.(106, 107)

2.3.1.3 CPIPS

In the UK, a national hip surveillance programme only exists in Scotland. The CPIPS programme was introduced across Scotland in 2013. This surveillance system provides CP children aged 2-16 years with equal access to a standardised, protocol-based management system that utilises physical and radiological examinations to monitor the risk of hip displacement.(11)

Once a child has been diagnosed with CP, the GMFCS level should be confirmed. At two years of age, the child attends the CPIPS clinic for physical examination every six months until the age of six. Beyond this, the child attends the CPIPS clinic annually until age sixteen, at which age the child is no longer monitored. The frequency of clinics is increased in the presence of 'red values' on the traffic light

system. The traffic light system compares values from a child's physical examination to a pre-defined set of values, described by the three traffic light colours, which correspond to different levels of severity and action (Figure 2.3). Green values do not usually require any action and represent normal or almost normal values; amber values require a review of the individual's current management plan; red values necessitate a referral for further assessment from the orthopaedic department. A hip x-ray is taken in every child with diagnosed CP at the age of two, six and sixteen. The frequency of radiographic evaluation in-between these ages is dependent upon the severity of CP as described by the GMFCS level, and age.(11)

Five years after the implementation of CPIPS, 1646 children were actively monitored through the programme. The prevalence of hip displacement decreased significantly by 55%, and the prevalence of complete dislocation decreased by 48%. Early surgical intervention successfully treated hip subluxation in 94% of the patients who required it.(102) The CPIPS system has now been enrolled across England and the other UK nations, though different areas and regions are at different stages of implementation.(11)

Figure 2.3 CPIPS Traffic Light System for physical examination (11)

GMFCS I – III	RED	AMBER	GREEN
Hip abduction/knee ext	<30°	30° - 40°	>40°
Popliteal angle	>50°	40° - 50°	<40°
Knee extension	<0°		180° /0°
Dorsiflexion with flexed knee	<10°	10° - 20°	>20°
Dorsiflexion with extended knee	0<°	0° - 10°	>10°
Hip internal rotation	<30°	30° - 40°	>40°
Hip external rotation	3<0°	30° - 40°	>40°
Duncan Ely test	<100°	100° -120°	>120°
Hip extension	<10°		>10°

GMFCS IV – V	RED	AMBER	GREEN
Hip abduction	<20°	20° - 30°	>30°
Popliteal angle	>60°	40° - 60°	<40°
Knee extension	>10° fixed flexion	0°/10 fixed flexion	180° /0°
Dorsiflexion with flexed knee	<0°	0° - 10°	>10°
Dorsiflexion with extended knee	<-10°	-10° - 0°	>0°
Hip internal rotation	<30°	30° - 40°	>40°
Hip external rotation	<30°	30° - 40°	>40°
Ely test	<90°	90° - 110°	>110°
Hip extension	<-10°	-10° - 0°	>0°

2.4 Artificial Intelligence

2.4.1.1 Background

AI describes “a system’s ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation.”(108) The notion of an intelligent computer was first described by a British mathematician called Alan Turing. In 1950, Turing, known for breaking the enigma code during World War II, published an article describing how to test the intelligence of a machine. The ‘Turing test’ assesses the capacity of a machine to perform cognitive tasks at a human level.(109-111) The term ‘artificial intelligence’ was first used by a computer scientist called John McCarthy, who used it the name the Dartmouth Summer Research Project on Artificial Intelligence in 1956. Following this eight-week-long conference, the field of AI saw a significant improvement in success for nearly two decades, with machines such as ELIZA, a natural language processing tool capable of conversing with humans, and the General Problem Solver program, which was capable of solving simple problems.(110) Despite the creation of these intelligent machines, the field of AI came to a halt in 1969, as the AI methods were too advanced for the computer processing power available at the time.(112) Since this time, the field of AI has seen many highs and lows.(113) The recent revolution in AI can be attributed to an increase in computer processing power and the availability of data for analysis and learning.(114) Currently, applications of AI can be seen in smart voice-assistant devices, self-driving cars, weather forecasting, Google search and many other devices spanning virtually every field and industry.(115-119)

2.4.1.2 Machine Learning

Machine Learning is a subset of AI that allows computers to learn and make predictions from a dataset without being explicitly programmed to do so.(120-122) Algorithms are the processes by which calculations and problem-solving operations are performed.(123) Algorithms are employed by Machine Learning models to produce an output. The data fed into the Machine Learning model is known as the input, and the predictions or descriptions made by the Machine Learning model are known as the output. There are three main methods of learning that teach Machine Learning models to make predictions or solutions: supervised learning, unsupervised learning and reinforcement learning.(124) It is important to note that algorithms are not exclusive to specific types of learning and can be used in multiple different models, given that the algorithm is compatible with that model.

Supervised learning is the most common type of training and uses labelled data to teach the Machine Learning model. This means that a sample dataset is given containing predetermined input values that

have already been paired with corresponding output values. The machine uses this labelled dataset to learn the pattern and make predictions.(125) In unsupervised learning, the input data is unlabelled, meaning the output data is not known for any values. Therefore, unsupervised learning makes predictions by drawing inferences and finding hidden patterns from the input data. Semi-supervised learning is a method in which a partially labelled dataset is used to train the Machine Learning model. This training method infers the unlabelled values from labelled samples. Semi-supervised models may be used to reduce the cost of labelling data and are often used in image retrieval systems. Reinforcement learning does not learn from a dataset. It uses a trial-and-error feedback system that learns from external responses to make predictions. This model is often used for gaming and navigation.(124, 126)

Supervised learning can be further divided into two broad categories: regressions models and classification models. Regression models and classification models solve different types of problems; as a result, the algorithms used by these models produce different types of data. In regression models, the algorithms map input values to output values containing continuous data. In classification models, the algorithms map input values to output values containing discrete categorical data.(127) Commonly used supervised learning algorithms include Decision Trees (DTs), Naïve Bayes, Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Logistic Regression, K-Nearest Neighbor (KNN) and Random Forests (RFs).(128) Some supervised learning applications include spam detection, predictive analysis and image recognition.(129-131)

Similar to supervised learning, unsupervised learning can be divided two: clustering models and association models.(132) Clustering models and association models use different approaches to produce an output. Clustering models categorise data into groups based on similarities and differences.(133) Association models discover relationships between variables in a dataset.(134) Examples of unsupervised learning algorithms include K-Means Clustering, KNN, Anomaly Detection, ANNs, Principle Component Analysis and Singular Value Decomposition.(135) Use cases of unsupervised learning include market research, pattern recognition, fraud detection and social network analysis, predictive analysis.(136)

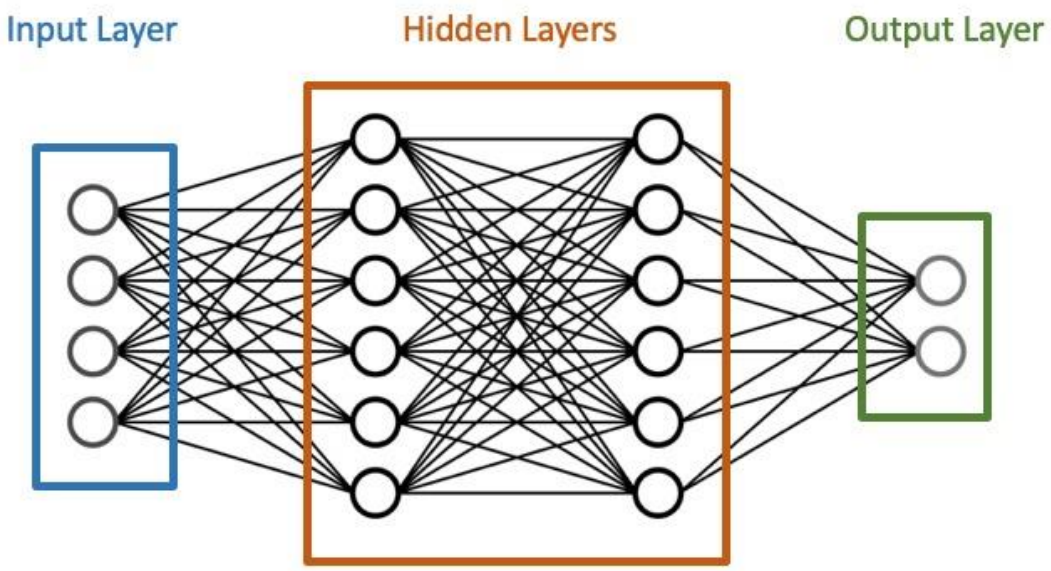
2.4.1.3 Deep Learning

Deep learning is a subset of Machine Learning that uses ANNs with multiple layers to learn from large datasets and make accurate predictions.(137, 138) An ANNs consists of interconnected nodes arranged in multiple layers, imitating neurons in the human brain (Figure 2.4). (139) The input data is first fed into the network's input layer, the information is then processed in the hidden layer, and a corresponding prediction is produced in the output layer. This process is called forward propagation. When a neural

network flows unidirectionally from input layer to output layer, it is known as a feedforward neural network. Each node in the network represents a different feature of the input dataset. A feature is any aspect of the input data that can be measured. A weight is assigned to each node, conveying the importance of its feature and its relationships to the final output or prediction. Therefore, heavier weighted nodes contribute more to the final output of a model.(140)

Deep Learning models consist of multiple hidden layers, allowing more complex operations to occur. Some Deep Learning neural networks can operate bi-directionally through a process called backpropagation. Backpropagation allows a Deep Learning algorithm to increase a model's predictive accuracy over time. By performing error calculations every time a prediction is produced, Deep Learning algorithms can work backwards and adjust the weight of the interconnected nodes to provide a more precise output.(141) This ability to automatically tweak internal parameters allows Deep Learning models to automate the process of feature engineering. Feature engineering, traditionally undertaken by human experts, involves manipulating data to identify features and ranking them by importance before feeding the processed data into a Machine Learning model to produce weighted, accurate predictions and solutions.(142) Therefore, Deep Learning models decrease the need for human intervention and data preparation and produce increasingly accurate predictions and patterns.

Figure 2.4 Artificial neural network model



Various types of neural networks are required for the broad range of Deep Learning applications and data types. Two commonly used Deep Learning neural networks are convolutional neural networks (CNNs) and recurrent neural networks (RNNs).(143) CNNs are feedforward neural networks designed to process pixel data. They can easily detect features and patterns in any given image. Thus, CNNs are mainly used in image analysis and object detection applications.(144, 145) RNNs work differently to traditional feedforward neural networks and CNNs. They have a looping mechanism that acts like a memory, allowing previously learned information from prior inputs to affect current inputs and outputs. As a result of this memory-like mechanism, RNNs are great at processing sequential data such as audio or text and find applications in language translation, speech recognition and generation of image descriptions.(146)

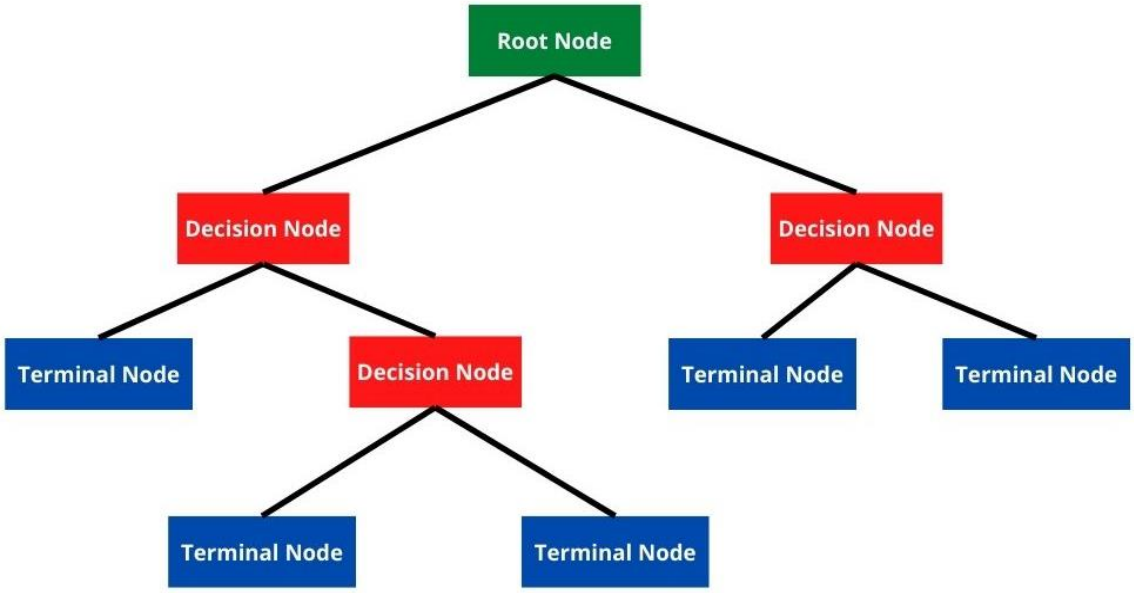
2.4.1.4 Ensemble Learning

Ensemble learning is a strategy of learning that can be applied to multiple Machine Learning models to produce more robust and accurate predictions. In ensemble learning, multiple Machine Learning models are combined to create a single model. The composite output produced by this combined model neutralises the high variance and bias produced by individual models. (147) Once Machine Learning models are part of an ensemble learning model, they are called base learners. Ensemble learning can be achieved via two approaches: bagging and boosting(148, 149) In bagging, also known as bootstrap aggregation, base learners are trained in parallel to each other, but in boosting, base learners are trained sequentially one after another. Bagging reduces the variance in a model, whilst boosting improves the accuracy of a model.(150) Bagging, unlike boosting, uses an equal weight voting system, meaning base learners have an equal weighting in the output of the model. In boosting, each base learner is trained on data that a previous base learner has modified. As a result, base learners do not carry an equal weighting in the function of these models.(151) The most commonly used boosting algorithm is Adaboost.(152) Other boosting algorithms include gradient boosting and XGBoost.(153, 154)

Random Forests are commonly used bagging algorithms that combine multiple DTs to produce an output. The RF model produces a final output from the prediction of the majority of decision trees.(155) A decision tree is a supervised learning algorithm, applicable to regression and classification models, that makes predictions by learning decision rules from labelled training data. Decision rules are applied repeatedly, finding the best split to arrange most, if not all, of the data into a hierarchical flowchart structured like a tree. Decision trees consist of nodes and branches representing all possible decisions and consequences (Figure 2.5).(156, 157) Unfortunately, decision trees have a high variance, meaning small variations in the data can have significant changes in the model.(158) They are also prone to overfitting, which occurs when a model performs too well on training data, resulting in a failure to

perform accurately on new data.(159) Random Forests overcome these issues by using multiple decision trees, producing highly accurate predictions on large datasets and reducing the risk of overfitting.(160) Random Forests can also be used to estimate missing data, as they can maintain high accuracy even when a large proportion of data is missing.(161) The applications of RFs can be seen in bioinformatics, medicine and economics.(162)

Figure 2.5 Decision tree algorithm. The root node starts the algorithm by evaluating a variable that is best able to split the data. Decision nodes occur where another variable is evaluated however, they are not the final output. Terminal nodes are the final node in a tree and represent the potential outputs of the algorithm.



2.5 AI Medical Image Analysis

2.5.1.1 Background

The medical field has seen a surge in the implementation of AI technology.(163) Various applications of AI have been seen across most specialities, including cardiology, respiratory, endocrinology, nephrology, gastroenterology, neurology, oncology, histopathology and medical imaging.(164) These applications can be categorised as either virtual or physical.(165) Virtual applications of AI can be seen in health management systems, disease detection, medical imaging and clinical trial design.(166-168) Physical applications of AI can be seen in robot-assisted surgery, nanobot drug delivery systems and care of the elderly or handicapped.(169-171) In 2004, ANNs were the most commonly used algorithms

in medicine and were used to determine the diagnosis and prognosis of multiple conditions. Specific tasks performed by ANNs include classification of diseases, analysis of radiological images, interpretation of data and prediction of patient survival and outcome.(172) Random Forestss and SVMs were also popular Machine Learning methods in the past. However, in recent years there has been a shift towards CNNs. Currently, CNNs and RFs are the most commonly used algorithms in the medical field.(173)

With the disproportionate growth of medical imaging data in recent years, radiologists have had to adapt by increasing their productivity to unsustainable rates, making errors in judgement an inevitable occurrence.(174-176) The advent of AI in medical imaging has been much needed to meet this increasing workload without errors. The Machine Learning models in this field can be classified into two types: models requiring feature engineering and models not requiring feature engineering.(177) Since Deep Learning models are the only models that do not require feature engineering, they can be used synonymously with 'models not requiring feature engineering'. Deep Learning models, specifically CNNs, are the most commonly used models in medical imaging.(178) This could be due to the reduced need for human intervention and the increasingly accurate predictions they can make. Furthermore, Deep Learning models have been found to perform better than Machine Learning models, which require feature engineering, and can match the performance of trained radiologists.(179-181) These advancements give machines the ability to interpret images, performing tasks such as: abnormality detection; disease classification; monitoring disease progression; and segmentation, which is the process of extracting desired structures from an image.(182-186)

2.5.1.2 Medical Imaging

The use of AI to interpret medical images can be seen in most specialities. The most common application of AI in chest radiographs and CT scans is the detection and classification of lung nodules.(187-190) Other applications include the detection of various pathologies on chest radiographs(191-194) and CT texture analysis associated with pulmonary diseases(195, 196). In brain imaging analysis, applications are mainly seen in magnetic resonance imaging (MRI) scans for the following tasks: classification of disorders such as Alzheimer's disease, Huntington's disease, mild cognitive impairment and schizophrenia(197-201); and detection, classification and segmentation of lesions and tumours(202-206). In ophthalmology, AI is used to detect abnormalities and diseases(207-209) or segment ocular anatomy on colour fundus images(210, 211). In breast disease, AI applications involve the detection, classification and risk calculation of tumours in the breast; modalities include mammograms, tomosynthesis and ultrasounds scans.(212-219) There is a wide range of applications in cardiac imaging; however, left ventricular segmentation in an MRI scan is the most commonly

researched AI application.(220-224) Abdominal imaging applications primarily involve segmentation of the liver, kidneys, bladder and pancreas on CT scans.(225-230) Additionally, applications in the colon include the detection of polyps and colitis on CT scans or during colonoscopy.(231-235) Medical image analysis using AI is not limited to the specialities mentioned above and can be seen in many other specialties such as foetal medicine, dermatology and oncology.(236-241)

2.5.1.3 Musculoskeletal Imaging

The applications of AI in musculoskeletal imaging cover various modalities, including radiographs(242), CT scans(243), MRI scans(244) and dual-energy X-ray absorptiometry (DEXA) scans(245). There are multiple applications of AI in musculoskeletal imaging, from detecting abnormalities to diagnosing and classifying diseases. Studies assess an AI model's performance by calculating the area under the curve (AUC) in addition to accuracy, specificity and sensitivity. The AUC is a summary of the Receiver Operator Characteristic (ROC) curve, which is an evaluation metric used to assess a model's ability to distinguish between binary classes (0 or 1). The ROC curve is a probability curve that is created by plotting sensitivity against the specificity at various threshold values. In studies assessing AI models, the AUC is used to describe a model's diagnostic accuracy and is limited to values between 0 and 1. An AUC of 1.0 indicates a perfect prediction and an AUC of 0.5 indicates an even chance of predicting correctly or incorrectly(246) (Table 2.2).

Table 2.2 Relationship between AUC and Diagnostic Accuracy (246)

AUC	Diagnostic Accuracy
0.9 - 1.0	Excellent
0.8 - 0.9	Very good
0.7 - 0.8	Good
0.6 - 0.7	Sufficient
0.5 - 0.6	Bad
< 0.5	Test not useful

Fracture detection is an important research topic in AI, with a large publicly available dataset containing over 40,000 labelled images to accelerate the development of AI fracture detection tools.(247) Some AI models can detect and classify fractures in multiple anatomical regions such as the ankle, hand, hip, spine, wrist, ulna, femur and humerus, with AUCs ranging between 0.94-1.0 and accuracies ranging between 77% to 98%.(248) They can also detect fractures in the spine and is especially useful for reporting vertebral body compression fractures, which may go unnoticed by radiologists.(249, 250) Previously, a limitation of CNN-based fracture detection models was the non-transferable nature of its learning; each model must be retrained if analysing a new anatomical model.(251) However, recently, an ensemble of 10 CNNs has been used to detect fractures across 16 anatomical sites, with near perfect predictions (mean AUCs, >0.98) in over half of the fracture sites.(252)

Osteoarthritis is a condition in which the implementation of AI would have incredibly beneficial consequences, given the rising incidence of osteoarthritis in England.(253) Currently, AI models can detect hip osteoarthritis on pelvic radiographs achieving an accuracy and precision of 90.2% and 84.7%, respectively.(254) They can also grade the severity of osteoarthritis on knee radiographs, according to the Kellgren-Lawrence grading scale, achieving a diagnostic accuracy equivalent to human experts.(255) The majority of research in this osteoarthritis focuses on the knee, while other locations, such as the hand, have not been explored as much.(256)

Some AI models can also detect abnormalities in the knee joint, including meniscal tears and anterior cruciate ligament ruptures, achieving AUCs as high as 0.97 for meniscal tears and 0.89 and 0.94 for partial and complete ACL tears, respectively.(257, 258)

In orthopaedic oncology, AI models have been developed to automatically detect lytic, sclerotic, and metastatic bone lesions in the spine and sclerotic lesions of the ribs. However, sensitivity values are lower than those in fraction detection models.(259-262) AI can also differentiate between soft tissue tumours, such as lipoma and liposarcoma, on MRI with near perfect predictions (AUC, 0.92) and good accuracy (0.88).(263)

Multiple studies report on developing SVM models that can assess trabecular bone strength using trabecular bone microarchitecture on DEXA scans and MRIs.(264-266) These AI models will act as a framework to automate the assessment of osteoporosis.(267) Recently, AI-interpreted MRI scans have been used in conjunction with Fracture Risk Assessment Tool (FRAX) scores, bone mineral density and physical traits to predict osteoporotic bone fractures.(268)

Bone age has has also been predicted using AI. Some AI models have been developed that can predict skeletal maturity from paediatric hand radiographs to the accuracy levels on par with expert radiologists.(269) The ideal environment for AI to predict skeletal maturity is in combination with the interpretation of a radiologist, which has proven to produce more accurate results when compared to solo interpretations from either AI or radiologists.(270)

2.5.1.4 Pelvic Imaging

Currently, AI models are being built that can automatically detect specific landmarks on images, giving machines the ability to make accurate diagnoses and calculate radiological geometric measurements to assess the severity of the condition. Traditionally, the best Machine Learning model for automatic landmark detection is an RF model, which works exceptionally well for image segmentation.(271-274) RF models have been applied in the automatic segmentation of the pelvis and femur on pelvic radiographs with a success rate of 98%.(275) Recently, CNN models have also been used to detect landmarks in the hip and other anatomical regions.(276-282) Other hip-related tasks performed by AI models can include: severity grading and prediction of hip osteoarthritis(283, 284); detection of total hip replacement prosthesis designs(285); detection and classification of hip fractures(286); segmentation of the hip cartilage, proximal femur and pelvis(287-289); classification of hip dysplasia status(290). Only a few AI models have been developed to calculate measurements from the detected pelvic landmarks.

Developmental dysplasia of the hip (DDH) is a common paediatric hip disorder in which AI has been implemented. Image analysis in DDH is an essential yet taxing task that orthopaedic surgeons undertake.(291, 292) Automating this process will significantly decrease the workload of surgeons and yield more consistent results. A CNN model, built using 10,000 pelvic radiographs, has been reported to detect pelvic landmarks and measure the AcI for the diagnosis of DDH in infants aged one month to six years.(293) Landmarks were detected by first identifying morphological features from the local neighbourhood around the landmark due to the difficulty in identifying landmarks on the misshapen pelvis in DDH. The three landmarks plotted on each hip were the tri-radiate cartilage centre, the acetabulum superolateral margin and the femoral head. Using these landmarks, Hilgenreiner's line and Perkin's line were drawn, and the AcI was calculated within an error of 5°. This CNN model achieved highly precise landmark detection scores and performed better than human experts on illness diagnosis.(293) Another CNN model, developed 11,473 pelvic radiographs, can diagnose DDH by automatically measuring Sharp's angle.(294) Two landmarks on either hip were detected to calculate the measurements, including the lower edge of the teardrop and the outer edge of the acetabulum. The CNN model's performance was similar to that of the surgeons' but required considerably less time. The

Sharp's angle calculated by the model was statistically similar to the surgeons' calculations. The model also achieved a similar diagnostic sensitivity, specificity, and accuracy to that of the surgeons'.(294) Furthermore, a CNN model has been used to calculate the alpha and beta angles on ultrasound scans, with 93% and 85% of estimates being calculated with errors under 5 degrees, respectively.(295)

Total hip arthroplasty (THA) is a surgical procedure that requires preoperative planning to identify various anatomical points and measurements.(296) A CNN model has been used to aid the preoperative THA planning by detecting different landmarks and measuring the centre-edge angle (CEA), NSA and abduction angle.(297) Seven hundred and seven cases were used to train the model. Eleven landmarks were identified on each hip, including the teardrop, upper acetabular rim, ischial tuberosity, lesser trochanter, greater trochanter, femoral head centre, lower acetabular rim, top acetabulum, femoral neck centre, upper femoral shaft axis and lower femoral shaft axis. Measurements were calculated in 0.3 seconds.(297)

A CNN model has also been used to identify the landmarks needed to calculate CEA, Tönnis angle, Sharp's angle, and femoral head extrusion index (FHEI).(298) Of the 1260 images included in the study, 1060 images were used to train the model, and 200 images were used to test the model. The five landmarks identified on each hip include the outermost point of the acetabulum, the innermost point of subchondral sclerosis, the inferior point of teardrop, the femoral head centre and a point on a circle plotted by a dedicated labelling software. Landmark performance was assessed using Percentage of Correct Key points (PCK), which shows the percentage of predicted landmarks that fall near the reference radiologists' landmarks. The model produced a PCK range of 87%-100% in a 3mm radius region around the reference radiologists' landmarks. The measurement calculations of the model and expert radiologists correlated well and were consistent between the two groups across all four measurements.(298)

2.5.1.5 Cerebral Palsy Hips

Children with CP often have pelvic radiographs to assess the severity of hip disease further. The interpretation of hip radiographs in CP involves taking manual measurements, such as RMP, which can be inconsistent. Little research has been conducted on the automatic interpretation of radiographs in CP. Only one article reports on the development of a model that can assess hip radiographs in CP by measuring RMP.(299) In this study, the CNN model was trained using a small dataset with limited labelled data. Sixty-seven radiographs were used for training and validation, and 55 radiographs for testing. However, a fine-tuning method using iterative weight updates was used to compensate for the limited dataset. The model detected four landmarks on each hip: the rostral aspect of the open triradiate

cartilage through Hilgenreiner's line (A_1, A_2), the lateral margins of the acetabulum through Perkin's lines (B_1, B_2), and the lateral and medial edges of femoral head ossific nucleus perpendicular to Hilgenreiner's line (C_1, C_2, D_1, D_2). Using these landmarks, the model is able to calculate the RMP and classify the hips in two ways. The hips were classified as both normal or displaced using a threshold RMP value of 30%, and as requiring surgery or not requiring surgery using a threshold RMP of 40%. The classification of hip displacement achieved a sensitivity and specificity of 87.8% and 93.4%, whilst the classification of requiring surgery achieved a sensitivity and specificity of 63.2% and 94.5%. The model also has good reliability achieving an intra-class correlation coefficient of 0.91 when comparing the model's predicted RMP to that of the raters in the study. Measurements were calculated within 5 seconds.(299)

2.6 Summary

Hip disease in CP is a generic term that can be used to describe hip displacement, hip migration, hip subluxation, hip dislocation and hip dysplasia. Hip subluxation is a likely occurrence in this disease and left untreated can result in hip dislocation. This can decrease the quality of life in children with CP and hinder them from carrying out daily activities. Various measurements have been used to assess the severity of hip displacement on a radiograph. A standard set of measurements does not exist, but RMP has been regarded as the gold standard measurement. This condition can be classified using the GMFCS, and CP hips can be classified using either Reimers' classification system or the Melbourne Cerebral Palsy Hip Classification System. Management of CP hip disease can be split into non-surgical and surgical. The three different types of surgical procedures are preventative, reconstructive and salvage.

Hip surveillance programmes were introduced to detect hip displacement early in the disease. This allows children with CP to have interventions earlier and prevent severe disease and dislocation. The first country to introduce a national hip surveillance programme was Sweden. Since then, multiple other countries have adopted a similar programme, including Scotland.

The term 'AI' is a broad and describes any machine that can interpret data, learn from that data, and achieve a goal based on that knowledge. Machine Learning is a subset of AI that can be trained via supervised, unsupervised or reinforcement learning. Each type of learning may use different algorithms to create a Machine Learning model; however, algorithms are not exclusive to a specific learning method. Deep Learning is a subset of Machine Learning that utilises algorithms called ANNs, allowing models to achieve complex tasks without human interaction. CNNs are one type of Deep Learning model that can be used to analyse images. Ensemble learning is a strategy that can be used to combine

multiple Machine Learning models to increase accuracy and decrease bias. Random Forests are a commonly used ensemble model that utilises algorithms called DTs to produce highly accurate results even when data is missing.

Artificial Intelligence models can interpret medical images by performing tasks such as detection, classification, monitoring and segmentation. Applications of AI can be seen in most specialties, including neurology, respiratory, ophthalmology, breast disease, cardiology and gastroenterology. In musculoskeletal imaging, AI has been used to: detect and classify fractures; detect and classify osteoarthritis; detect abnormalities in the knee joint; detect bone tumours; assess bone strength and osteoporosis; and predict skeletal maturity. AI has also been used to detect various landmarks on hip radiographs allowing models to calculate measurements such as AcI, Sharp's angle, CEA, NSA, abduction angle, Tönnis angle and FHEI. Recently, a CNN model has been developed that can detect RMP. However, little research has been conducted regarding the automatic calculation of critical measurements in CP hips. Therefore, it is important to explore the development of different AI models to automate hip radiograph assessment and the calculation of measurements. In order to build a clinically useful model, the key measurements that need to be measured when assessing CP hip radiographs must first be identified. In the next chapter, a systematic review is performed to first form a comprehensive list of the measurements that are used to assess hip disease in CP.

Chapter 3: A systematic review of the radiographic measurements used in the assessment of cerebral palsy hips

3.1 Introduction

Multiple radiographic measurements can be used to assess the geometry of various aspects of the hip. In cerebral palsy (CP), these measurements can be used to help determine the severity of disease and guide decisions regarding treatment.(300) Currently, the evaluation of CP hip radiographs vary greatly across clinical practice and research, owing to the vast selection of radiographic measurements. This heterogeneity can restrict the effectiveness of comparisons that can be drawn between studies, increase the risk of reporting bias and decrease the relevance and effectiveness of research and clinical practice.(301) With the recent success of hip surveillance programmes(56, 98-102), there is an increased need for accurate methods of radiographic evaluation and an increased amount of data available for research. These issues highlight the need to establish a standard set of measurements for the reporting of CP hip radiographs. This will encourage uniformity in the assessment of CP hip radiographs and support collaboration between centres. Furthermore, a core set of measurements will guide the development of an artificial intelligence (AI) system for the automatic evaluation of hip radiographs by enabling researchers to focus on the most important measurements.

In order to form this core measurement set (CMS), a comprehensive list of all the different radiographic measurements must first be identified from the literature, after which a Delphi study is conducted to select the most important measurements for the CMS. To date, one systematic review has assessed the validity and reliability of common radiological measurements used to assess CP hips, such as Reimers' migration percentage (RMP), acetabular index (AcI), femoral neck-shaft angle (NSA), femoral head-shaft angle (HSA) and femoral anteversion, using several imaging modalities including radiography, computerised tomography (CT), magnetic resonance imaging (MRI) and ultrasound.(302) However, no studies have identified a comprehensive list of the radiographic measurement reported in the literature. The aim of this chapter was to systematically identify these measurements for use in a Delphi survey, which will form a CMS. This CMS will then be used to inform the development of an artificial intelligence (AI) system which will automatically assess the hip radiographs in children with CP.

3.2 Methods

A systematic literature review was performed per the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.(303) The following electronic databases were searched for published literature: PubMed, SCOPUS and Web of Science (Appendix 1). The search strategy is included in the supplementary material. Relevant papers published in the English language from 1st January 2011 onwards were searched.

3.2.1 Eligibility criteria

Studies were eligible if they met the following criteria:

Inclusion criteria

1. Studies focussing on hip disease in cerebral palsy.
2. Studies using measurements to assess pelvic radiographs.
3. Studies with a patient population:
    a. ≥ 50 study participants
    b. aged ≤ 18 years
4. Case series, retrospective studies, cross-sectional studies, cohort studies or randomised control studies.
5. Studies must be published in the English language.
6. Published from 1st January 2011.

Exclusion criteria

1. Studies solely focussing on the diagnostic accuracy or reliability of a measurement.
2. Studies without a full text article and only an abstract.

3.2.2 Study selection process

The review team consisted of three individuals: Mr Prince Josiah Sajanthan Joseph (PJSJ), MPhil student at the University of Liverpool; Miss Sundus Tahir Masudi (STM), Medical Student at the University of Liverpool; and Mr Mohammed Khattak (MK), Academic Clinical Fellow in Trauma and Orthopaedic Surgery at the University of Liverpool. Articles from the online databases were downloaded to Endnote X9, and duplicate articles were checked for and excluded.(304) The deduplicated articles were exported to Rayyan, and the title and abstract of each article were independently screened by two reviewers (PJSJ and STM) according to the inclusion/exclusion

criteria.(305) Disagreements between the reviewers were resolved via discussion. If a resolution was not reached, the article was included for full-text analysis.

Articles remaining after the screening process were reviewed in the full-text format against the inclusion/exclusion criteria to confirm eligibility. Articles excluded through this process were not recorded, with reasons provided. Any disagreements during the full-text analysis were resolved by consulting the third reviewer (MK).

3.2.3 Data extraction

The following data fields were independently extracted from each article by two authors (PJSJ and STM): title, author, year, journal of publication, location, study type, population size, number of hips studied, age, sex, duration of follow-up, measurement used, time point of measurement, verbatim definition of measurement, use of visual explanations for measurements and primary intervention. A study was considered to have defined a measurement if it provided a definition of the measurement or a visual explanation.

3.2.4 Quality assessment and risk of bias

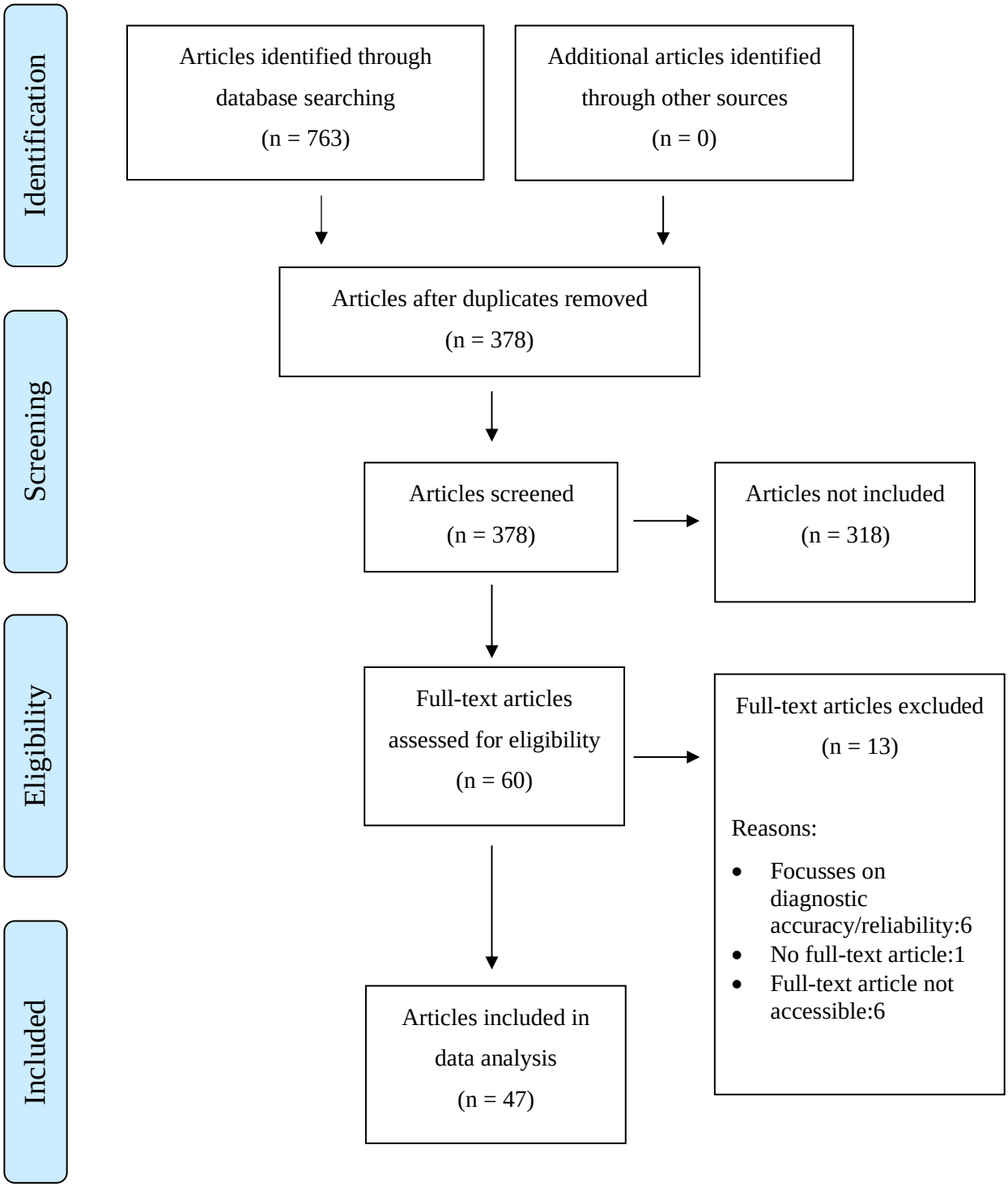
The purpose of this systematic review was to identify all the radiographic measurements reported in the literature irrespective of study quality; therefore, an appraisal of the methodological quality of the studies or risk of bias assessment was not undertaken. Although this review was submitted to PROSPERO, it was deemed out of scope and has therefore not been registered on their database.

3.3 Results

3.3.1 Literature Search Results

The initial literature search identified a total of 763 articles. After removing duplicates, 378 abstracts were screened using the eligibility criteria, after which 60 articles were further assessed for eligibility using the full-text. Six of the 60 papers were immediately excluded as the full-text article was not accessible (Appendix 2). In total, 47 of the 60 articles remained for the final analysis (Appendix 3). The PRISMA flow diagram for the selection process is presented in Figure 3.1.

Figure 3.1 PRISMA flow diagram



3.3.2 Study characteristics

There were 29 retrospective studies, nine prospective studies and nine cross-sectional studies. The most common intervention being studied was surgery (n=22). Other primary interventions studied include hip surveillance programmes (n=18) and physiotherapy (n=1).

Over half of the studies were conducted in Europe (n=25); however, studies were also completed in Asia (n=9), North America (n=9), Australia (n=3) and South America (n=1).

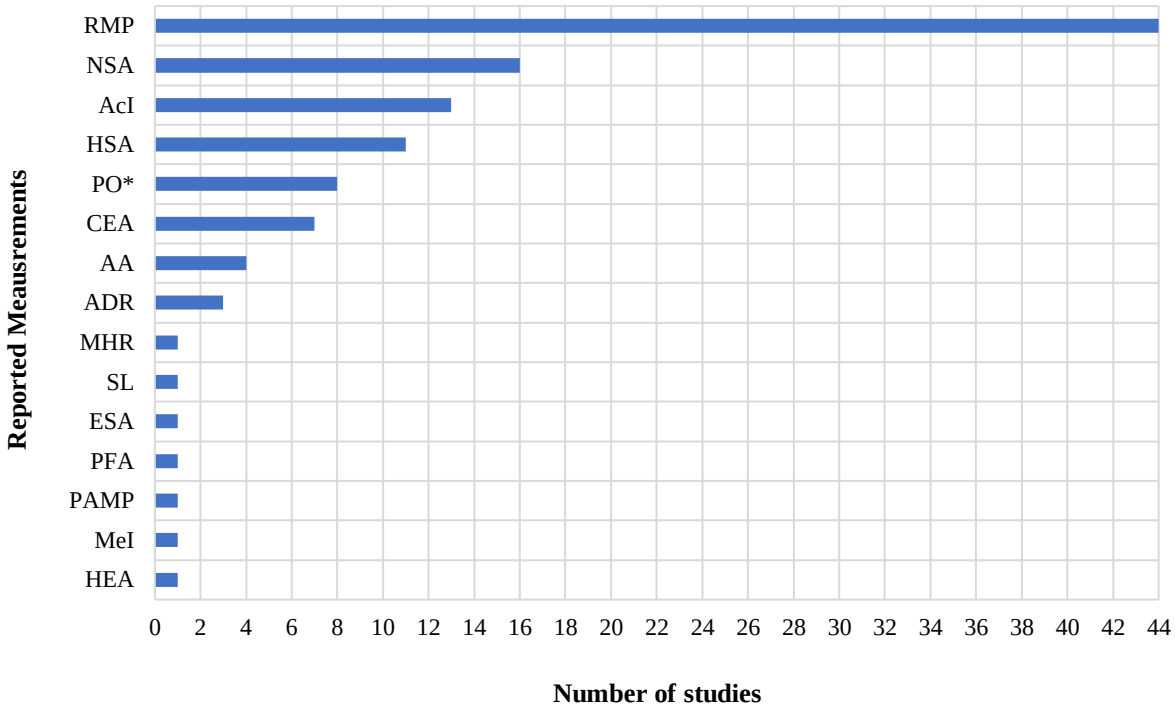
The number of CP patients included per study ranged between 50 and 1171 patients. The average age of the study population ranged between 2.0 and 14.6 years. Most studies had a higher proportion of males ranging from 44% to 75% of the patient population. The average duration of follow-up ranged between 1.2 and 12.8 years.

Of the studies where surgery was the primary intervention, 95% took radiographic assessments before and after surgery (n=21), with one study also taking radiographic images intraoperatively. Of the studies with hip surveillance programmes as the primary intervention, 50% of studies analysed the radiograph taken at the beginning of the programme or the latest radiograph available at the time of the study (n=9). The full list of study details and patient characteristics can be found for each study in Appendix 4.

3.3.3 Measurements

Fifteen distinct measurements were identified from the 105 reported measurements obtained across 47 studies. The median number of measurements reported per study was two (range=1-5, IQR=2). Appendix 5 lists all the measurements reported for each study. RMP (n=44, 94%) was the most common measurement. Other reported measurements include NSA (n=16, 34%), AcI (n=13, 28%), HSA (n=11, 23%), pelvic obliquity (PO) (n=8, 17%)*, centre edge angle (CEA) (n=7, 15%), Sharp's angle or acetabular angle (AA) (n=4, 8.5%), acetabular depth ratio (ADR) (n=3, 6.4%), Mose hip ratio (MHR) (n=1, 2.1%), Shenton's line (SL) (n=1, 2.1%), epiphyseal shaft angle (ESA) (n=1, 2.1%), pelvic femoral angle (PFA) (n=1, 2.1%), pelvic adjusted migration percentage (PAMP) (n=1, 2.1%), medialization index (MeI) (n=1, 2.1%), and Hilgenreiner epiphyseal angle (HEA) (n=1, 2.1%). The list of identified measurements are presented in Figure 3.2.

Figure 3.2 Total number of studies per measurement. Forty-seven studies were reviewed, identifying 15 distinct measurements. RMP was the most common measurement.



None of the identified measurements were consistently reported across every study. Measurements were undefined in 60% of studies (n=28). Definitions of measurements were homogeneous across all studies that provided one (Appendix 5).

3.4 Discussion

This systematic review is the first to identify all the different measurements that have been reported to assess hip radiographs in children with CP in the last ten years. Across the 47 included studies, 15 distinct measurements have been identified. Of these 15 measurements, RMP is the most common, appearing in 44 studies (96%). For comparison, this is 28 more studies (60%) than the second most common measurement, highlighting the dominant use of RMP over other radiographic measurements. This suggests a widespread agreement on the use of RMP; however, there is little agreement beyond this.

3.4.1 Radiographic Measurements

RMP has been referred to as the gold standard for the assessment of hip displacement in CP.(53, 54, 306) The results from this study support this opinion with RMP appearing in the majority of included studies. Furthermore, Shore et al. found RMP to be the most reliable measure when compared to AcI and AA, obtaining excellent inter-and intra-rater reliability.(52) Craven et al. found RMP to be high reliable when measured in young preschool-aged children.(53) Additionally, studies have shown that rater experience has no effect on the calculation of RMP, further supporting its reliability and repeatability.(307, 308) Thus, it can be concluded that RMP is a robust measurement that is widely used in the assessment of CP hip disease and it is highly likely to be included in the CMS.

Other potential candidates for the CMS include NSA, AcI and HSA. These were the only other measurements, beyond RMP, that were included in over 20% of studies. Although NSA, AcI and HSA are reported in significantly fewer studies than RMP, they assess different aspects and deformities of the hip in comparison to RMP and can therefore be used alongside RMP as supplement measurements. NSA or HSA can be beneficial when coxa valga deformities are present,(46, 50) whilst AcI can be useful when deciding whether or not to surgically reconstruct the acetabulum in the presence of hip dysplasia.(2, 309) It has been reported that AcI is insufficient at predicting hip displacement(310) and has fewer studies researching it as a primary objective in the context of CP hips when compared to NSA and HSA. This suggests that AcI is of lesser importance and may therefore be the least likely measurement to be included in CMS from the potential candidates identified. Lee et al. found that correlation with RMP was higher in NSA than HSA, and concluded that NSA is a more clinically relevant measurement in the assessment of proximal femur deformities in CP when compared to HSA.(311) This is finding is reflective of how widespread the reporting of these two measurements are, with NSA appearing in five more studies (11%) than HSA. These findings suggest that NSA is more likely to be included in the CMS than HSA.

Head-Shaft Angle has some controversy relating to its prognostic value in hip displacement and studies have reported conflicting results. Hermanson et al. report that a high HSA is a risk factor for hip displacement in children with CP and developed a risk score called the Uppföljningsprogram för cerebral pares (CPUP) Hip Score, which predicts the risk of hip displacement in children with CP to a high discriminatory accuracy.(51, 312) However, Chougule et al. did not find a correlation between HSA and RMP and concluded that the use of HSA in routine x-rays does not add any value.(306) Terjesen et al. report that HSA is not a predictor of hip displacement in children under five years of age; however, after the age of five, an association between HSA and hip displacement can be seen.(313) Conversely, van der List et al. report that HSA is a valuable predictor of hip displacement before the

age of four but not after.(314) Head-Shaft Angle is a well-researched measurement; however, its importance in the assessment of CP hip disease remains unresolved.

It seems unlikely that measurements reported in fewer studies (i.e. under 20%) will be included in the CMS, as this suggests that researchers do not regard these measurements with high importance. This may be especially true for measurements that were only reported in one study, such as MHR, SL, ESA, PFA, PAMP, MeI and HEA.

3.4.2 Rationale for eligibility criteria

An inclusion criterion of 50 or more participants was established to ensure that only large studies were included. This ensured that small studies focusing on newly proposed or unvalidated measurements were not included and limited the number papers that had to be screened, allowing this systematic review to be completed within the project timeline. An age range of ≤ 18 years was established due to the variation in the age of inclusion across the literature. A systematic review assessing the effectiveness of hip surveillance programmes and factors associated with the progression of hip displacement, reported that most studies focussed on children and adolescents up to 18 years, half of the studies focussed on younger children under 12 years, and a few studies focussed on young adults over 18 years.(101) Therefore, having an upper limit of 18 years ensured that the review broadly encompassed studies of "children" and increases the generalisability of the CMS. Additionally, only papers published in the English language were included as the research team did not have a reliable means to accurately translate articles written in other languages and foreign terminology for the measurements may lead to ambiguity regarding the identity of those measurements. Furthermore, only articles published in the last 10 years were included, as the aim of this systematic review was the identify all the different measurements that are currently used.

3.4.3 Strengths and limitations

Strengths of this review include a broad search strategy used to identify all relevant studies, the use of two reviewers during the study selection and data extraction process, reducing human error and rater bias. However, limitations arise from the restricted full-text search and the eligibility criteria. Only libraries accessible through the University of Liverpool were used to gain access to full-text articles; other methods such as inter-library loans were not used, limiting the number of full-text articles analysed in the review. Furthermore, the inclusion criteria included a participant number that was 50 or more and required articles to be published in the English language. These two criteria exclude a vast number of articles that may have been included if there was a longer time frame to complete the review

and a means to accurately translate non-English language studies. However, although this limitation may affect the extent of the measurements identified in this review, it is unlikely to affect the formation of the CMS as important measurements are likely to have been used in studies with 50 or more participants published in the last 10 years. Additionally, a Delphi process has been planned to identify important measurements that have not been established through this systematic review.

Another limitation of this study arises from the uncertainty regarding the exact identity of certain measurements from studies which had a poor quality of reporting. A lack of clear definitions or visual explanations to accompany a reported measurement restricted the ability of the study authors to consistently identify distinct measurements across all studies. Therefore, assumptions were made based on the objectives of the study and the application of the measurement in question in order to determine undefined and poorly defined measurements as either identical or distinct from other measurements. Furthermore, the citation given for each measurement was compared between studies to identify duplicate measurements with different names. If citations did not match, the references of the cited paper were searched further until a well-informed decision could be made. These issues made it difficult to resolve if poorly reported measurements were duplicates of other measurements or separate measurements with subtle variations (i.e. HSA and ESA). Of the 15 identified measurements, only RMP had consistent and clear definitions and citations. Beyond this, neither a consistent definition nor a consistent citation was used for a measurement. To avoid exclusion of any measurements, poorly defined measurements were treated as separate measurements and will be included in the Delphi survey.

Future studies should assess the quality of reporting and risk of bias to highlight areas for improvement when reporting radiographic measurements of the hip in CP. Another limitation of this study arises from the selection of articles published only in the English language, which could account for why most studies were conducted in Europe and North America.

3.4.4 Recommendation

A formal Delphi study should be conducted to determine the opinions of orthopaedic surgeons and physiotherapists on the most important measurements in the assessment of CP hip disease, following the COMET (Core Outcome Measures in Effectiveness Trials) guidelines(315). This systematic review has identified 15 candidates to form the CMS.

3.5 Conclusion

This is the only systematic review to identify and summarise all the different measurements used to assess hip radiographs in children with CP. Reimers' Migration Percentage is the most common measurement and is likely to be in the CMS. NSA, AcI and HSA are also reported often and are good candidates for the core set. A Delphi study is needed to create a CMS, which will inform the development of an AI system for the automatic analysis of CP hips. This chapter has achieved the initial objectives to systematically identify all the radiographic measurements used to assess hip disease in children with CP.

Chapter 4: Development of a Core Measurements Set for the assessment of Cerebral Palsy Hips: a Delphi study

4.1 Introduction

The radiographic assessment of cerebral palsy (CP) hip disease involves using measurements to help quantify the severity of the condition. In the previous chapter, 15 measurements were identified in the literature. Of these measurements, Reimers' Migration Percentage (RMP) is the most commonly reported measurement appearing in over 90% of studies. It has also been referred to as the 'gold standard' measurement for the radiographic assessment of hip displacement and is embedded in the radiograph protocol of every hip surveillance programme.(54, 56, 99-102, 316) However, the reporting of RMP is not standardised across all studies, and the recording of additional measurements alongside RMP differs between studies and surveillance programmes. A core set of measurements for the standardised assessment of CP hip radiographs in research and clinical practice has not yet been developed. Without a core measurement set (CMS), comparisons cannot be readily made between studies or surveillance programmes.

The term 'core measurement set' has been coined from the term 'core outcome set' (COS). The Core Outcome Measures in Effectiveness Trials (COMET) handbook has defined a COS as "an agreed standardised collection of outcomes which should be measured and reported, as a minimum, in all trials for a specific clinical area".(315) A CMS serves a similar purpose, consisting of measurements as opposed to outcomes; however, its application can be extended to clinical practice and trials due to its diagnostic nature. Therefore, a CMS can be defined as an agreed standardised collection of measurements which should be recorded, as a minimum, in all clinical trials and assessments for a specific clinical area. The formation of a CMS will resolve the heterogeneity in the reporting of measurements whilst ensuring the inclusion of essential measurements.

This chapter aims to form a CMS for the assessment of hip radiographs in children with CP, which can serve as a minimum requirement in clinical studies and hip surveillance programmes. A CMS can facilitate multicentre research to be conducted and uniformity in the monitoring of CP children. The CMS can also serve as a framework for developing software that automatically analyses hip radiographs in children with CP. Building this software around the most important measurements will improve its clinical impact and facilitate implementation into hip surveillance programmes.

4.2 Methods

The Delphi method was used to form the CMS.(317) This consists of two stages: (1) a two-round Delphi survey to score the identified measurements based on importance; and (2) a final consensus meeting to discuss any remaining measurements that have not reached consensus and form the CMS. The Delphi survey was created using the 15 measurements identified in the previous chapter, including RMP, femoral Neck-Shaft Angle (NSA), acetabular index (AcI), femoral head-shaft angle (HSA), pelvic obliquity (PO)*, centre-edge angle (CEA), Sharp's angle or acetabular angle (AA), acetabular depth ratio (ADR), Mose hip ratio (MHR), Shenton's line (SL), epiphyseal shaft angle (ESA), pelvic femoral angle (PFA), pelvic adjusted migration percentage (PAMP), medialization index (MeI) and Hilgenreiner epiphyseal angle (HEA). The study was conducted and supervised by a dedicated steering committee consisting of: Mr Prince Josiah Sajanthan Joseph (PJSJ), MPhil student at University of Liverpool; Mr Mohammed Khattak (MK), Academic Clinical Fellow at the University of Liverpool; and Professor Daniel Perry (DP), Professor & Honorary Consultant Orthopaedic Surgeon at Alder Hey Children's Hospital and University of Liverpool.

4.2.1 Participants

The Delphi study was conducted with orthopaedic surgeons and physiotherapists with a specialist interest in CP. Participants were eligible to take part from any country, and they were recruited via the investigators' networks and newsletters disseminated by the British Society for Children's Orthopaedic Surgery (BSCOS).(318) Participants were only allowed to take part in round two of the Delphi survey if they participated in round one. Invitations for round two were sent via personalised emails to participants from round one. Eligible participants were asked to complete an online Delphi questionnaire using Microsoft Forms (Microsoft, Redmond, USA) (Appendices 6 and 7). Twenty participants were sought to take part as a minimum. PJSJ was responsible for sending out invitations and creating the online Delphi questionnaire, which was reviewed by MK and DP.

4.2.2 Delphi process and Definitions

4.2.2.1 Overview

Participants were given four weeks to complete each round. Participants who failed to complete the online questionnaire by the end of week two were sent a personalised follow-up email, if possible, to reduce attrition rates. Another personalised email was sent at the end of week three if the participant

had still not completed the questionnaire. Failure to complete the questionnaire during the four-week window resulted in exclusion from participation in further rounds. Follow-up emails were sent by PJSJ.

Each measurement was scored in the questionnaire using the Grading of Recommendations, Assessment, Development and Evaluations approach.(319) A 9-point Likert scale was used to score the listed measurements, with 1–3 considered 'not important', 4–6 considered 'important but not critical', and 7–9 considered 'critically important'.

4.2.2.2 Definition of Consensus

Consensus definitions were based on the '70/15%' consensus framework described in the COMET handbook: version 1.0.(315) For a measurement to be included in the CMS, the majority of participants (>70%) should score the measurement in question as being 'critically important', with only a small minority (< 15%) considering it to be 'not important'. For a measurement to be excluded from the CMS, the majority of participants (> 70%) should score the measurement in question as ‘'not important', with only a small minority (<15%) considering it to be 'critically important'. Any measurements that did not reach the consensus threshold were considered equivocal and were discussed in a final consensus meeting. The threshold for consensus was predefined to avoid bias once the responses have been collected.

4.2.2.3 Delphi Round One

Participants were asked to score the list of measurements followed by an optional free-text section where any additional measurements not currently listed could be suggested for round two. Results from round one were checked for measurements that had reached consensus for exclusion from round two. Additional measurements that were suggested by participants were assessed by the steering committee (PJSJ and DP) for any duplicate measurements that were synonymous for measurements that were already included in the survey.

4.2.2.4 Delphi Round Two

Participants were presented with anonymised data from round one, including the number of respondents, distribution of scores for all listed measurements and measurements that had reached consensus in round one. The remaining equivocal measurements were rescored acknowledging the information from round one, and newly added measurements were scored for the first time. Results from round two were analysed to identify measurements that had reached consensus. Each measurement

from both rounds was classed as 'consensus in', 'consensus out' or 'no consensus' per the consensus criteria.

4.2.2.5 Consensus Meeting

An online consensus setting exercise, chaired by PJSJ, was completed by the Consensus Focus Group, consisting of participants from the Delphi survey with expertise in CP hip disease and the steering committee (PJSJ and DP). The Consensus Focus Group formed a final CMS following a discussion regarding the results from round two with subsequent voting. A Nominal Group Technique was used to discuss the equivocal measurements, allowing individuals in the Consensus Focus Group to discuss and justify opposing views to make an informed decision.

4.3 Results

4.3.1 Participant Characteristics

The 15 measurements identified in the systematic review were presented to 22 participants in the first round of the Delphi process, including 21 orthopaedic surgeons (95%) and one physiotherapist (4.5%). Participation came from five countries, including the United Kingdom (UK) (n=17, 77%), United States of America (USA) (N=2, 9.1%), Netherlands (n=1, 4.5%), India (n=1, 4.5%) and Thailand (n=1, 4.5%). Of these 22 participants, one participant (4.5%) did not also take part in the second round of the Delphi process. Participation in round two was also received from five countries, including the UK (n=17, 81%), USA (N=1, 4.8%), Netherlands (n=1, 4.8%), India (n=1, 4.8%) and Thailand (n=1, 4.8%). Twenty orthopaedic surgeons (95%) and one physiotherapist (4.8%) took part in the second round. Table 4.1 shows the number of participants taking part from different participant groups. The median time for survey completion was 2 minutes and 46 seconds in round one and 4 minutes and 19 seconds in round two.

Table 4.1 Number of participants taking part in the Delphi survey.

Participants group	Round 1	Round 2
Orthopaedic surgeons	21	20
Physiotherapists	1	1
UK	17	17
USA	2	1
Netherlands	1	1
India	1	1
Thailand	1	1

4.3.2 Round One and Two

Figure 4.1 presents a summary of the Delphi responses received over the two rounds. At the end of round one, two of the 15 measurements had reached consensus. RMP was voted to be included in the CMS, and MHR was voted to be excluded from the CMS. Additionally, two new measurements were suggested for at the end of round one: Sourcil Tönnis angle (STA) and Femoral head shape/congruency (FHS). None of the suggested measurements were duplicates of an already included measurement; therefore, both STA and FHS were included in round two. Furthermore, it was suggested by one of the participants that pelvic obliquity, which was initially included in this study, should excluded as it is not used to measure hip migration in CP and is only considered if the hips are not in the correct position in the radiograph. The steering committee agreed with this opinion; therefore, pelvic obliquity was prospectively excluded from future rounds of the Delphi.

In round two, 14 measurements were presented, excluding the two measurements that reached consensus and the one measurement that was removed upon review, and including the two newly suggested measurements. Of these 14 measurements, five measurements reached consensus. Measurements that reached consensus to be excluded from the CMS include CEA, ADR, HEA, PFA and MeI, and no measurements reached consensus for inclusion in the CMS. After the two rounds of the Delphi process, one measurement was voted for inclusion in the CMS, six measurements were voted out of the CMS, and nine measurements did not reach consensus. The full set of anonymised responses for each round are presented in Appendices 8 and 9.

4.3.3 Final Consensus Meeting

The nine remaining equivocal measurements were discussed in the final consensus meeting. Of these nine measurements, one measurement was included in the CMS, and all other measurements were excluded from the CMS. The consensus group decided that HSA should be included in the CMS. There was a debate between HSA and NSA, as both were broadly identified to be important throughout the Delphi study. However, NSA is known to be unreliable and harder to reproduce, given the influence of hip rotation.(320) Furthermore, HSA has greater utility as part of the risk calculation in hip displacement using the Uppföljningsprogram för cerebral pares (CPUP) Hip Score.(51) Neck-Shaft Angle and AcI were discussed as potential candidates for inclusion but were eventually voted out of the CMS. NSA was eventually excluded as it was considered unreliable due to the significant influence hip rotation has on the accuracy of the measurement. Although, AcI was recognised as a useful measurement, it was ultimately excluded from the CMS as it was not considered critically important.

The final consensus group reasoned that when any signs of dysplasia are present, surgeons are likely to operate on the acetabulum without needing to quantify the severity of acetabular coverage using AcI. Therefore, it was concluded that AcI is not a necessary measurement, but it would be advantageous to build an AI system measuring AcI for future research. Overall, of the nine equivocal measurements, HSA was the only measurement that was included in the CMS. At the end of the consensus meeting, the final CMS consisted of RMP and HSA (Figure 4.2).

Figure 4.1 Summary of Delphi responses over two rounds. A total of 16 measurements were scored over the two rounds with RMP reaching 'consensus in' and MHR, CEA, ADR, HEA, PFA and MeI reaching 'consensus out'.

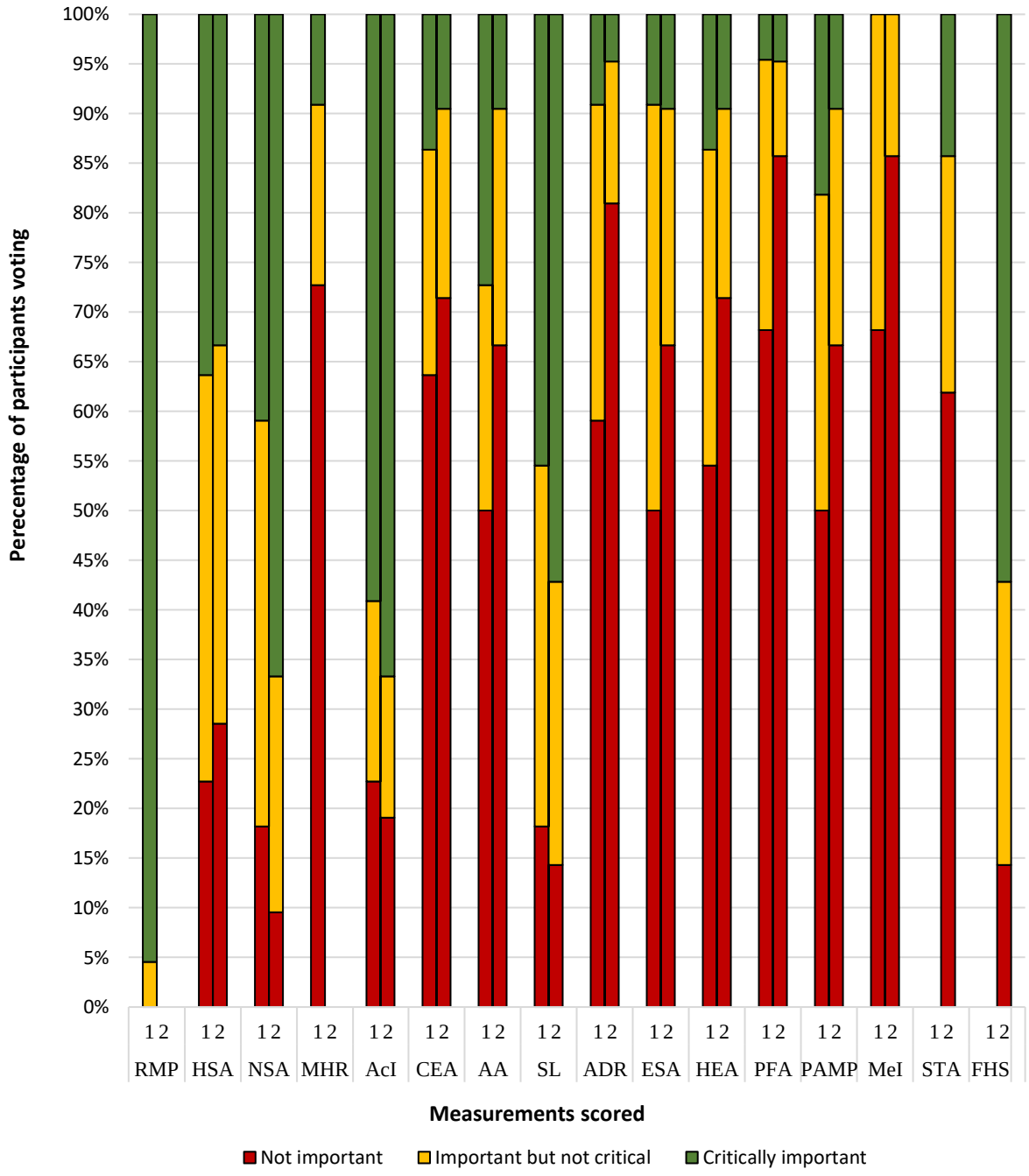
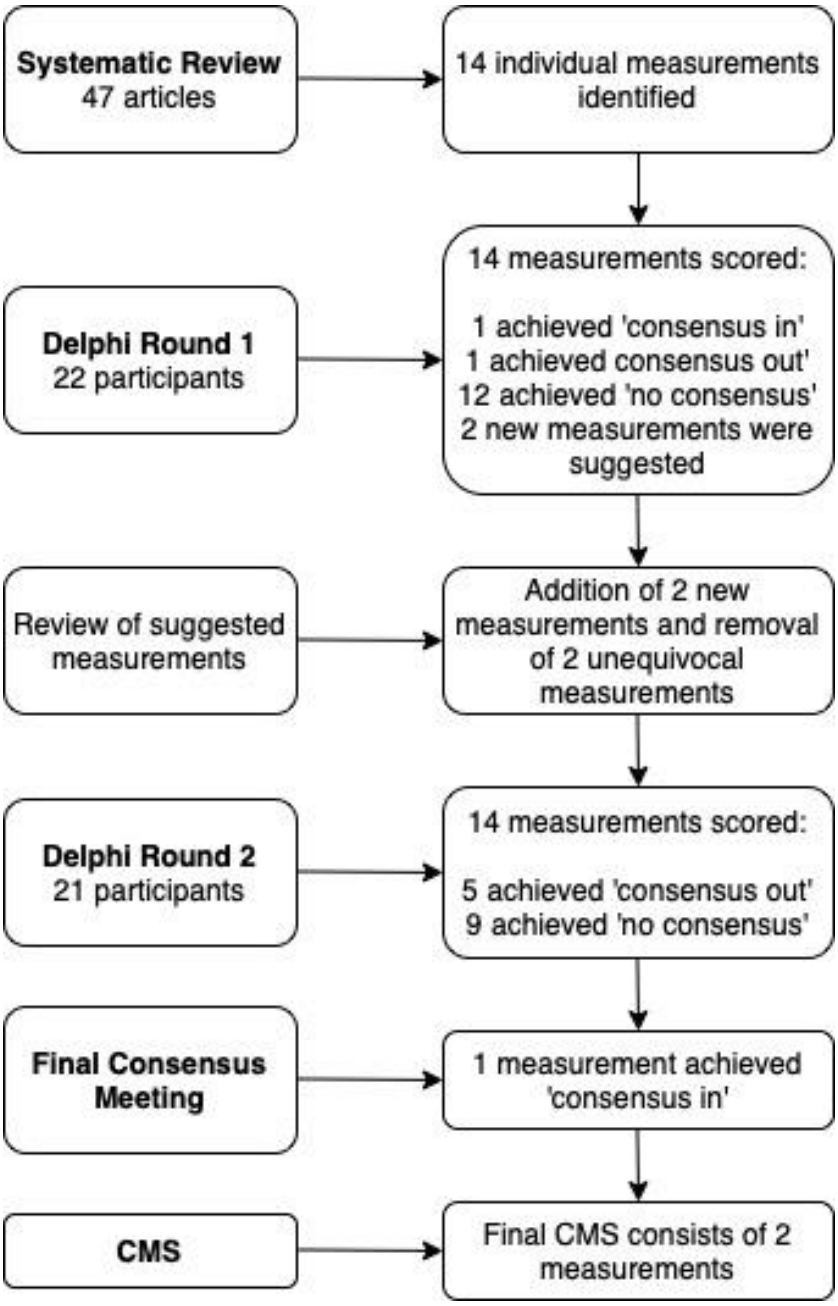


Figure 4.2 Overview of the development of the CMS



4.4 Discussion

This is the first study to identify a set of core measurements for reporting in studies on hip disease in CP. Most participants were orthopaedic surgeons, and the country with the most representation was the UK. Sixteen measurements were presented over the two rounds, and seven measurements reached consensus. Of these seven measurements, RMP was voted in, and MHR, CEA, ADR, HEA, PFA and

MeI were voted out. Following the final consensus meeting, HSA was added to the CMS, and all other measurements were voted out. The final CMS consists of RMP and HSA.

4.4.1 Interpretation of Results

Reimers' Migration Percentage is a commonly used measurement that has been considered the 'gold standard' in the radiographic assessment of hip displacement.(54) Furthermore, it reached 'consensus in' during the first round of the Delphi study, with 95% of participants considering it 'critically important' with a score of nine out of nine. Consequently, this measurement will likely be widely accepted in future projects. HSA was the fourth most commonly reported measurement identified in the systematic review. As outlined in the previous chapter, there is some controversy regarding association the association between HSA and hip displacement.(306, 313, 314, 321) The Delphi responses reflect this disagreement reaching 'no consensus' in both rounds. However, despite the debate surrounding HSA, the ability to successfully calculate a risk score for hip displacement using HSA suggests there must be some correlation.(51) The CPUP hip score has also been assessed in multiple populations, achieving a high discriminatory accuracy indicating its ability to differentiate between individuals with different levels of risk for hip displacement, further supporting its significance.(51, 322) In line with this opinion, HSA was deemed 'critically important' by the Consensus Focus Group to warrant inclusion in the CMS. Head-Shaft Angle is the only other measurement needed, in addition to RMP, to calculate the CPUP Hip Score, which is used to predict the probability of developing hip displacement in the next five years in children with an RMP of >40%.(51) Furthermore, the measuring of HSA in addition to RMP is already embedded in the protocols of a few national hip surveillance programmes such as CPUP and CPIPS.(11, 56) Given the ability to successfully calculate a risk score using RMP and HSA, the addition of HSA is a trend that other hip surveillance programmes are likely to adopt, making it a natural addition to the CMS.

Other commonly reported measurements identified in the systematic review include NSA and AcI. Neither of these measurements reached consensus after the two rounds of the Delphi; however, they were both very close to being voted in during the second round, with 67% of participants rating them as 'critically important' measurements. In the previous chapter, NSA and AcI were the second and third most commonly reported measurements, respectively. Despite these measurements being reported more often than HSA, they did not make the CMS.

Neck-Shaft Angle requires both hips to be internally rotated to get an accurate measurement. An insufficiently rotated hip joint will result in an inaccurate measurement.(323) Although some mathematical solutions have been developed to help correct this rotational effect, the correctional

outcome cannot be reliably verified.(324) In addition to the effect of rotation on NSA, there are other shortcomings that further support its exclusion from the CMS. Boese et al. found a high variance in the reporting of NSA in their systematic review and identified inconsistent methods of measurement as the main issue for this.(324) An inconsistently reported measurement may not successfully facilitate comparisons between studies or achieve true uniformity in the reporting of CP hip radiographs. Furthermore, Foroohar et al. stated a preference for HSA over NSA, suggesting that NSA may underestimate the deformity of the proximal femur given the valgus position of the femoral head in comparison to the femoral neck.(321)

The significance of AcI has been challenged in multiple studies and its exclusion from the CMS is expected. Hägglund et al. report that radiographic signs of acetabular dysplasia are usually only detected in the later stages of hip displacement, which suggests that AcI is not a good predictor of hip disease in CP.(325) Terjesen et al. corroborate these results but found AcI to be a significant prognostic factor in hip displacement.(31) Despite this, Terjesen et al. suggest only using AcI as a supplement to RMP and also note that AcI can be extremely hard to measure accurately in children with fixed flexion deformities that cause anterior pelvic tilts.(31) DiFazio et al. and Spencer et al. also acknowledge the difficulties of accurately measuring AcI in children with CP.(326, 327)

Four of the six measurements that were voted out after the Delphi survey, namely MHR, HEA, PFA and MI, were also the least commonly reported measurements in the systematic review, with only one study reporting each of these measurements. RMP was the only measurement that was included after the two rounds of the Delphi survey and was the most commonly reported measurement in the systematic review appearing in over 90% of studies. This suggests an association between how frequently a measurement is reported and its importance in the assessment of CP hips. Less commonly reported measurements are less important than more commonly reported measurements, and vice versa.

The vast majority of participants were orthopaedic surgeons, with only one physiotherapist taking part. This dominance is likely due to the relative predominance of orthopaedic surgeons carrying out the interpretation of CP hip radiographs. Physiotherapists are usually more involved in the physical examination of CP children; however, in recent years, their role has changed, with an increasing involvement in the radiological aspect of a child's examination.(11)

4.4.2 Implications for Research and Clinical Practice

This CMS will directly inform the development of an automated system for measuring hip disease in children with CP. The automatic calculation of critical measurements and a hip displacement risk score

will be especially useful in surveillance programmes. The automatic calculation of risk scores for hip displacement will enable children to be monitored without needing input from clinicians and make consistent and accurate calculations. Thus far, AI radiograph analysis software has not been developed using a CMS. This study provides a framework ensuring the inclusion of critical measurements and landmarks for clinically impactful AI software.

The formation of a CMS will standardise the evaluation of hip radiographs in CP children. This will improve the quality of reporting and allow for the calculation of risk scores for hip displacement. A standardised CMS also allows for comparisons to be drawn between studies. Additionally, the CMS can serve as a minimum set of measurements to be reported in hip surveillance programmes. This way, research can be conducted by combining and comparing data from different programmes.

4.4.3 Strengths and Limitations

This study had very high retention rates, with only one participant from round one not taking part in round two. This is likely due to the online format of the survey and the small amount of time required to complete the survey allowing individuals to participate in a time-efficient and convenient manner. Hall et al. have found personalised email initiations to be the most effective method for recruitment and retention, and report on the negative effects of a three-round Delphi study on retention.(328) This study only consisted of two rounds and used personalised emails to recruit and remind participants to take part, further explaining the high retention rates.

Moreover, this study could only be conducted with one stakeholder group; therefore, it is free from attrition bias. Although orthopaedic surgeons and physiotherapists have different roles in the care of CP patients, their perspectives on the importance of different radiographic measurements are not distinct, as both roles have the same objectives when assessing CP hip radiographs. Therefore, they come under the same stakeholder group preventing the possibility of an attrition bias. Contrary to the COMET guidelines, patient involvement was not appropriate in this study as this Delphi process assessed the diagnostic importance of geometric measurements as opposed to clinical trial-related outcomes assessing the effectiveness or safety of interventions, procedures or treatments that are directly relevant to patients.(315)

As highlighted in the previous chapter, measurements are not always reported to a high quality in the literature, lacking sufficient definitions and references. This has resulted in the reporting of multiple measurements that may or may not be synonymous with each other. Additionally, some of the listed measurements may not be well-known, with seven of the identified radiographic measurements only

appearing in one study. The use of multiple terms to describe the same measurement and the identification of less commonly known measurements may have resulted in some confusion when the measurements were scored during the Delphi survey. This is a limitation of this study. If an unfamiliar measurement was not researched thoroughly by a participant, the importance of that measurement may have been underestimated. However, if a measurement was unfamiliar to most participants, it is unlikely for that measurements to be of significant value, reducing the effect of this limitation.

Another limitation includes the generalisability of the views of the Consensus Focus Group to the rest of the Delphi participants. Since the Consensus Focus Group consisted of fewer individuals than the total number of participants, the opinions and conclusions reached in the final consensus meeting may not represent the other participants. Other limitations exist as a result of the online format of the Delphi survey and the lack of support for multiple languages. This study is biased towards English-speaking individuals with access to a computer with an internet connection.

4.5 Conclusion

In conclusion, this chapter has established a CMS for the reporting of radiographic measurements in the assessment of hip disease in children with CP. This CMS consists of two measurements: RMP and HSA. These measurements will directly inform the development of AI software for the automatic analysis of hip radiographs in children with CP and help standardise the evaluation of hip radiographs in CP across clinical practice and research. This chapter has achieved the initial objective to identify the most important measurements needed to assess hip disease in CP and form a CMS via a Delphi study.

Chapter 5: An automated system for the evaluation of pelvic radiographs in cerebral palsy: a preliminary study

5.1 Introduction

With the introduction of artificial intelligence (AI) in medical image analysis, computers are able to automate the diagnosis and classification of diseases, enabling clinicians to work efficiently and effectively.(329) However, the automatic analysis of hip radiographs in children with cerebral palsy (CP) has not been well studied. The manual calculation of radiographic measurements is a frequently performed task that would benefit from becoming automated given the growing number of hip surveillance programmes.(99-102, 316)

To date, only one study has looked into the automatic assessment of hip displacement in CP.(299) Pham et al. used a transfer learning approach on two deep learning models to measure Reimers' migration percentage on anteroposterior pelvis radiographs. More specifically, they used a fine-tuning approach in which two pre-trained convolutional neural networks, containing weights and biases specific to different datasets, were adjusted using a supervised learning to create a new model. One of the pre-trained models was tuned to detect eight reference landmarks, and the other pre-trained model was used to obtain more precise landmark coordinates.(299) As of yet, no research has been published investigating the use of a machine learning model that has been trained from scratch on datasets consisting of pelvic radiographs from children with CP. Although unsupervised learning models are the more popular choice in medical image analysis(178), this technique requires vast amounts of data. Therefore, supervised learning models are more feasible when building AI models from scratch and working with smaller datasets.

Random Forest Regression-Voting is a technique that has been used in supervised ensemble learning models to outline and segment skeletal structures on radiographs.(273) These outlines can then be used to perform shape analysis, including: (1) the construction of statistical shape models(330), which describe a range of anatomical variations from a mean bone shape for a given population, or (2) the automatic calculation of geometric measurements(331). The Random Forest Regression-Voting technique has been used to outline bone contours and landmarks in various structures, including the skull(332), teeth(333), hand(334), proximal femur(335), knee(336) and foot(337). It has also been used to calculate conventional geometric measurements such as: head diameter, neck width, shaft width and

neck-shaft angle from hip radiographs(331); hip–knee–ankle angle from knee radiographs(338); and calcaneal tilt, cuboid height and Meary's angle from foot radiographs(337). Furthermore, Random Forest Regression-Voting has been used to automatically classify the hip on pelvic radiographs as being affected by Legg-Calvé-Perthes disease or not.(339)

This chapter aims to explore the use Random Forest Regression-Voting technique to automate the calculation of Reimers' migration percentage (RMP) in CP hip radiographs. The proposed framework works by automatically identifying key structural landmarks of the hip and measuring the relevant anatomical distances needed to calculate a given radiographic measurement. This study is part of a larger project in which AI is used analyse multiple childhood diseases affecting the hip including Legg-Calvé-Perthes disease, slipped capital femoral epiphysis, developmental dysplasia of the hip and CP hip disease. The long-term objective is to create a system in which children's hip radiographs can be automatically diagnosed with one of these four conditions. In this chapter, a preliminary analysis conducted assessing the AI model's ability to calculate RMP from CP hip radiographs; the reliability of the dataset used to train the AI model will also be assessed.

5.2 Methods

5.2.1 Dataset

The anteroposterior pelvic radiographs of 1650 children were retrospectively selected from the Picture Archiving Computer Systems (PACS) records stored at Alder Hey Children's Hospital. The inclusion criteria were: (1) age between 2 and 16 years, and (2) a diagnosis of CP with a GMFCS level be between 1 and 5. The exclusion criterion was an unidentifiable femoral head or acetabulum. Radiographs that matched the inclusion criteria were completely anonymised and exported to the annotating software, where they were screened using the exclusion criteria. The harmonised arrangement for the Governance Arrangements for Research Ethics Committees (GAfREC) has judged that the use of anonymised information in this way is acceptable without requiring patient consent.(340) The Alder Hey Caldicott guardian approved the use of radiographs.

5.2.2 Training Dataset Preparation

Each radiograph was manually annotated by two observers using the BoneFinder® annotation software.(273) Trainer 1, Mr Prince Josiah Sajanthan Joseph (PJSJ), was an MPhil student at the University of Liverpool with no prior experience in radiograph analysis; Trainer 2, Dr Peter Thompson (PT), was a computer science research associate at the University of Manchester with a year of

experience in radiograph analysis. The pelvis and proximal femur in a radiograph were annotated separately. Trainer 1 was responsible for annotating the left proximal femur, and Trainer 2 was responsible for annotating the right proximal femur and the pelvis. The right proximal femur was flipped prior to annotation by Trainer 2, resembling a left proximal femur, as the AI model was not programmed to differentiate between the left and right femur. Therefore, the left and right femurs were treated generically as the femur after completing the annotations. Each radiographic structure was only annotated once. The pelvis and proximal femurs were outlined by placing a set of points along the contour of the bone or at key landmark positions. Once the AI trainers annotated all 1650 radiographs, the labelled images were used to train the AI model to outline and segment the pelvis and the proximal femur automatically. Each annotation point represents a specific anatomical structure or landmark (Tables 5.1 and 5.2). Sixty annotation points, numbered 0 to 59, were used to plot the outline of the pelvis (Figure 5.1) and 42 annotation points, numbered 0 to 41, were used to plot the outline of the proximal femur omitting the greater and lesser trochanter (Figure 5.2). Key annotation points were chosen on either side of the triradiate cartilage and at each end of the acetabular roof to segment Hilgenreiner's line and the acetabular roof, both of which are needed to calculate RMP.

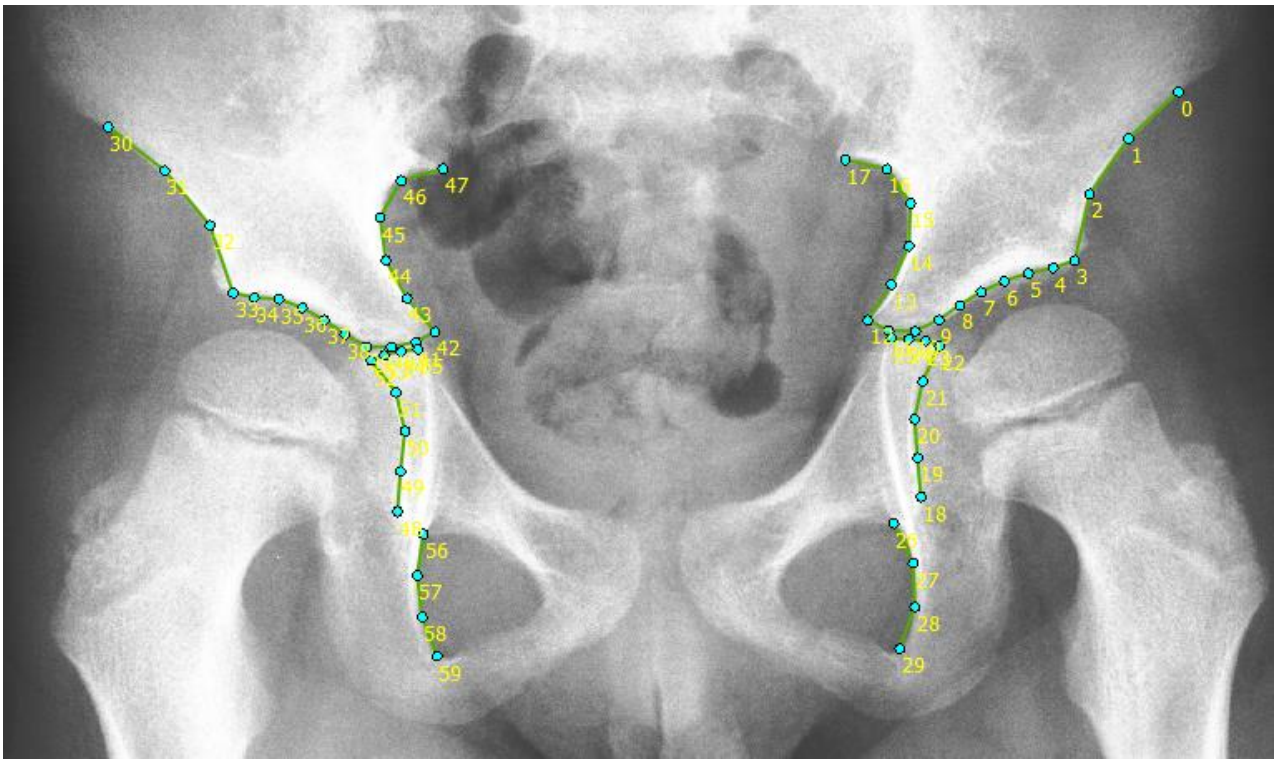
Table 5.1 Description of annotation points plotted on the pelvis.

Pelvis	
Points	**Description**
0-3	Extends out the edge of the pelvis and continue to the first point of discontinuity.
30-33	
3-9	Follows the acetabular roof from its lateral to medial end.
33-39	
9-12 & 22-25	Follows the visible edge of the bone on either side of the triradiate cartilage, with 9 being placed at the medial end of the acetabular roof. Note that the points on the lower side may appear above those on the upper side due to parallax.
39-42 & 52-55	
12-17	Follow the pelvic brim from the triradiate cartilage to the corner point where the ilium meets the sacrum.
42-47	
18-22	Follow the lower part of the acetabulum until it vanishes or merges with the horizontal line of the of the pubic bone.
48-52	
26-29	Follow the inside edge of the ischium from where it crosses the ramus superior to the corner before it fuses with the pubis.
56-59	

Table 5.2 Description of annotation points plotted on the proximal femur.

Proximal Femur	
Points	**Description**
0-6	The proximal side of the femoral shaft starting at the intersection with the lesser trochanter. If the lesser trochanter is not visible, the most consistent estimate of its location was taken.
7	The inflection where the convex part of the femoral neck starts.
8-12	Follows the convex part of the femoral neck to the physis.
12&31	Spans the gap between the femoral head and neck in cases where they are not yet fused. In cases where they are fused, the points were placed on the visible seam. If there is no visible seam, the most consistent estimate of its location was taken.
13&41	
31-41	Follows the edge of the femoral head.
13-23	Follows the lateral side of the femoral neck to the corner formed by the femoral shaft and neck.
24-30	Follows the lateral side of the femoral shaft up to point 30, which is directly opposite to point 0 on the other side of the femoral shaft.

Figure 5.1 Manually annotated pelvis. Annotation points (blue) are numbered 0-59. Curves (green) are formed by a line drawn between adjacent annotation points.



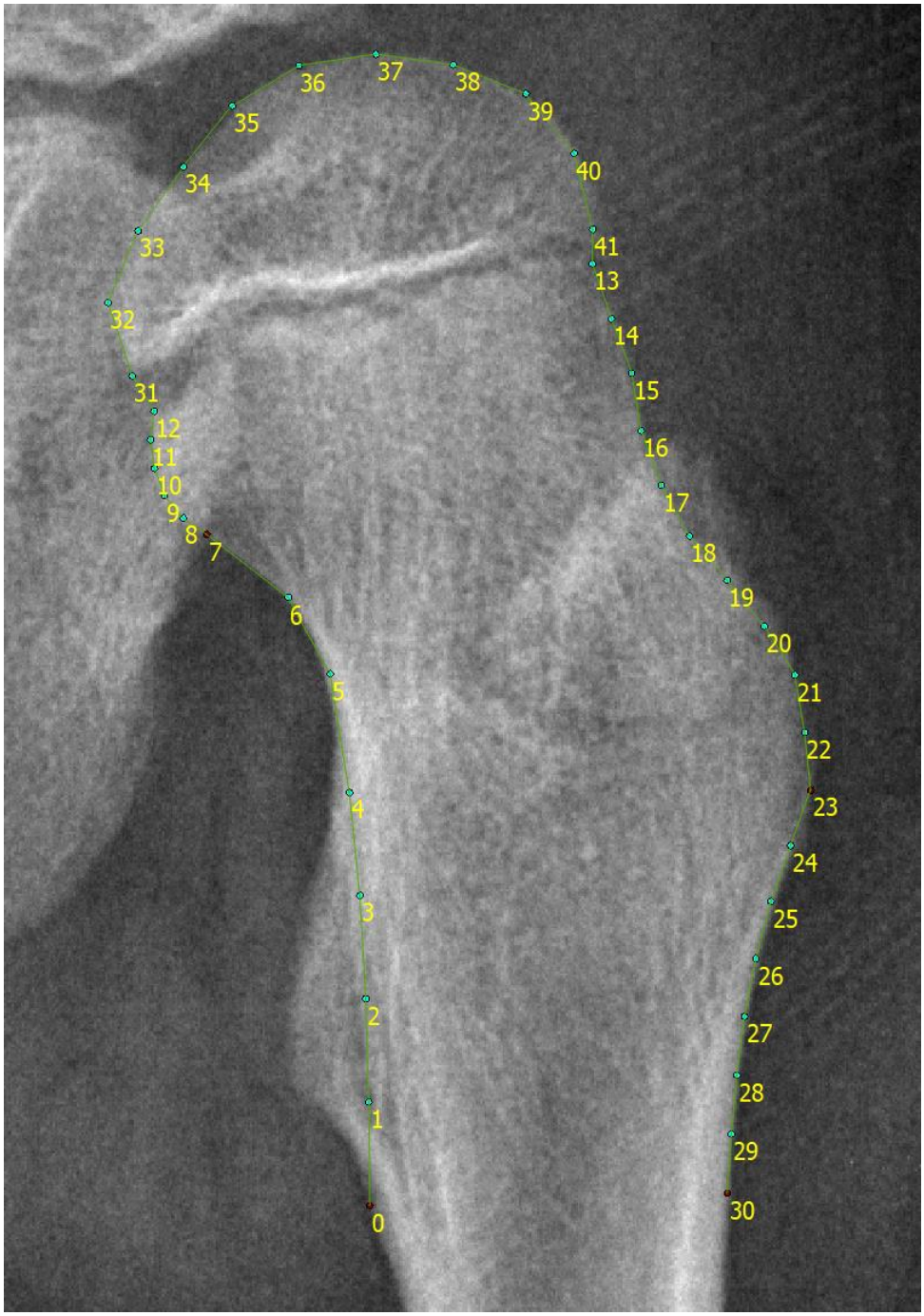


Figure 5.2 Manually annotated left proximal femur. Annotation points (blue) are numbered 0-41. Curves (green) are formed by a line drawn between adjacent annotation points.

5.2.3 Artificial Intelligence Model

The AI model was developed in collaboration with engineers from the University of Manchester, using previous models that were trained to analyse, outline and segment radiographs of healthy patients, patients with Legg-Calvé-Perthes disease and patients with developmental dysplasia of the hip. The

model automatically detects the structural landmarks of the hip using a Hough-like approach and then locally refines the positions of each point by applying Random Forest Regression-Voting in a Constrained Local Model framework.(273) Each radiograph is interpreted as a grid, with each annotation point representing a coordinate. The AI model can calculate a given radiographic measurement by measuring the distances between key landmarks. Currently, the AI model has been programmed to calculate RMP by identifying Hilgenreiner's line, defined as the horizontal line that passes through the triradiate cartilage on each side of the pelvis, and the acetabular roof, defined as a line of impaction running from the triradiate cartilage to the end of the socket. The final model will also calculate the femoral head-shaft angle (HSA)(321), the other measurement in the core measurement set.

5.2.4 Evaluation

5.2.4.1 AI Model

The performance of the AI model was assessed by: (1) comparing the inter-observer reliability of automatically derived RMP values generated by the model and manually derived RMP values generated by five different clinicians; and (2) calculating the point-to-point distance between the automatically annotated points plotted by the AI model and the manually annotated points plotted by the AI trainers during the preparation of the training dataset.

Intraclass correlation coefficients (ICC) and their 95% confidence intervals (CI) were used to assess the inter-observer reliability between (1) the five clinicians that provided manually calculated RMP values and (2) the RMP values derived automatically and manually. The manually derived RMP values were generated using the mean RMP values of the five clinicians for each image, and the automatically derived RMP values were generated using the AI model. The five clinicians that provided the manual RMP measurements were all orthopaedic specialty trainee doctors. Each clinician used custom-made software, which enabled them to calculate the RMP by drawing Hilgenreiner's line, Perkin's line, and two parallel lines marking the medial and lateral edges of the femoral head. Each clinician received training and had practice measurements validated by the research group before recording measurements for the study.

Each of the ICC estimates and their 95% CI were calculated using SPSS statistical package version 27 (SPSS Inc, Chicago, IL) based on two-way mixed effects, absolute agreement, single rater/measurement model per the McGraw and Wong guidelines(341). The inter-observer reliability was interpreted using guidelines published by Koo et al., which state that: "the 95% confident interval of the ICC estimate (not the ICC estimate itself) should be used as the basis to evaluate the level of reliability using the

following general guideline: *Values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability*." (342)

In this study, the point-to-point distance (mm) was measured as the Euclidian distance between two matching annotation points in the same radiograph, where each annotation point is plotted by a different annotator. Matching annotation points correspond to the same structural landmark descriptions in Tables 5.1 and 5.2. Based on previous studies,(273, 339, 343) a general guideline was formed to assess the reliability of different point-to-point distance values: *Point-to-point distances under 1mm indicate excellent reliability, point-to-point distances between 1mm and 2mm indicate very good reliability, point-to-point distances between 2mm and 3mm indicate good reliability, point-to-point distances between 3mm and 4mm indicate adequate reliability, and point-to-point distances over 4mm indicate poor reliability.*

In order to assess the performance of the AI model, the point-to-point distances for annotation points that were automatically plotted by the AI model were compared to annotation points that were manually plotted by the AI trainers during the dataset preparation stage of the study. A total of 102 point-to-point distances were calculated for each radiograph for the 60 annotation points of the pelvis (numbered 0-59, as mentioned in Table 5.1) and the 42 annotations points of the proximal femur (numbered 0-41, as mentioned in Table 5.2). Descriptive statistics (including mean, standard error, standard deviation, median, 90^{th} percentile, 95^{th} percentile and 99^{th} percentile) were calculated using custom-made software to summarise: (1) the mean point-to-point distance for each annotation point across all radiographs; and (2) the overall mean point-to-point distance across all annotation points and all radiographs.

5.2.4.2 Training Dataset

The inter-and intra-observer reliability of the training dataset was assessed by comparing the point-to-point distance between annotations plotted by Trainer 1 and Trainer 2. In order to test the inter-and intra-observer reliability, radiographs were reannotated by the AI trainers as the radiographic structures in each radiograph were annotated individually during the preparation stage. For this preliminary study, the reliability of the left proximal femur was assessed, and 42 point-to-point distances were calculated in each radiograph (numbered 0-41, as described in Table 5.2). Since Trainer 1 was solely responsible for the annotation of the left proximal femur in the training dataset, the left proximal femur was annotated for the first time by Trainer 2 in order to assess the inter-observer reliability, and for a second time by Trainer 1 in order to assess the intra-observer reliability. The intra-observer analysis was conducted with a three-month gap between the first and second annotations of Trainer 1. Descriptive

statistics were calculated, as mentioned before, for each observation pair: Trainer 1's initial annotations and Trainer 2's annotations (T1a-T2); Trainer 1's repeat annotations and Trainer 2's annotations (T1b-T2); and Trainer 1's initial and repeat annotations (T1a-T1b). Cumulative distributions of point-to-point distance thresholds were also determined for each observation pair using a stepwise increase in distance from 0.03mm to 39.97mm.

5.3 Results

5.3.1 AI Model Performance

5.3.1.1 Reimers' Migration Percentage

A total of 450 new images, separate to the images used in the training cohort, were randomly selected to assess the AI model's ability to automatically calculate RMP and outline and segment the pelvis and proximal femur. Table 5.3 presents the descriptive statistics and inter-observer reliability of five clinicians. The highest mean RMP was recorded by Clinician 1 (mean = 31.99, SD = 25.25) and the lowest mean RMP was recorded by Clinician 2 (mean = 28.48, SD = 25.96). "Excellent" inter-observer reliability was observed between all five clinicians (ICC = 0.946, 95% CI = 0.936-0.954).

Table 5.3 Descriptive statistics and inter-observer reliability of RMP measurements (%) derived from five clinicians.

Observers	Mean	SD[1]	Range	ICC[2]	95% confidence interval	
					Lower bound	Upper bound
Clinician 1	31.99	25.25	0-100	0.946	0.936	0.954
Clinician 2	28.48	25.96	0-100			
Clinician 3	30.16	25.48	0-100			
Clinician 4	28.66	25.04	0-100			
Clinician 5	31.09	24.15	0-100			

[1]Standard deviation, [2] Intraclass correlation coefficient

Table 5.4 presents the descriptive statistics and inter-observer reliability of the AI and manually derived measurements. The AI measurements a mean RMP (mean = 32.27, SD = 25.57) that was 2.19% higher than the manual measurement (mean = 30.08, SD = 24.67). The inter-observer reliability between the AI measurements and manual measurements was 'good to excellent' (ICC = 0.893, 95% CI = 0.870-

0.912). The full set of RMP values recorded by each observer for 450 images is presented in Appendix 10.

Table 5.4 Descriptive statistics and inter-observer reliability of RMP measurements (%) derived automatically by the AI model and manually by the five clinicians. The 'manual' RMP measurements were formed using the mean RMP of the five clinicians for each image.

Observers	Mean	SD[1]	Range	ICC[2]	95% confidence interval	
					Lower bound	Upper bound
Manual	30.08	24.67	0-100	0.893	0.870	0.912
AI	32.27	25.57	0-100			

[1]Standard deviation, [2]Intraclass correlation coefficient

5.3.1.2 Point-to-Point Distance

The descriptive statistics of the overall mean point-to-point distances for each anatomical structure are presented in Table 5.5. The mean overall point-to-point distance was adequate (mean = 3.64mm, SD = 4.22mm), indicating a moderate inter-observer reliability between the AI annotations and the manual annotations. The pelvis had a lower mean point-to-point distance (mean = 2.51mm, SD = 2.15mm) in comparison to the proximal femur (mean = 5.26mm, SD = 8.51mm). The AI model achieved an overall mean point-to-point distance within 6.15mm, 8.23mm and 26.5mm for 90%, 95% and 99% of all images, respectively. The pelvis recorded lower mean point-to-point distances for 90%, 95% and 100% of radiographs (90% = 4.84mm, 95% = 6.00mm, 99% = 8.94mm) than the proximal femur (90% = 8.69mm, 95% = 12.2mm, 99% = 57.2mm).

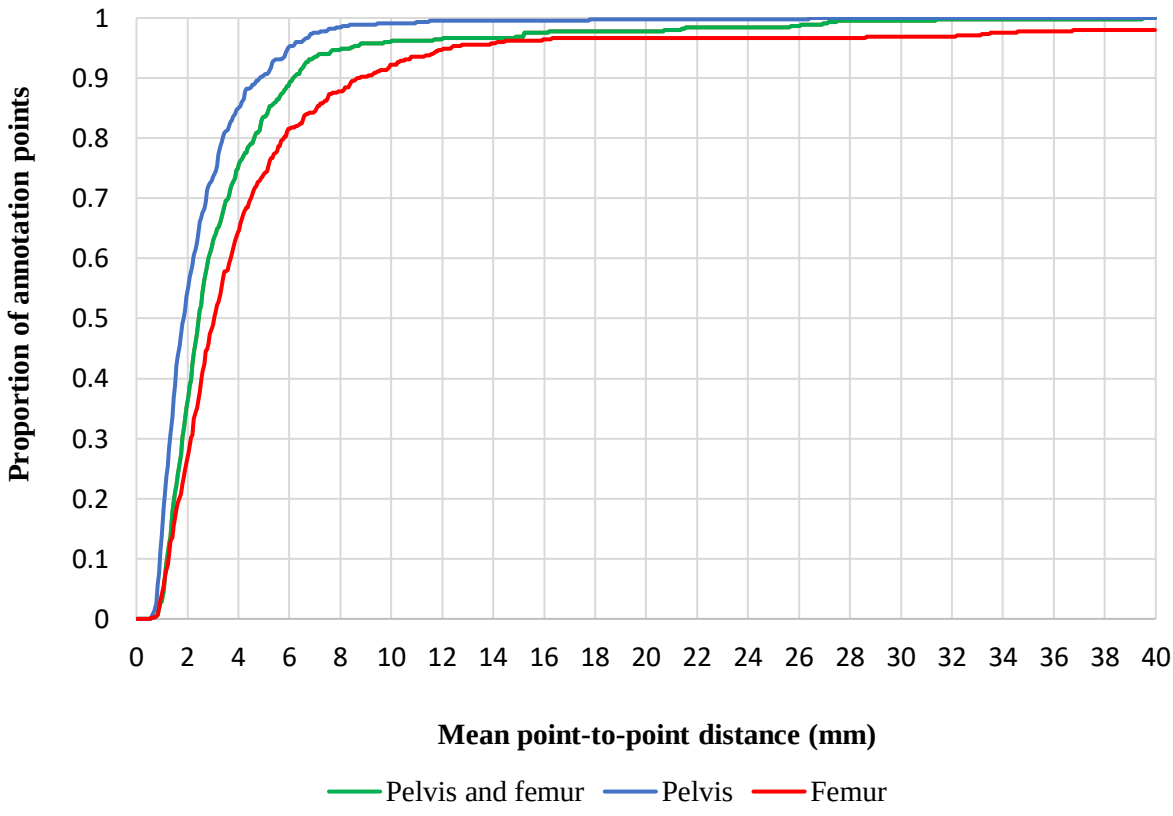
Table 5.5 Descriptive statistics of the overall point-to-point distances for different anatomical structures.

Structure	Mean	Standard Error	Standard Deviation	Median	90%	95%	99%
Pelvis and femur	3.64	0.199	4.22	2.49	6.15	8.23	26.5
Pelvis	2.51	0.102	2.15	1.89	4.85	6.00	8.94
Femur	5.26	0.402	8.51	3.07	8.69	12.2	57.2

In the pelvis, the lowest mean point-to-point distance was recorded point 56 (mean = 1.63mm, SD = 3.30mm), and the highest point-to-point distance was recorded for point 0 (mean = 4.01mm, SD = 5.61mm). In the proximal femur, the lowest mean point-to-point distance was recorded point 37 (mean = 3.63mm, SD = 8.64mm), and the highest point-to-point distance was recorded for point 0 (mean = 11.2mm, SD = 13.0mm). The descriptive statistics for the point-to-point distances per annotation point between the annotations completed by the AI model and the observers who helped to train the model for the pelvis and proximal femur across 450 images are presented in Appendices 11 and 12.

Figure 5.3 presents a cumulative distribution function graph showing the proportion of annotation points falling within a given mean point-to-point distance for each anatomical structure. In the pelvis, points 57, 35 and 41 recorded the lowest mean point-to-point distance for 90%, 95% and 99% of radiographs, respectively (90% = 3.18mm, 95% = 5.41mm, 99% = 11.2mm), and point 0 achieved the highest mean point-to-point distance for 90%, 95% and 99% of radiographs (90% = 9.47mm, 95% = 15.4mm, 99% = 27.3mm). In the proximal femur, points 38, 8 and 31 recorded the lowest mean point-to-point distance for 90%, 95% and 99% of radiographs, respectively (90% = 6.27mm, 95% = 9.55mm, 99% = 43.3mm), and point 0, 0 and 23 achieved the highest mean point-to-point distance for 90%, 95% and 99% of radiographs, respectively (90% = 28.4mm, 95% = 35.6mm, 99% = 72.3mm).

Figure 5.3 Cumulative distribution function graph showing the proportion of annotation points falling under a given point-to-point distance for each anatomical structure.



5.3.2 Reliability of Training Dataset

A total of 50 radiographs were randomly selected to assess the inter-and intra-observer reliability of the left proximal femur annotations from the training dataset. The descriptive statistics of the overall mean point-to-point distance for each observation pair are presented in Table 5.6. The mean overall point-to-point distance was generally low for every observation pair, with T1b-T2 and T1a-T1b achieving very good reliability and T1a-T2 achieving good reliability. The lowest overall point-to-point distance was seen between T1a-T1b (mean = 1.45mm, SD = 0.929mm), and the highest overall point-to-point distance was seen between T1a-T2 (mean = 2.20mm, SD = 1.98mm). Observation pair T1a-T1b achieved the lowest mean point-to-point distance for 90%, 95% and 99% of radiographs (90% = 3.25mm, 95% = 3.47mm, 99% = 3.83mm), and observation pair T1a-T2 achieved the highest mean point-to-point distance for 90%, 95% and 99% of radiographs (90% = 4.81mm, 95% = 6.38mm, 99% = 8.36mm).

Table 5.6 Descriptive statistics of the overall point-to-point distance for different observation pairs.

	Observation pairs	Mean	Standard Error	Standard Deviation	Median	90%	95%	99%
Inter-observer	**T1a[1]-T2[2]**	2.20	0.283	1.98	1.37	4.81	6.38	8.36
	T1b[3]-T2	1.93	0.242	1.7	1.37	4.31	6.17	7.38
Intra-observer	**T1a-T1b**	1.45	0.133	0.929	1.28	3.25	3.47	3.83

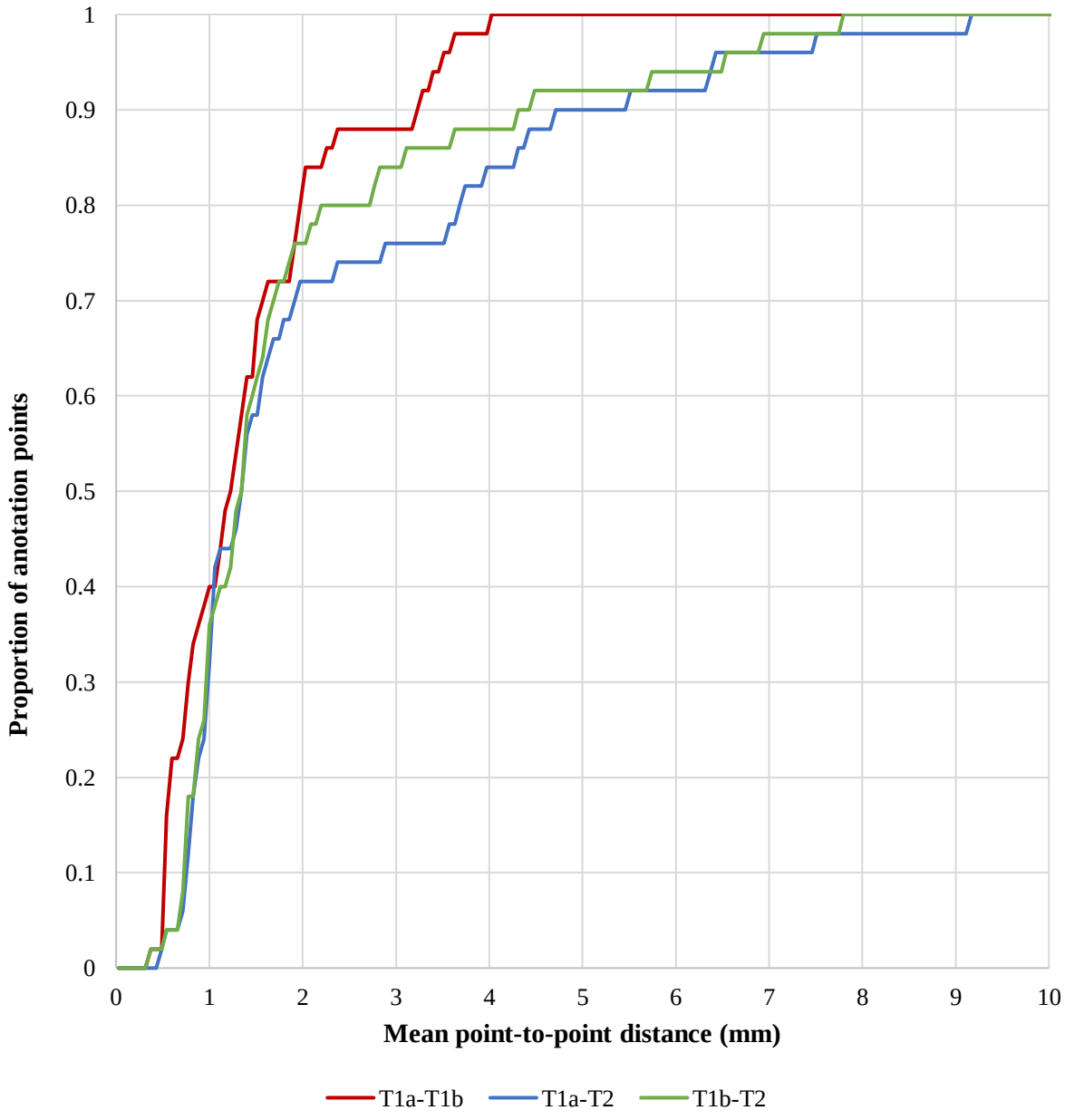
[1]Trainer 1's first proximal femur annotations, [2] Trainer 2's first proximal femur annotations, [3]Trainer 1's second proximal femur annotations.

The full set of descriptive statistics for the point-to-point distances per annotation point between each observation pair are presented in Appendices 13, 14 and 15. In observation pair T1a-T2, the lowest mean point-to-point distance was recorded for Point 15 (mean = 1.13mm, SD = 1.03mm) and the highest mean point-to-point distance was recorded for Point 0 (mean = 4.45mm, SD = 5.20mm). In observation pair T1b-T2, the lowest mean point-to-point distance was recorded for Point 38 (mean = 1.02mm, SD = 1.19mm) and the highest mean point-to-point distance was recorded for Point 30 (mean = 3.78mm, SD = 4.24mm). In observation pair T1a-T1b, the lowest mean point-to-point distance was recorded for

Point 38 (mean = 0.565mm, SD = 1.07mm) and the highest mean point-to-point distance was recorded for Point 0 (mean = 2.82mm, SD = 3.58mm).

Figure 5.4 presents a cumulative distribution function graph for the proportion of annotation points that fall within a given point-to-point distance. The intra-observation pair had lower mean point-to-point distances overall compared to the inter-observation pairs, and the mean point-to-point distances were generally lower in T1b-T2 than T1a-T2 amongst the inter-observation pairs.

Figure 5.4 Cumulative distribution function graph showing the proportion of annotation points falling under a given point-to-point distance for each observation pair.



For observation pair T1a-T2, points 39, 14 and 14 achieved the lowest mean point-to-point distance for 90%, 95% and 99% of radiographs, respectively (90% = 1.93mm, 95% = 2.93mm, 99% = 4.40mm), and points 0, 0 and 32 achieved the highest mean point-to-point distance for 90%, 95% and 99% of radiographs, respectively (90% = 11.9mm, 95% = 16.3mm, 99% = 23.8mm).

For observation pair T1b-T2, points 39, 38 and 39 achieved the lowest mean point-to-point distance for 90%, 95% and 99% of radiographs, respectively (90% = 1.68mm, 95% = 2.43mm, 99% = 4.55mm), and points 30, 30 and 32 achieved the highest mean point-to-point distance for 90%, 95% and 99% of radiographs, respectively (90% = 11.0mm, 95% = 12.8mm, 99% = 23.1mm).

For observation pair T1a-T1b, points 37, 38 and 16 achieved the lowest mean point-to-point distance for 90%, 95% and 99% of radiographs, respectively (90% = 0.88mm, 95% = 1.56mm, 99% = 2.97mm), and points 0, 30 and 0 achieved the highest mean point-to-point distance for 90%, 95% and 99% of radiographs, respectively (90% = 8.08mm, 95% = 9.72mm, 99% = 14.7mm).

5.4 Discussion

This study has demonstrated that this AI model can achieve 'good to excellent' inter-observer reliability in the measurement of RMP when compared with clinicians, despite only achieving an adequate point-to-point distances when compared to the AI trainers' annotations. The AI trainers achieved a good inter- and intra-observer reliability, with the intra-observer annotations achieving lower point-to-point distances than the inter-observer annotations. Moreover, between the inter-observer analyses, the observation pair containing Trainer 1's repeat annotations achieved a lower point-to-point distance than Trainer 1's initial annotations.

5.4.1 Interpretation of Results

Our AI model's ability to outline and segment the pelvis and femur was adequate. When comparing the annotations between the AI model and clinicians, the annotation point with the highest point-to-point distance in both the pelvis and the femur was point 0. In the pelvis, point 0 corresponds to the first point of discontinuity along the edge of the pelvis (Table 5.1). In the proximal femur, point 0 corresponds to the proximal side of the femoral shaft where the lesser trochanter intersects the femur (Table 5.2). Point 0 is located at the end of a curve in both the pelvis and the femur, suggesting that the AI model is less reliable near these points (Figures 5.1 and 5.2).

Furthermore, of the annotation points ranking in the top five highest mean point-to-point distances, the vast majority, namely points 0, 30, 1 and 17 in the pelvis and points 0, 30, 1, 29 and 2 in the femur, are placed at or near the end of a curve. This further corroborates the theory that the AI model struggles to accurately plot annotation points at or near the end of a curve. The reliability of the training dataset for the left proximal femur was also substandard for points at or near the end of a curve. The top five annotations points with the highest mean point-to-point distances in the inter-and intra-observation pairs were points 0, 30, 1, 29 and 2, with point 0 recording the highest mean point-to-point distance. This directly corresponds with the top five points that the AI model performed poorly on, which suggests that the inferior performance of the AI model for points at or near the end of a curve could have resulted from the poor consistency in the training dataset amongst the AI trainers. This could be either as opposed to or in addition to computational insufficiencies in the AI model.

Another possible explanation for the high point-to-point distances in points placed at or near the end of a curve can include ambiguity in the definition of those points. The pelvis and femur are large bone structures that do not need to be outlined and segmented entirely to calculate measurements in CP hip radiographs. Cut-off points, which are often annotations points at or near the end of curves, were identified for both the pelvis and femur; however, ambiguity in the definition of these cut-off points may have decreased the inter-and intra-observer reliability of annotation points. For example, point 0 in the pelvis, which represents the first point of discontinuity along the edge of the pelvis, has a wide area within which a point can be plotted whilst corresponding to the predefined point definitions. Although these cut-off points directly affect the mean point-to-point distance of a model's annotations, they are not directly used in the calculation of radiographic measurements. Therefore, inconsistency in the annotation of these points may not significantly affect the AI model's primary output, RMP. This is supported by the 'good to excellent' ICC estimates achieved by our model despite the poor mean point-to-point distances recorded for points at or near the end of curves.

Variability in a landmarks' visibility or structure may have also accounted for the high mean point-to-point distances observed in points at or near the end of a curve. The inter-and intra-observer analysis of the proximal femur amongst the AI trainers, and the AI model and clinicians, showed that point 0 was the least reliably annotated point. This could have resulted from an inaccurate identification of the anatomical landmark for the femur's point 0, which corresponds with the intersection of the lesser trochanter and the proximal side of the femoral shaft. This landmark was highly variable between radiographs as the lesser trochanter was not always visible in every image due to either the rotation of the femur, the presence of surgical screw and plate devices or a femoral deformity. In cases where this landmark was not visible, an estimate was taken, increasing the likelihood of a poor inter-and intra-observer reliability.

This variability in the radiographic presentation of the proximal femur may also explain the AI model's superior ability to identify the pelvis when compared to the femur. As mentioned previously, estimates were taken when the femur's point 0 was not visible, increasing the likelihood of inconsistent annotations in the training dataset. Using inconsistent annotations to train the AI model will decrease its accuracy. This mismatch in performance could have resulted from the additional trainer used to annotate the proximal femur in the training dataset. For the pelvis, only annotations from Trainer 2 were used; however, for the proximal femur, Trainer 1's annotations of the left femur and Trainer 2's annotations of the right femur were used. This increases the variability of the annotations used to teach the AI model to outline the proximal femur when compared to the pelvis, as inter-observer annotations are likely to have poorer reliability than intra-observer annotations. This is finding is further supported by the higher mean point-to-point distances obtained from the proximal femur annotations when compared to the pelvis annotations. This combination of an increased irregularity in the presentation of the femur and the use of multiple trainers to annotate the femur in the training dataset significantly increases the variability of the femur annotations used to train the AI model, thereby decreasing the model's accuracy for this structure.

Examples of CP-related proximal femoral deformities include femoral anteversion and coxa valga.(344) Unlike pelvic deformities, which occur as a result of hip displacement(64), femoral deformities have been shown to contribute to hip displacement(47) and occur independently of hip displacement(33). Therefore, femoral deformities may be observed more often than pelvic deformities, further explaining the mismatch in the AI model's ability to identify those structures. Beyond this, the AI model's inferior annotation of the proximal femur can be further explained by a reduced likelihood for the pelvis to be incorrectly positioned in a radiograph or feature orthopaedic tools from a previous surgery when compared to the femur.

The inter-observer reliability between Trainer 1 and Trainer 2 was higher when comparing Trainer 2's annotations with Trainer 1's second annotations (T1b-T2) as opposed to Trainer 1's first annotations (T1a-T2). This difference in reliability could be explained by the increased experience Trainer 1 would have gained by the time the repeat annotations were required. This suggests that the training dataset's reliability increases as the trainers' experience increases.

5.4.2 Comparison to other AI models

Pham et al.(299), who developed the only other AI model for the automated assessment of CP hip disease, recorded an ICC and 95% CI of 0.91 [0.87,0.94] and 0.91 [0.86,0.94] when comparing the RMP values between their AI model and their manual observers. This translates to a 'good to excellent'

reliability, which is on par with our model. Although the ICC estimates appear to give Pham et al's model the edge over ours, several limitations render our model preferable in comparison. The data used to test and train Pham et al's model included radiographs in which the majority had triradiate cartilages that had not yet fused. This restricts the generalisability of the model to calculate RMP measurements in children with CP past the age of 10.

Additionally, the eight landmarks identified as reference points for RMP measurements are tailored toward radiographs in which the triradiate cartilage has not yet fused, further reducing the effectiveness of Pham et al's model in clinical scenarios. Although our model also included annotations points that corresponded to landmarks along the triradiate cartilage, the inclusion of other distinctive points on the pelvis enabled our model to analyse radiographs in which the triradiate cartilage was ossified, as the model could capture the lateral symmetry of the pelvis using these additional points and produce a better estimate of its orientation in the image. Furthermore, the landmarks identified via the 102 annotations points outlining the pelvis and proximal femur can be used to calculate multiple measurements in addition to RMP, including acetabular index(2), neck-shaft angle(46) and HSA(321). Pham et al's model cannot be expanded to include additional radiographic measurements as the reference landmarks used in their model are specific to RMP.

Lindner et al.(330, 343) and Davison et al.(339) assessed the ability of their AI models to successfully annotate a proximal femur using the mean point-to-curve distance as opposed to the mean point-to-point distance. The point-to-curve distance is a more accurate measure when assessing linear structures, such as the femoral shaft, as the exact location of an annotation point along a linear structure is not necessarily important as long as the outline of the shape is correct. This is because minor differences in the point-to-point distance along a line will not significantly affect the shape of a straight annotation curve outlining a linear structure; however, it will affect the specific point-to-point distance. Conversely, the point-to-point distance is a more accurate measure when assessing non-linear structures, such as the femoral head and acetabulum, as the exact location of each point is extremely important, and minor differences in the point-to-point distance can significantly affect the arc of an annotation curve in these structures. This is a limitation of these previously mentioned studies as the point-to-point distance was not reported. Therefore, the AI model's ability to outline non-linear structures, such as the femoral head, was not appropriately assessed in each of these studies.(330, 339, 343) This is also a limitation of this study, as the point-to-curve distance was not assessed in addition to the point-to-point distance despite the presence of a combination of linear and non-linear structures in each radiograph. Furthermore, this limitation may further explain the poorer reliability seen in annotation points at or near the end of curves, as the end of curves are usually linear structures.

5.4.3 Limitations

Although, the AI model's ability to calculate RMP can be interpreted as 'good to excellent' using Koo et al's interpretation(342), there is not enough evidence to determine whether the AI model is on par with humans. Therefore, the accuracy of the model may still need improvement for use in clinical practice given the 2.19% difference in the mean RMP values calculated by the AI model and the clinicians. Additionally, official guidelines for the interpretation of point-to-point distances have yet not been published; therefore, the conclusions derived from this study regarding the AI model's ability to outline and segment structures may not be accurate, even though the definitions used were based on previous studies. Furthermore, there is currently not enough evidence to determine whether the AI model can make a clinically significant difference; however, even if there isn't a clinically significant difference, there are advantages to implementing an AI system to automatically analyse CP hip radiographs, and it could be financially beneficial to the NHS and help to standardise access to CP care.

Other limitations of this study include the generalisability of the AI model's performance to the level of experienced orthopaedic surgeons or physiotherapists. Although all the clinicians participating in this study were trained and had practice annotations validated, experienced surgeons with more years of experience may have generated different RMP values that may be more accurate than trainee doctors. However, Demir et al. studied the effect of experience and expertise on the inter-and intra-observer reliability of RMP in children with cerebral palsy and concluded that neither experience nor expertise affected the measurement of RMP.(308) This suggests that the measurements calculated by orthopaedic trainees are sufficient to generalise results to experienced orthopaedic surgeons. However, this excellent inter-and intra-observer reliability may not be extended to other measurements, namely the HSA. Therefore, in future studies, it would be beneficial to compare the performance of an AI model to that of experienced orthopaedic surgeons or physiotherapists.

Additionally, due to a lack of repeat annotation data, the inter-and intra-observer reliability was only assessed for the left proximal femur in this preliminary study. This may not represent the reliability of the right proximal femur annotations and, therefore, the proximal femur annotations fed to the machine. Although the results are likely similar to those of the left proximal femur, this cannot be assumed. This highlights one of the disadvantages of using a supervised learning method to train an AI model. The time-consuming nature of this method restricts the extent to which a preliminary analysis can be conducted on the inter-and intra-observer reliability of the manually annotated radiographs.

Furthermore, the patient age range in this preliminary analysis is different to the patient age range used to search for studies in the systematic review of chapter three. Images used in this chapter were obtained

from patients aged 2-16 years as per the CPIPS handbook(43), since the research group aims to integrate this AI model into the CPIPS system upon completion. One limitation of this focussed age range arises from the inability to use this AI model in other countries or regions where "children" with CP may include individuals beyond this age range. However, the core measurement set was built using an age range of ≤ 18 years in the systematic review, which is more representative of the age of inclusion across the literature.(101)

5.5 Conclusion

In conclusion, this chapter explored the use of Random Forest Regression-Voting to build an AI model capable of automatically calculating the RMP from hip radiographs in children with CP. Our AI model achieved a 'good to excellent' inter-observer reliability when compared to five trained clinicians and was able to outline the pelvis better than the femur. This AI model can also be programmed to calculate other radiographic measurements and assess radiographs in which the triradiate cartilage and physis of the femur have fused, providing a good alternative to the other AI models. The structural landmarks that our AI model and trainers struggled with the most were located at or near the end of an annotation curve. Future research should report both point-to-point distances and point-to-curve distances when analysing anatomical structures with linear and non-linear edges, assess AI models against experienced healthcare professionals and focus on the accuracy of the proximal femur given its increased variability in radiographs. This chapter achieved the initial objective of assessing the performance of the AI model as well as the reliability of the data used to train the model. However, this analysis did not include HSA, which is part of the CMS. Therefore, this chapter did not achieve the initial objective of assessing the performance of the AI model at calculating all the measurements in the CMS; although, it did assess the performance of the AI model at calculating RMP.

Chapter 6: Conclusion

This thesis aimed to identify the most important radiographic measurements used to assess hip disease in cerebral palsy (CP) and explore the automatic calculation of these measurements using artificial intelligence (AI) software. This was achieved by conducting: a systematic review, which identified all the different radiographic measurements currently used in the literature; a Delphi study, which formed a core measurement set (CMS) including the most important measurements in the opinion of orthopaedic surgeons and physiotherapists; and a preliminary analysis of our newly proposed AI system, which assessed the performance and reliability of the model.

The systematic review identified 15 distinct measurements from the literature, including Reimers' migration percentage (RMP), Femoral neck-shaft angle, Acetabular index, Femoral head-shaft angle (HSA), Centre edge angle, Sharp's angle or acetabular angle, Acetabular depth ratio, Mose hip ratio, Shenton's line, Epiphyseal shaft angle, Pelvic femoral angle, Pelvic adjusted migration percentage, Medialization index, and Hilgenreiner epiphyseal angle. Of these measurements, RMP was by far the most reported amongst the included studies. These 15 measurements then formed the basis for the Delphi study, which produced a final CMS with RMP and HSA. The preliminary analysis of our AI system demonstrated a 'good to excellent' inter-observer reliability between RMP values generated automatically by the model and manually by five clinicians, an adequate ability to automatically outline and segment the pelvis and proximal femur, and a 'good to very good' inter-and intra-observer reliability amongst the manually annotated radiographs in the training dataset used to teach the AI model. Having proven the feasibility of an AI system for calculating these measurements, this thesis has successfully identified a framework to automate the analysis of hip radiographs in children with CP.

The development of a CMS has provided the necessary measurements needed to build a functional AI system. Moreover, this CMS will reduce heterogeneity in the reporting of CP hip radiographs and facilitate meta-analysis whilst reducing the risk of reporting bias. Our AI model has contributed to the limited research in the field of CP hips by proposing a new technique for the calculation of radiographic measurements and demonstrating the competence of AI models in the calculation of these measurements. Additionally, the automatic analysis of CP hip radiographs will significantly impact the management of children with CP, especially if implemented in national hip surveillance programmes. Potential benefits include reduced costs to healthcare providers, consistent diagnoses, and standardised access to CP care. In regions with limited access to healthcare provisions, children with CP may not be monitored regularly via radiographs resulting in poorer outcomes. An automated system can help flatten

these inequalities and populate national databases facilitating population-wide research. Optimising the operational efficiency of national hip surveillance programmes can improve the management and outcome of children with CP.

Upcoming prototypes of our AI model have been planned to additionally measure HSA, after which a complete analysis of the model can be conducted. Future research will look at the implementation of the Uppföljningsprogram för Cerebral Pares (CPUP) hip score into our proposed model.(51) The calculation of the CPUP hip score can be easily applied given the inclusion of RMP and HSA in our model's programming. Automatically calculating the risk of hip displacement would further diminish the need for human intervention, as clinicians will not need to review the AI-derived measurements in order to assess the risk of hip displacement in a child. Furthermore, in addition to the automatic calculation of geometric measurements, future research should also focus on quantifying the shape and appearance of CP hips via the formation of Statistical Shape Models(345) and Statistical Appearance Models(346). This detailed level of analysis would enable AI systems to automatically differentiate between complex and subtle variations in the hip, further improving the prediction and diagnosis of CP hip disease.

This chapter has achieved the initial objectives to highlight the implications of the thesis results and suggest topics for future research. In conclusion, this thesis has investigated the use of radiographic measurements in the assessment of hip disease in CP and explored the application of AI-powered software to automate the calculation of these measurements.

References

1. Krigger KW. Cerebral palsy: an overview. American family physician. 2006;73(1):91-100.
2. Cooke P, Cole W, Carey R. Dislocation of the hip in cerebral palsy. Natural history and predictability. The Journal of bone and joint surgery British volume. 1989;71(3):441-6.
3. Penner M, Xie WY, Binepal N, Switzer L, Fehlings D. Characteristics of pain in children and youth with cerebral palsy. Pediatrics. 2013;132(2):e407-e13.
4. Bagg MR, Farber J, Miller F. Long-Term Follow-up of Hip Subluxation in Cerebral Palsy Patients. Journal of Pediatric Orthopaedics. 1993;13(1):32-6.
5. Ramstad K, Jahnsen RB, Terjesen T. Severe hip displacement reduces health-related quality of life in children with cerebral palsy: A population-based study of 67 children. Acta orthopaedica. 2017;88(2):205-10.
6. Boldingh EJ, Jacobs-van der Bruggen MA, Bos CF, Lankhorst GJ, Bouter LM. Determinants of hip pain in adult patients with severe cerebral palsy. Journal of pediatric Orthopaedics B. 2005;14(2):120-5.
7. Onimus M, Allamel G, Manzone P, Laurain J. Prevention of hip dislocation in cerebral palsy by early psoas and adductors tenotomies. Journal of pediatric orthopedics. 1991;11(4):432-5.
8. Phelps WM. Prevention of acquired dislocation of the hip in cerebral palsy. JBJS. 1959;41(3):440-8.
9. Picciolini O, Albisetti W, Cozzaglio M, Spreafico F, Mosca F, Gasparroni V. "Postural Management" to prevent hip dislocation in children with cerebral palsy. Hip International. 2009;19(6_suppl):56-62.
10. Sharrard W, Allen J, Heaney S. Surgical prophylaxis of subluxation and dislocation of the hip in cerebral palsy. The Journal of bone and joint surgery British volume. 1975;57(2):160-6.
11. Gaston MS. CPIPS: musculoskeletal and hip surveillance for children with cerebral palsy. Paediatrics and Child Health. 2019;29(11):489-94.
12. Sato Y, Takegami Y, Asamoto T, Ono Y, Hidetoshi T, Goto R, et al. Artificial intelligence improves the accuracy of residents in the diagnosis of hip fractures: a multicenter study. BMC musculoskeletal disorders. 2021;22(1):1-10.
13. Kang Y-J, Yoo J-I, Cha Y-H, Park CH, Kim J-T. Machine learning–based identification of hip arthroplasty designs. Journal of orthopaedic translation. 2020;21:13-7.
14. Karnuta JM, Haeberle HS, Luu BC, Roth AL, Molloy RM, Nystrom LM, et al. Artificial intelligence to identify arthroplasty implants from radiographs of the hip. The Journal of Arthroplasty. 2021;36(7):S290-S4. e1.
15. Gorelik N, Gyftopoulos S. Applications of artificial intelligence in musculoskeletal imaging: from the request to the report. Canadian Association of Radiologists Journal. 2021;72(1):45-59.
16. Sankar C, Mundkur N. Cerebral palsy-definition, classification, etiology and early diagnosis. The Indian Journal of Pediatrics. 2005;72(10):865-8.

17. Rosenbaum P, Paneth N, Levinton A, Goldstein M, Bax M, Damiano D, et al. The definition and classification of cerebral palsy. NeoReviews. 2006;7(11):e569.
18. Howard C, McKibbin B, Williams L, Mackie I. Factors affecting the incidence of hip dislocation in cerebral palsy. The Journal of bone and joint surgery British volume. 1985;67(4):530-2.
19. Soo B, Howard JJ, Boyd RN, Reid SM, Lanigan A, Wolfe R, et al. Hip displacement in cerebral palsy. JBJS. 2006;88(1):121-9.
20. Ramstad K, Terjesen T. Hip pain is more frequent in severe hip displacement: a population-based study of 77 children with cerebral palsy. Journal of Pediatric Orthopaedics B. 2016;25(3):217-21.
21. McIVOR WC, SAMILSON RL. Fractures in patients with cerebral palsy. JBJS. 1966;48(5):858-66.
22. SAMILSON RL, Tsou P, AAMOTH G, GREEN WM. Dislocation and subluxation of the hip in cerebral palsy: pathogenesis, natural history and management. JBJS. 1972;54(4):863-73.
23. Knapp Jr DR, Cortes H. Untreated hip dislocation in cerebral palsy. Journal of Pediatric Orthopaedics. 2002;22(5):668-71.
24. Michael-Asalu A, Taylor G, Campbell H, Lelea L-L, Kirby RS. Cerebral palsy: diagnosis, epidemiology, genetics, and clinical update. Advances in pediatrics. 2019;66:189-208.
25. Oskoui M, Coutinho F, Dykeman J, Jette N, Pringsheim T. An update on the prevalence of cerebral palsy: a systematic review and meta-analysis. Developmental Medicine & Child Neurology. 2013;55(6):509-19.
26. Romeo DM, Sini F, Brogna C, Albamonte E, Ricci D, Mercuri E. Sex differences in cerebral palsy on neuromotor outcome: a critical review. Developmental Medicine & Child Neurology. 2016;58(8):809-13.
27. Paneth N, Hong T, Korzeniewski S. The descriptive epidemiology of cerebral palsy. Clinics in perinatology. 2006;33(2):251-67.
28. Hack M, Costello DW. Trends in the rates of cerebral palsy associated with neonatal intensive care of preterm children. Clinical obstetrics and gynecology. 2008;51(4):763-74.
29. Huser A, Mo M, Hosseinzadeh P. Hip surveillance in children with cerebral palsy. Orthopedic Clinics. 2018;49(2):181-90.
30. Hägglund G, Lauge-Pedersen H, Wagner P. Characteristics of children with hip displacement in cerebral palsy. BMC musculoskeletal disorders. 2007;8(1):1-6.
31. Terjesen T. The natural history of hip development in cerebral palsy. Developmental medicine & child neurology. 2012;54(10):951-7.
32. Larnert P, Risto O, Hägglund G, Wagner P. Hip displacement in relation to age and gross motor function in children with cerebral palsy. Journal of children's orthopaedics. 2014;8(2):129-34.
33. Chang CH, Wang YC, Ho PC, Hwang AW, Kao HK, Lee WC, et al. Determinants of Hip Displacement in Children With Cerebral Palsy. Clinical Orthopaedics and Related Research®. 2015;473(11):3675-81.
34. Stanley FJ, Blair E, Alberman E. Cerebral palsies: epidemiology and causal pathways: Cambridge University Press; 2000.

35. Eunson P. Aetiology and epidemiology of cerebral palsy. Paediatrics and Child Health. 2012;22(9):361-6.
36. O'Callaghan ME, MacLennan AH, Gibson CS, McMichael GL, Haan EA, Broadbent JL, et al. Epidemiologic associations with cerebral palsy. Obstetrics & Gynecology. 2011;118(3):576-82.
37. Reddihough DS, Collins KJ. The epidemiology and causes of cerebral palsy. Australian Journal of physiotherapy. 2003;49(1):7-12.
38. Agarwal A, Verma I. Cerebral palsy in children: An overview. Journal of clinical orthopaedics and trauma. 2012;3(2):77-81.
39. Miller F. Etiology of Hip Displacement in Children with Cerebral Palsy. Cerebral Palsy. 2020:1881-91.
40. Eldessouky A, Smeda G. Hip Dislocation in Cerebral Palsy: Treatment Options. J Orthop Res Physiother. 2016;2:026.
41. Hosseinzadeh P, Baldwin K, Minaie A, Miller F. Management of hip disorders in patients with cerebral palsy. JBJS reviews. 2020;8(3):e0148.
42. Marcström A, Hägglund G, Alriksson-Schmidt AI. Hip pain in children with cerebral palsy: a population-based registry study of risk factors. BMC musculoskeletal disorders. 2019;20(1):1-10.
43. (CPIPS) CPIPS. Origins and Development Core Dataset Clinical Assessment. 2017.
44. Pountney T, Green EM. Hip dislocation in cerebral palsy. Bmj. 2006;332(7544):772-5.
45. Reimers J. The stability of the hip in children: a radiological study of the results of muscle surgery in cerebral palsy. Acta Orthopaedica Scandinavica. 1980;51(sup184):1-100.
46. Bobroff ED, Chambers HG, Sartoris DJ, Wyatt MP, Sutherland DH. Femoral anteversion and neck-shaft angle in children with cerebral palsy. Clinical Orthopaedics and Related Research®. 1999;364:194-204.
47. Robin J, Graham HK, Selber P, Dobson F, Smith K, Baker R. Proximal femoral geometry in cerebral palsy: a population-based cross-sectional study. The Journal of bone and joint surgery British volume. 2008;90(10):1372-9.
48. Oetgen ME, Ayyala H, Martin BD. Treatment of hip subluxation in skeletally mature patients with cerebral palsy. Orthopedics. 2015;38(4):e248-e52.
49. Tönnis D. Normal values of the hip joint for the evaluation of X-rays in children and adults. Clinical orthopaedics and related research. 1976(119):39-47.
50. SOUTHWICK WO. Osteotomy through the lesser trochanter for slipped capital femoral epiphysis. JBJS. 1967;49(5):807-35.
51. Hermanson M, Hägglund G, Riad J, Rodby-Bousquet E, Wagner P. Prediction of hip displacement in children with cerebral palsy: development of the CPUP hip score. The bone & joint journal. 2015;97(10):1441-4.
52. Shore BJ, Martinkevich P, Riazi M, Baird E, Encisa C, Willoughby K, et al. Reliability of radiographic assessments of the hip in cerebral palsy. Journal of Pediatric Orthopaedics. 2019;39(7):e536-e41.
53. Craven A, Pym A, Boyd RN. Reliability of Radiologic Measures of Hip Displacement in a Cohort of Preschool-aged Children With Cerebral Palsy. Journal of Pediatric Orthopaedics. 2014;34(6):597-602.

54. Wek C, Chowdhury P, Smith C, Kokkinakis M. Is the Gothic Arch a reliable radiographic landmark for migration percentage in children with cerebral palsy? Journal of Children's Orthopaedics. 2020;14(5):397-404.
55. Terjesen T. Development of the hip joints in unoperated children with cerebral palsy: a radiographic study of 76 patients. Acta orthopaedica. 2006;77(1):125-31.
56. Hägglund G, Andersson S, Düppe H, Lauge-Pedersen H, Nordmark E, Westbom L. Prevention of dislocation of the hip in children with cerebral palsy: the first ten years of a population-based prevention programme. The Journal of Bone and Joint Surgery British volume. 2005;87(1):95-101.
57. Scrutton D, Baird G, Smeeton N. Hip dysplasia in bilateral cerebral palsy: incidence and natural history in children aged 18 months to 5 years. Developmental medicine and child neurology. 2001;43(9):586-600.
58. Robin J, Graham HK, Baker R, Selber P, Simpson P, Symons S, et al. A classification system for hip disease in cerebral palsy. Developmental Medicine & Child Neurology. 2009;51(3):183-92.
59. Connelly A, Flett P, Graham HK, Oates J. Hip surveillance in Tasmanian children with cerebral palsy. Journal of paediatrics and child health. 2009;45(7-8):437-43.
60. Terjesen T, Runden TØ, Tangerud Å. Ultrasonography and radiography of the hip in infants. Acta Orthopaedica Scandinavica. 1989;60(6):651-60.
61. Noordin S, Umer M, Hafeez K, Nawaz H. Developmental dysplasia of the hip. Orthopedic Reviews. 2010;2(2).
62. Bayusentono S, Choi Y, Chung CY, Kwon S-S, Lee KM, Park MS. Recurrence of hip instability after reconstructive surgery in patients with cerebral palsy. JBJS. 2014;96(18):1527-34.
63. Cho Y, Park ES, Park HK, Park JE, Rha D-w. Determinants of hip and femoral deformities in children with spastic cerebral palsy. Annals of Rehabilitation Medicine. 2018;42(2):277.
64. Flynn JM, Miller F. Management of hip disorders in patients with cerebral palsy. JAAOS-Journal of the American Academy of Orthopaedic Surgeons. 2002;10(3):198-209.
65. Selva G, Miller F, Dabney KW. Anterior hip dislocation in children with cerebral palsy. Journal of Pediatric Orthopaedics. 1998;18(1):54-61.
66. Sugano N, Noble PC, Kamaric E. A comparison of alternative methods of measuring femoral anteversion. Journal of computer assisted tomography. 1998;22(4):610-4.
67. Cans C. Surveillance of cerebral palsy in Europe: a collaboration of cerebral palsy surveys and registers. Developmental Medicine & Child Neurology. 2000;42(12):816-24.
68. Rana M, Upadhyay J, Rana A, Durgapal S, Jantwal A. A systematic review on etiology, epidemiology, and treatment of cerebral palsy. International Journal of Nutrition, Pharmacology, Neurological Diseases. 2017;7(4):76.
69. Eggink H, Kremer D, Brouwer OF, Contarino MF, van Egmond ME, Elema A, et al. Spasticity, dyskinesia and ataxia in cerebral palsy: are we sure we can differentiate them? European Journal of Paediatric Neurology. 2017;21(5):703-6.

70. Blair E, Stanley F. Interobserver agreement in the classification of cerebral palsy. Developmental Medicine & Child Neurology. 1985;27(5):615-22.
71. Palisano R, Rosenbaum P, Walter S, Russell D, Wood E, Galuppi B. Development and reliability of a system to classify gross motor function in children with cerebral palsy. Developmental medicine & child neurology. 1997;39(4):214-23.
72. Palisano RJ, Hanna SE, Rosenbaum PL, Russell DJ, Walter SD, Wood EP, et al. Validation of a model of gross motor function for children with cerebral palsy. Physical therapy. 2000;80(10):974-85.
73. Palisano R, Rosenbaum P, Bartlett D, Livingston M. Gross Motor Function Classification System: expanded and revised 2007. Canada, CanChild Centre for Childhood Disability Research, McMaster University. 2010.
74. Shrader M, Koenig A, Falk M, Belthur M, Boan C. An independent assessment of reliability of the melbourne cerebral palsy hip classification system. Journal of children's orthopaedics. 2017;11(5):334-8.
75. Gose S, Sakai T, Shibata T, Akiyama K, Yoshikawa H, Sugamoto K. Verification of the Robin and Graham classification system of hip disease in cerebral palsy using three-dimensional computed tomography. Developmental Medicine & Child Neurology. 2011;53(12):1107-12.
76. Robb JE, Hägglund G. Hip surveillance and management of the displaced hip in cerebral palsy. Journal of Children's Orthopaedics. 2013;7(5):407-13.
77. Noonan KJ, Jones J, Pierson J, Honkamp NJ, Leverson G. Hip function in adults with severe cerebral palsy. JBJS. 2004;86(12):2607-13.
78. Martinsson C, Himmelmann K. Effect of Weight-Bearing in Abduction and Extension on Hip Stability in Children With Cerebral Palsy. Pediatric Physical Therapy. 2011;23(2):150-7.
79. Pountney TE, Mandy A, Green E, Gard PR. Hip subluxation and dislocation in cerebral palsy–a prospective study on the effectiveness of postural management programmes. Physiotherapy Research International. 2009;14(2):116-27.
80. Hoffer M. Management of the hip in cerebral palsy. JBJS. 1986;68(4):629-31.
81. Shore BJ, Graham HK. Management of Moderate to Severe Hip Displacement in Nonambulatory Children with Cerebral Palsy. JBJS Reviews. 2017;5(12):e4.
82. Graham HK, Selber P. Musculoskeletal aspects of cerebral palsy. The Journal of bone and joint surgery British volume. 2003;85(2):157-66.
83. Graham HK, Rosenbaum P, Paneth N, Dan B, Lin JP, Damiano DL, et al. Cerebral palsy. Nat Rev Dis Primers. 2016;2:15082.
84. Brunner R, Baumann JU. Long-Term Effects of Intertrochanteric Varus-Derotation Osteotomy on Femur and Acetabulum in Spastic Cerebral Palsy: An 11- to 18-Year Follow-up Study. Journal of Pediatric Orthopaedics. 1997;17(5):585-91.
85. Wright D, Whyne C, Hardisty M, Kreder HJ, Lubovsky O. Functional and anatomic orientation of the femoral head. Clinical Orthopaedics and Related Research®. 2011;469(9):2583-9.
86. Miller F, Girardi H, Lipton G, Ponzio R, Klaumann M, Dabney KW. Reconstruction of the Dysplastic Spastic Hip with Peri-ilial Pelvic and Femoral Osteotomy Followed by Immediate Mobilization. Journal of Pediatric Orthopaedics. 1997;17(5):592-602.

87. McNerney NP, Mubarak SJ, Wenger DR. One-Stage Correction of the Dysplastic Hip in Cerebral Palsy with the San Diego Acetabuloplasty: Results and Complications in 104 Hips. Journal of Pediatric Orthopaedics. 2000;20(1):93.
88. Dega W, Król J, Polakowski L. Surgical treatment of congenital dislocation of the hip in children: a one-stage procedure. JBJS. 1959;41(5):920-34.
89. Boldingh EJ, Bouwhuis CB, van der Heijden-Maessen HC, Bos CF, Lankhorst GJ. Palliative hip surgery in severe cerebral palsy: a systematic review. Journal of Pediatric Orthopaedics B. 2014;23(1):86-92.
90. Castle M, Schneider C. Proximal femoral resection-interposition arthroplasty. JBJS. 1978;60(8):1051-4.
91. Queally J, Abdulkarim A, Mulhall K. Total hip replacement in patients with neurological conditions. The Journal of Bone and Joint Surgery British volume. 2009;91(10):1267-73.
92. Gabos PG, Miller F, Galban MA, Gupta GG, Dabney K. Prosthetic interposition arthroplasty for the palliative treatment of end-stage spastic hip disease in nonambulatory patients with cerebral palsy. Journal of Pediatric Orthopaedics. 1999;19(6):796.
93. Hogan KA, Blake M, Gross RH. Subtrochanteric valgus osteotomy for chronically dislocated, painful spastic hips. JBJS. 2006;88(12):2624-31.
94. Leet AI, Chhor K, Launay F, Kier-York J, Sponseller PD. Femoral head resection for painful hip subluxation in cerebral palsy: is valgus osteotomy in conjunction with femoral head resection preferable to proximal femoral head resection and traction? Journal of Pediatric Orthopaedics. 2005;25(1):70-3.
95. Root L, Goss J, Mendes J. The treatment of the painful hip in cerebral palsy by total hip replacement or hip arthrodesis. The Journal of Bone and Joint surgery American Volume. 1986;68(4):590-8.
96. Souza RCd, Mansano MV, Bovo M, Yamada HH, Rancan DR, Fucs PMdMB, et al. Hip salvage surgery in cerebral palsy cases: a systematic review. Revista Brasileira de Ortopedia. 2015;50:254-9.
97. Tachdjian MO, Minear WL. Hip dislocation in cerebral palsy. JBJS. 1956;38(6):1358-64.
98. Hollung SJ. Surveillance of cerebral palsy in Norway; a national registry-based study. 2019.
99. Rasmussen HM, Nordbye-Nielsen K, Møller-Madsen B, Johansen M, Ellitsgaard N, Pedersen CR, et al. The Danish cerebral palsy follow-up program. Clinical epidemiology. 2016;8:457.
100. BritishColumbiaHipSurveillancePlanningCommittee. British Columbia Consensus Statement on Hip Surveillance for Children with Cerebral Palsy: Information for health care professionals caring for children with cerebral palsy.
101. Wynter M, Gibson N, Willoughby KL, Love S, Kentish M, Thomason P, et al. Australian hip surveillance guidelines for children with cerebral palsy: 5-year review. Developmental Medicine & Child Neurology. 2015;57(9):808-20.
102. Wordie SJ, Robb JE, Hägglund G, Bugler KE, Gaston MS. Hip displacement and dislocation in a total population of children with cerebral palsy in Scotland: status after five years' hip surveillance. The Bone & Joint Journal. 2020;102(3):383-7.

103. Hägglund G, Alriksson-Schmidt A, Lauge-Pedersen H, Rodby-Bousquet E, Wagner P, Westbom L. Prevention of dislocation of the hip in children with cerebral palsy: 20-year results of a population-based prevention programme. The bone & joint journal. 2014;96(11):1546-52.
104. Elkamil AI, Andersen GL, Hägglund G, Lamvik T, Skranes J, Vik T. Prevalence of hip dislocation among children with cerebral palsy in regions with and without a surveillance programme: a cross sectional study in Sweden and Norway. BMC musculoskeletal disorders. 2011;12(1):1-7.
105. Dobson F, Boyd R, Parrott J, Nattrass G, Graham H. Hip surveillance in children with cerebral palsy: impact on the surgical management of spastic hip disease. The Journal of bone and joint surgery British volume. 2002;84(5):720-6.
106. Jeglinsky I, Alriksson-Schmidt AI, Hägglund G, Ahonen M. Prevalence and treatment of hip displacement in children with cerebral palsy in Finland. Journal of Children's Orthopaedics. 2022;16(2):128-35.
107. Gordon G, Simkiss D. A systematic review of the evidence for hip surveillance in children with cerebral palsy. The Journal of Bone and Joint Surgery British volume. 2006;88(11):1492-6.
108. Kaplan A, Haenlein M. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons. 2019;62(1):15-25.
109. Hodges A. Alan Turing: the enigma. Alan Turing: The Enigma: Princeton University Press; 2014.
110. Haenlein M, Kaplan A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. California management review. 2019;61(4):5-14.
111. Turing AM, Haugeland J. Computing machinery and intelligence. The Turing Test: Verbal Behavior as the Hallmark of Intelligence. 1950:29-56.
112. Minsky M, Papert S. An introduction to computational geometry. Cambridge tiass, HIT. 1969;479:480.
113. Muthukrishnan N, Maleki F, Ovens K, Reinhold C, Forghani B, Forghani R. Brief history of artificial intelligence. Neuroimaging Clinics. 2020;30(4):393-9.
114. Khakurel J, Penzenstadler B, Porras J, Knutas A, Zhang W. The rise of artificial intelligence under the lens of sustainability. Technologies. 2018;6(4):100.
115. Weaver CE, Lazaros EJ, Zhao JJ, Davison CB, Truell AD. The Internet of Things: An overview of selected smart home technology. Issues in Information Systems. 2020;21(2):43.
116. Yaqoob I, Khan LU, Kazmi SA, Imran M, Guizani N, Hong CS. Autonomous driving cars in smart cities: Recent advances, requirements, and challenges. IEEE Network. 2019;34(1):174-81.
117. McGovern A, Elmore KL, Gagne DJ, Haupt SE, Karstens CD, Lagerquist R, et al. Using artificial intelligence to improve real-time decision-making for high-impact weather. Bulletin of the American Meteorological Society. 2017;98(10):2073-90.
118. Beel J, Gipp B, editors. Google Scholar's ranking algorithm: an introductory overview. Proceedings of the 12th international conference on scientometrics and informetrics (ISSI'09); 2009: Rio de Janeiro (Brazil).

119. Nti IK, Adekoya AF, Weyori BA, Nyarko-Boateng O. Applications of artificial intelligence in engineering and manufacturing: A systematic review. Journal of Intelligent Manufacturing. 2021:1-21.
120. Samuel A. Eight-move opening utilizing generalization learning,(See Appendix B, Game G-43.1 Some Studies in Machine Learning Using the Game of Checkers). IBM Journal. 1959:210-29.
121. Mitchell TM. Machine learning. McGraw-hill New York; 1997.
122. Alpaydin E. Introduction to machine learning: MIT press; 2020.
123. Lindebaum D, Vesa M, Den Hond F. Insights from "the machine stops" to better understand rational assumptions in algorithmic decision making and its implications for organizations. Academy of Management Review. 2020;45(1):247-63.
124. El Naqa I, Murphy MJ. What is machine learning? machine learning in radiation oncology: Springer; 2015. p. 3-11.
125. Learned-Miller EG. Introduction to supervised learning. I: Department of Computer Science, University of Massachusetts. 2014:3.
126. Ongsulee P, editor Artificial intelligence, machine learning and deep learning. 2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE); 2017: IEEE.
127. Nasteski V. An overview of the supervised machine learning methods. Horizons b. 2017;4:51-62.
128. Caruana R, Niculescu-Mizil A, editors. An empirical comparison of supervised learning algorithms. Proceedings of the 23rd international conference on Machine learning; 2006.
129. Crawford M, Khoshgoftaar TM, Prusa JD, Richter AN, Al Najada H. Survey of review spam detection using machine learning techniques. Journal of Big Data. 2015;2(1):1-24.
130. Labhsetwar S. Predictive analysis of customer churn in telecom industry using supervised learning. ICTACT Journal on Soft Computing. 2020;10(2):2054-60.
131. Pellegretti P, Roli F, Serpico SB, Vernazza G. Supervised learning of descriptions for image recognition purposes. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1994;16(1):92-8.
132. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction: Springer; 2009.
133. Greene D, Cunningham P, Mayer R. Unsupervised learning and clustering. Machine learning techniques for multimedia: Springer; 2008. p. 51-90.
134. Cios KJ, Swiniarski RW, Pedrycz W, Kurgan LA, editors. Unsupervised learning: Association rules. Data Mining; 2007: Springer.
135. Celebi ME, Aydin K. Unsupervised learning algorithms: Springer; 2016.
136. Berry MW, Mohamed A, Yap BW. Supervised and unsupervised learning for data science: Springer; 2019.
137. Hinton GE. Learning multiple layers of representation. Trends in cognitive sciences. 2007;11(10):428-34.
138. Du X, Cai Y, Wang S, Zhang L, editors. Overview of deep learning. 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC); 2016: IEEE.

139. Jain AK, Mao J, Mohiuddin KM. Artificial neural networks: A tutorial. Computer. 1996;29(3):31-44.
140. Da Silva IN, Spatti DH, Flauzino RA, Liboni LHB, dos Reis Alves SF. Artificial neural networks. Cham: Springer International Publishing. 2017;39.
141. Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G. Backpropagation and the brain. Nature Reviews Neuroscience. 2020;21(6):335-46.
142. Nargesian F, Samulowitz H, Khurana U, Khalil EB, Turaga DS, editors. Learning Feature Engineering for Classification. Ijcai; 2017.
143. LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521(7553):436-44.
144. Albawi S, Mohammed TA, Al-Zawi S, editors. Understanding of a convolutional neural network. 2017 international conference on engineering and technology (ICET); 2017: Ieee.
145. O'Shea K, Nash R. An introduction to convolutional neural networks. arXiv preprint arXiv:151108458. 2015.
146. Schmidt RM. Recurrent neural networks (rnns): A gentle introduction and overview. arXiv preprint arXiv:191205911. 2019.
147. Zhang C, Ma Y. Ensemble machine learning: methods and applications: Springer; 2012.
148. Breiman L. Bagging predictors. Machine learning. 1996;24(2):123-40.
149. Freund Y, Schapire RE, editors. Experiments with a new boosting algorithm. icml; 1996: Citeseer.
150. Opitz D, Maclin R. Popular ensemble methods: An empirical study. Journal of artificial intelligence research. 1999;11:169-98.
151. Kotsiantis S, Pintelas P. Combining bagging and boosting. International Journal of Computational Intelligence. 2004;1(4):324-33.
152. Schapire RE. Explaining adaboost. Empirical inference: Springer; 2013. p. 37-52.
153. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Frontiers in neurorobotics. 2013;7:21.
154. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.
155. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.
156. Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society. 2004;18(6):275-85.
157. Song Y-Y, Ying L. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry. 2015;27(2):130.
158. Dietterich TG, Kong EB. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Citeseer; 1995.
159. Bramer M. Avoiding overfitting of decision trees. Principles of data mining. 2007:119-34.
160. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC bioinformatics. 2006;7(1):1-13.
161. Tang F, Ishwaran H. Random forest missing data algorithms. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2017;10(6):363-77.

162. Liu Y, Wang Y, Zhang J, editors. New machine learning algorithm: Random forest. International Conference on Information Computing and Applications; 2012: Springer.
163. Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. Journal of family medicine and primary care. 2019;8(7):2328-31.
164. Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow. Frontiers in medicine. 2020;7:27.
165. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017;69:S36-S40.
166. Shen J, Zhang CJ, Jiang B, Chen J, Song J, Liu Z, et al. Artificial intelligence versus clinicians in disease diagnosis: systematic review. JMIR medical informatics. 2019;7(3):e10010.
167. Le E, Wang Y, Huang Y, Hickman S, Gilbert F. Artificial intelligence in breast imaging. Clinical radiology. 2019;74(5):357-66.
168. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. Trends in pharmacological sciences. 2019;40(8):577-91.
169. Panesar S, Cagle Y, Chander D, Morey J, Fernandez-Miranda J, Kliot M. Artificial intelligence and the future of surgical robotics. Annals of surgery. 2019;270(2):223-6.
170. Singh AV, Ansari MHD, Laux P, Luch A. Micro-nanorobots: important considerations when developing novel drug delivery platforms. Expert Opinion on Drug Delivery. 2019;16(11):1259-75.
171. Pouke M, Häkkilä J. Elderly healthcare monitoring using an avatar-based 3D virtual environment. International journal of environmental research and public health. 2013;10(12):7283-98.
172. Ramesh A, Kambhampati C, Monson JR, Drew P. Artificial intelligence in medicine. Annals of the Royal College of Surgeons of England. 2004;86(5):334.
173. Barragán-Montero A, Javaid U, Valdés G, Nguyen D, Desbordes P, Macq B, et al. Artificial intelligence and machine learning for medical imaging: A technology review. Physica Medica. 2021;83:242-56.
174. Boland GW, Guimaraes AS, Mueller PR. The radiologist's conundrum: benefits and costs of increasing CT capacity and utilization. European radiology. 2009;19(1):9-11.
175. McDonald RJ, Schwartz KM, Eckel LJ, Diehn FE, Hunt CH, Bartholmai BJ, et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Academic radiology. 2015;22(9):1191-8.
176. Fitzgerald R. Error in radiology. Clinical radiology. 2001;56(12):938-46.
177. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nature Reviews Cancer. 2018;18(8):500-10.
178. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical image analysis. 2017;42:60-88.
179. Paul R, Hawkins SH, Balagurunathan Y, Schabath M, Gillies RJ, Hall LO, et al. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. Tomography. 2016;2(4):388-95.

180. Cheng J-Z, Ni D, Chou Y-H, Qin J, Tiu C-M, Chang Y-C, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Scientific reports. 2016;6(1):1-13.
181. Ghafoorian M, Karssemeijer N, Heskes T, van Uden IW, Sanchez CI, Litjens G, et al. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. Scientific Reports. 2017;7(1):1-12.
182. Milea D, Singhal S, Najjar RP. Artificial intelligence for detection of optic disc abnormalities. Current Opinion in Neurology. 2020;33(1):106-10.
183. Alam M, Le D, Lim JI, Chan RV, Yao X. Supervised machine learning based multi-task artificial intelligence classification of retinopathies. Journal of clinical medicine. 2019;8(6):872.
184. Patriarche JW, Erickson BJ. Part 1. Automated change detection and characterization in serial MR studies of brain-tumor patients. Journal of Digital Imaging. 2007;20(3):203-22.
185. Chen CW, Luo J, Parker KJ. Image segmentation via adaptive K-mean clustering and knowledge-based morphological operations with biomedical applications. IEEE transactions on image processing. 1998;7(12):1673-83.
186. Ahishakiye E, Bastiaan Van Gijzen M, Tumwiine J, Wario R, Obungoloch J. A survey on deep learning in medical image reconstruction. Intelligent Medicine. 2021;1(03):118-27.
187. Riquelme D, Akhloufi MA. Deep learning for lung cancer nodules detection and classification in CT scans. AI. 2020;1(1):28-67.
188. Sim Y, Chung MJ, Kotter E, Yune S, Kim M, Do S, et al. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. Radiology. 2020;294(1):199-209.
189. Nasrullah N, Sang J, Alam MS, Mateen M, Cai B, Hu H. Automated lung nodule detection and classification using deep learning combined with multiple strategies. Sensors. 2019;19(17):3722.
190. Dey R, Lu Z, Hong Y, editors. Diagnostic classification of lung nodules using 3D neural networks. 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018); 2018: IEEE.
191. Bar Y, Diamant I, Wolf L, Greenspan H, editors. Deep learning with non-medical training used for chest pathology identification. Medical Imaging 2015: Computer-Aided Diagnosis; 2015: International Society for Optics and Photonics.
192. Shin H-C, Roberts K, Lu L, Demner-Fushman D, Yao J, Summers RM, editors. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
193. Zebin T, Rezvy S. COVID-19 detection and disease progression visualization: Deep learning on chest X-rays for classification and coarse localization. Applied Intelligence. 2021;51(2):1010-21.
194. Rahman T, Khandakar A, Kadir MA, Islam KR, Islam KF, Mazhar R, et al. Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. IEEE Access. 2020;8:191586-601.
195. Ohno Y, Aoyagi K, Takenaka D, Yoshikawa T, Ikezaki A, Fujisawa Y, et al. Machine learning for lung CT texture analysis: Improvement of inter-observer

agreement for radiological finding classification in patients with pulmonary diseases. European journal of radiology. 2021;134:109410.
196. Huang S, Lee F, Miao R, Si Q, Lu C, Chen Q. A deep convolutional neural network architecture for interstitial lung disease pattern classification. Medical & biological engineering & computing. 2020;58(4):725-37.
197. Suk H-I, Lee S-W, Shen D. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Structure and Function. 2015;220(2):841-59.
198. Kim J, Calhoun VD, Shim E, Lee J-H. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. Neuroimage. 2016;124:127-46.
199. Plis SM, Hjelm DR, Salakhutdinov R, Allen EA, Bockholt HJ, Long JD, et al. Deep learning for neuroimaging: a validation study. Frontiers in neuroscience. 2014;8:229.
200. Tikka SK, Singh BK, Nizamie SH, Garg S, Mandal S, Thakur K, et al. Artificial intelligence-based classification of schizophrenia: A high density electroencephalographic and support vector machine study. Indian Journal of Psychiatry. 2020;62(3):273.
201. Salehi AW, Baglat P, Sharma BB, Gupta G, Upadhya A, editors. A CNN model: earlier diagnosis and classification of Alzheimer disease using MRI. 2020 International Conference on Smart Electronics and Communication (ICOSEC); 2020: IEEE.
202. Birenbaum A, Greenspan H. Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks. Deep learning and data labeling for medical applications: Springer; 2016. p. 58-67.
203. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE transactions on medical imaging. 2016;35(5):1240-51.
204. Pan Y, Huang W, Lin Z, Zhu W, Zhou J, Wong J, et al., editors. Brain tumor grading based on neural networks and convolutional neural networks. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2015: IEEE.
205. Dou Q, Chen H, Yu L, Zhao L, Qin J, Wang D, et al. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. IEEE transactions on medical imaging. 2016;35(5):1182-95.
206. Ghafoorian M, Karssemeijer N, Heskes T, Bergkamp M, Wissink J, Obels J, et al. Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin. NeuroImage: Clinical. 2017;14:391-9.
207. Chen X, Xu Y, Wong DWK, Wong TY, Liu J, editors. Glaucoma detection based on deep convolutional neural network. 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC); 2015: IEEE.
208. Abràmoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Investigative ophthalmology & visual science. 2016;57(13):5200-6.

209. Van Grinsven MJ, van Ginneken B, Hoyng CB, Theelen T, Sánchez CI. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. IEEE transactions on medical imaging. 2016;35(5):1273-84.
210. Zhu C, Zou B, Zhao R, Cui J, Duan X, Chen Z, et al. Retinal vessel segmentation in colour fundus images using extreme learning machine. Computerized Medical Imaging and Graphics. 2017;55:68-77.
211. Liu Q, Zou B, Chen J, Ke W, Yue K, Chen Z, et al. A location-to-segmentation strategy for automatic exudate segmentation in colour retinal fundus images. Computerized medical imaging and graphics. 2017;55:78-86.
212. Amrane M, Oukid S, Gagaoua I, Ensari T, editors. Breast cancer classification using machine learning. 2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT); 2018: IEEE.
213. Wu J, Hicks C. Breast cancer type classification using machine learning. Journal of personalized medicine. 2021;11(2):61.
214. Al Bataineh A. A comparative analysis of nonlinear machine learning algorithms for breast cancer detection. International Journal of Machine Learning and Computing. 2019;9(3):248-54.
215. Assiri AS, Nazir S, Velastin SA. Breast tumor classification using an ensemble machine learning method. Journal of Imaging. 2020;6(6):39.
216. Nindrea RD, Aryandono T, Lazuardi L, Dwiprahasto I. Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: a meta-analysis. Asian Pacific journal of cancer prevention: APJCP. 2018;19(7):1747.
217. Fotin SV, Yin Y, Haldankar H, Hoffmeister JW, Periaswamy S, editors. Detection of soft tissue densities from digital breast tomosynthesis: comparison of conventional and deep learning approaches. Medical Imaging 2016: Computer-Aided Diagnosis; 2016: SPIE.
218. Dhungel N, Carneiro G, Bradley AP, editors. The automated learning of deep features for breast mass classification from mammograms. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2016: Springer.
219. Zhang Q, Xiao Y, Dai W, Suo J, Wang C, Shi J, et al. Deep learning based classification of breast tumors with shear-wave elastography. Ultrasonics. 2016;72:150-7.
220. Abdelmaguid E, Huang J, Kenchareddy S, Singla D, Wilke L, Nguyen MH, et al. Left ventricle segmentation and volume estimation on cardiac MRI using deep learning. arXiv preprint arXiv:180906247. 2018.
221. Emad O, Yassine IA, Fahmy AS, editors. Automatic localization of the left ventricle in cardiac MRI images using deep learning. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2015: IEEE.
222. Zhang L, Gooya A, Dong B, Hua R, Petersen SE, Medrano-Gracia P, et al., editors. Automated quality assessment of cardiac MR images using convolutional neural networks. International Workshop on Simulation and Synthesis in Medical Imaging; 2016: Springer.
223. Lessmann N, Išgum I, Setio AA, de Vos BD, Ciompi F, de Jong PA, et al., editors. Deep convolutional neural networks for automatic coronary calcium scoring in

a screening study with low-dose chest CT. Medical Imaging 2016: Computer-Aided Diagnosis; 2016: SPIE.
224. Moradi M, Guo Y, Gur Y, Negahdar M, Syeda-Mahmood T, editors. A cross-modality neural network transform for semi-automatic medical image annotation. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2016: Springer.
225. Altini N, Prencipe B, Cascarano GD, Brunetti A, Brunetti G, Triggiani V, et al. Liver, kidney and spleen segmentation from CT scans and MRI with deep learning: A survey. Neurocomputing. 2022;490:30-53.
226. Bobo MF, Bao S, Huo Y, Yao Y, Virostko J, Plassard AJ, et al., editors. Fully convolutional neural networks improve abdominal organ segmentation. Medical Imaging 2018: Image Processing; 2018: International Society for Optics and Photonics.
227. Gibson E, Giganti F, Hu Y, Bonmati E, Bandula S, Gurusamy K, et al. Automatic multi-organ segmentation on abdominal CT with dense v-networks. IEEE transactions on medical imaging. 2018;37(8):1822-34.
228. Christ PF, Elshaer MEA, Ettlinger F, Tatavarty S, Bickel M, Bilic P, et al., editors. Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. International conference on medical image computing and computer-assisted intervention; 2016: Springer.
229. Farag A, Lu L, Roth HR, Liu J, Turkbey E, Summers RM. A bottom-up approach for pancreas segmentation using cascaded superpixels and (deep) image patch labeling. IEEE Transactions on Image Processing. 2016;26(1):386-99.
230. Cha KH, Hadjiiski LM, Samala RK, Chan H-P, Cohan RH, Caoili EM, et al. Bladder cancer segmentation in CT for treatment response assessment: application of deep-learning convolution neural network—a pilot study. Tomography. 2016;2(4):421-9.
231. Wang P, Xiao X, Glissen Brown JR, Berzin TM, Tu M, Xiong F, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nature biomedical engineering. 2018;2(10):741-8.
232. Lee JY, Jeong J, Song EM, Ha C, Lee HJ, Koo JE, et al. Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets. Scientific reports. 2020;10(1):1-9.
233. Khorasani HM, Usefi H, Peña-Castillo L. Detecting ulcerative colitis from colon samples using efficient feature selection and machine learning. Scientific reports. 2020;10(1):1-9.
234. Näppi JJ, Hironaka T, Regge D, Yoshida H, editors. Deep transfer learning of virtual endoluminal views for the detection of polyps in CT colonography. Medical imaging 2016: computer-aided diagnosis; 2016: SPIE.
235. Tajbakhsh N, Gurudu SR, Liang J, editors. A comprehensive computer-aided polyp detection system for colonoscopy videos. International Conference on Information Processing in Medical Imaging; 2015: Springer.
236. Baumgartner CF, Kamnitsas K, Matthew J, Smith S, Kainz B, Rueckert D, editors. Real-time standard scan plane detection and localisation in fetal ultrasound using fully convolutional neural networks. International conference on medical image computing and computer-assisted intervention; 2016: Springer.

237. Thomsen K, Iversen L, Titlestad TL, Winther O. Systematic review of machine learning for diagnosis and prognosis in dermatology. Journal of Dermatological Treatment. 2020;31(5):496-510.
238. Barbu A, Lu L, Roth H, Seff A, Summers RM. An analysis of robust cost functions for CNN in computer-aided diagnosis. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 2018;6(3):253-8.
239. Ypsilantis P-P, Siddique M, Sohn H-M, Davies A, Cook G, Goh V, et al. Predicting response to neoadjuvant chemotherapy with PET imaging using convolutional neural networks. PloS one. 2015;10(9):e0137036.
240. Wang C, Yan X, Smith M, Kochhar K, Rubin M, Warren SM, et al., editors. A unified framework for automatic wound segmentation and analysis with deep convolutional neural networks. 2015 37th annual international conference of the ieee engineering in medicine and biology society (EMBC); 2015: IEEE.
241. Alansary A, Kamnitsas K, Davidson A, Khlebnikov R, Rajchl M, Malamateniou C, et al., editors. Fast fully automatic segmentation of the human placenta from motion corrupted MRI. International conference on medical image computing and computer-assisted intervention; 2016: Springer.
242. Chen H-Y, Hsu BW-Y, Yin Y-K, Lin F-H, Yang T-H, Yang R-S, et al. Application of deep learning algorithm to detect and visualize vertebral fractures on plain frontal radiographs. Plos one. 2021;16(1):e0245992.
243. Pranata YD, Wang K-C, Wang J-C, Idram I, Lai J-Y, Liu J-W, et al. Deep learning and SURF for automated classification and detection of calcaneus fractures in CT images. Computer methods and programs in biomedicine. 2019;171:27-37.
244. Hallinan JTPD, Zhu L, Yang K, Makmur A, Algazwi DAR, Thian YL, et al. Deep learning model for automated detection and classification of central canal, lateral recess, and neural foraminal stenosis at lumbar spine MRI. Radiology. 2021;300(1):130-8.
245. Hussain D, Al-Antari MA, Al-Masni MA, Han S-M, Kim T-S. Femur segmentation in DXA imaging using a machine learning decision tree. Journal of X-ray Science and Technology. 2018;26(5):727-46.
246. Šimundić A-M. Measures of diagnostic accuracy: basic definitions. ejifcc. 2009;19(4):203.
247. Rajpurkar P, Irvin J, Bagul A, Ding D, Duan T, Mehta H, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:171206957. 2017.
248. Langerhuizen DW, Janssen SJ, Mallee WH, Van Den Bekerom MP, Ring D, Kerkhoffs GM, et al. What are the applications and limitations of artificial intelligence for fracture detection and classification in orthopaedic trauma imaging? A systematic review. Clinical orthopaedics and related research. 2019;477(11):2482.
249. Burns JE, Yao J, Summers RM. Vertebral body compression fractures and bone density: automated detection and classification on CT images. Radiology. 2017;284(3):788.
250. Carberry GA, Pooler BD, Binkley N, Lauder TB, Bruce RJ, Pickhardt PJ. Unreported vertebral body compression fractures at abdominal multidetector CT. Radiology. 2013;268(1):120-6.

251. Chea P, Mandell JC. Current applications and future directions of deep learning in musculoskeletal radiology. Skeletal radiology. 2020;49(2):183-97.
252. Jones RM, Sharma A, Hotchkiss R, Sperling JW, Hamburger J, Ledig C, et al. Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. NPJ digital medicine. 2020;3(1):1-6.
253. Morgan OJ, Hillstrom HJ, Ellis SJ, Golightly YM, Russell R, Hannan MT, et al. Osteoarthritis in England: incidence trends from National Health Service hospital episode statistics. ACR Open Rheumatology. 2019;1(8):493-8.
254. Üreten K, Arslan T, Gültekin KE, Demir AND, Özer HF, Bilgili Y. Detection of hip osteoarthritis by using plain pelvic radiographs with deep learning methods. Skeletal Radiology. 2020;49(9):1369-74.
255. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Scientific reports. 2018;8(1):1-10.
256. Binvignat M, Pedoia V, Butte AJ, Louati K, Klatzmann D, Berenbaum F, et al. Use of machine learning in osteoarthritis research: a systematic literature review. RMD open. 2022;8(1):e001998.
257. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS medicine. 2018;15(11):e1002699.
258. Štajduhar I, Mamula M, Miletić D, Uenal G. Semi-automated detection of anterior cruciate ligament injury from MRI. Computer methods and programs in biomedicine. 2017;140:151-64.
259. O'Connor SD, Yao J, Summers RM. Lytic metastases in thoracolumbar spine: computer-aided detection at CT—preliminary study. Radiology. 2007;242(3):811-6.
260. Burns JE, Yao J, Wiese TS, Muñoz HE, Jones EC, Summers RM. Automated detection of sclerotic metastases in the thoracolumbar spine at CT. Radiology. 2013;268(1):69.
261. Wiese T, Burns J, Yao J, Summers RM, editors. Computer-aided detection of sclerotic bone metastases in the spine using watershed algorithm and support vector machines. 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2011: IEEE.
262. Wiese T, Yao J, Burns JE, Summers RM, editors. Detection of sclerotic bone metastases in the spine using watershed algorithm and graph cut. Medical Imaging 2012: Computer-Aided Diagnosis; 2012: SPIE.
263. Malinauskaite I, Hofmeister J, Burgermeister S, Neroladaki A, Hamard M, Montet X, et al. Radiomics and machine learning differentiate soft-tissue lipoma and liposarcoma better than musculoskeletal radiologists. Sarcoma. 2020;2020.
264. Sharma GB, Robertson DD, Laney DA, Gambello MJ, Terk M. Machine learning based analytics of micro-MRI trabecular bone microarchitecture and texture in type 1 Gaucher disease. Journal of Biomechanics. 2016;49(9):1961-8.
265. Huber MB, Lancianese SL, Nagarajan MB, Ikpot IZ, Lerner AL, Wismuller A. Prediction of biomechanical properties of trabecular bone in MR images with geometric features and support vector regression. IEEE Transactions on Biomedical Engineering. 2011;58(6):1820-6.

266. Yang C-C, Nagarajan MB, Huber MB, Carballido-Gamio J, Bauer JS, Baum TH, et al. Improving bone strength prediction in human proximal femur specimens through geometrical characterization of trabecular bone microarchitecture and support vector regression. Journal of electronic imaging. 2014;23(1):013013.
267. Sapthagirivasan V, Anburajan M. Diagnosis of osteoporosis by extraction of trabecular features from hip radiographs using support vector machine: an investigation panorama with DXA. Computers in biology and medicine. 2013;43(11):1910-9.
268. Ferizi U, Besser H, Hysi P, Jacobs J, Rajapakse CS, Chen C, et al. Artificial intelligence applied to osteoporosis: a performance comparison of machine learning algorithms in predicting fragility fractures from MRI data. Journal of Magnetic Resonance Imaging. 2019;49(4):1029-38.
269. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology. 2018;287(1):313-22.
270. Tajmir SH, Lee H, Shailam R, Gale HI, Nguyen JC, Westra SJ, et al. Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability. Skeletal radiology. 2019;48(2):275-83.
271. Gao Y, Shen D. Collaborative regression-based anatomical landmark detection. Physics in Medicine & Biology. 2015;60(24):9377.
272. Han D, Gao Y, Wu G, Yap P-T, Shen D, editors. Robust anatomical landmark detection for MR brain image registration. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2014: Springer.
273. Lindner C, Thiagarajah S, Wilkinson JM, Wallis GA, Cootes TF, Consortium a. Fully automatic segmentation of the proximal femur using random forest regression voting. IEEE transactions on medical imaging. 2013;32(8):1462-72.
274. Criminisi A, Shotton J, Robertson D, Konukoglu E, editors. Regression forests for efficient anatomy detection and localization in CT studies. International MICCAI workshop on medical computer vision; 2010: Springer.
275. Chen C, Zheng G, editors. Fully automatic segmentation of ap pelvis x-rays via random forest regression and hierarchical sparse shape composition. International conference on computer analysis of images and patterns; 2013: Springer.
276. Zhang J, Liu M, Shen D. Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. IEEE Transactions on Image Processing. 2017;26(10):4753-64.
277. Yang D, Zhang S, Yan Z, Tan C, Li K, Metaxas D, editors. Automated anatomical landmark detection ondistal femur surface using convolutional neural network. 2015 IEEE 12th international symposium on biomedical imaging (ISBI); 2015: IEEE.
278. Bier B, Unberath M, Zaech J-N, Fotouhi J, Armand M, Osgood G, et al., editors. X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2018: Springer.
279. Li Y, Alansary A, Cerrolaza JJ, Khanal B, Sinclair M, Matthew J, et al., editors. Fast multiple landmark localisation using a patch-based iterative network. International conference on medical image computing and computer-assisted intervention; 2018: Springer.

280. Xu Z, Huo Y, Park J, Landman B, Milkowski A, Grbic S, et al., editors. Less is more: Simultaneous view classification and landmark detection for abdominal ultrasound images. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2018: Springer.
281. Zhang J, Liu M, Wang L, Chen S, Yuan P, Li J, et al. Context-guided fully convolutional networks for joint craniomaxillofacial bone segmentation and landmark digitization. Medical image analysis. 2020;60:101621.
282. Bekkouch IEI, Aidinovich T, Vrtovec T, Kuleev R, Ibragimov B, editors. Multi-agent shape models for hip landmark detection in MR scans. Medical Imaging 2021: Image Processing; 2021: SPIE.
283. von Schacky CE, Sohn JH, Liu F, Ozhinsky E, Jungmann PM, Nardo L, et al. Development and validation of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs. Radiology. 2020;295(1):136.
284. Gielis W, Weinans H, Welsing PM, van Spil W, Agricola R, Cootes T, et al. An automated workflow based on hip shape improves personalized risk prediction for hip osteoarthritis in the CHECK study. Osteoarthritis and Cartilage. 2020;28(1):62-70.
285. Borjali A, Chen AF, Muratoglu OK, Morid MA, Varadarajan KM. Detecting total hip replacement prosthesis design on plain radiographs using deep convolutional neural network. Journal of Orthopaedic Research®. 2020;38(7):1465-71.
286. Krogue JD, Cheng KV, Hwang KM, Toogood P, Meinberg EG, Geiger EJ, et al. Automatic hip fracture identification and functional subclassification with deep learning. Radiology Artificial intelligence. 2020;2(2).
287. Schmaranzer F, Helfenstein R, Zeng G, Lerch TD, Novais EN, Wylie JD, et al. Automatic MRI-based three-dimensional models of hip cartilage provide improved morphologic and biochemical analysis. Clinical orthopaedics and related research. 2019;477(5):1036.
288. Zeng G, Zheng G. Deep learning-based automatic segmentation of the proximal femur from MR images. Intelligent Orthopaedics: Springer; 2018. p. 73-9.
289. . !!! INVALID CITATION !!! {}.
290. McEvoy FJ, Proschowsky HF, Müller AV, Moorman L, Bender-Koch J, Svalastoga EL, et al. Deep transfer learning can be used for the detection of hip joints in pelvis radiographs and the classification of their hip dysplasia status. Veterinary Radiology & Ultrasound. 2021;62(4):387-93.
291. Sewell M, Rosendahl K, Eastwood D. Developmental dysplasia of the hip. Bmj. 2009;339.
292. Woodacre T, Dhadwal A, Ball T, Edwards C, Cox P. The costs of late detection of developmental dysplasia of the hip. Journal of children's orthopaedics. 2014;8(4):325-32.
293. Liu C, Xie H, Zhang S, Mao Z, Sun J, Zhang Y. Misshapen pelvis landmark detection with local-global feature learning for diagnosing developmental dysplasia of the hip. IEEE Transactions on Medical Imaging. 2020;39(12):3944-54.
294. Li Q, Zhong L, Huang H, Liu H, Qin Y, Wang Y, et al. Auxiliary diagnosis of developmental dysplasia of the hip by automated detection of Sharp's angle on standardized anteroposterior pelvic radiographs. Medicine. 2019;98(52).

295. Hu X, Wang L, Yang X, Zhou X, Xue W, Cao Y, et al. Joint landmark and structure learning for automatic evaluation of developmental dysplasia of the hip. IEEE Journal of Biomedical and Health Informatics. 2021;26(1):345-58.
296. Colombi A, Schena D, Castelli CC. Total hip arthroplasty planning. EFORT Open Reviews. 2019;4(11):626-32.
297. Liu W, Wang Y, Jiang T, Chi Y, Zhang L, Hua X-S, editors. Landmarks detection with anatomical constraints for total hip arthroplasty preoperative measurements. International Conference on Medical Image Computing and Computer-Assisted Intervention; 2020: Springer.
298. Yang W, Ye Q, Ming S, Hu X, Jiang Z, Shen Q, et al. Feasibility of automatic measurements of hip joints based on pelvic radiography and a deep learning algorithm. European Journal of Radiology. 2020;132:109303.
299. Pham T-T, Le M-B, Le LH, Andersen J, Lou E. Assessment of hip displacement in children with cerebral palsy using machine learning approach. Medical & Biological Engineering & Computing. 2021;59(9):1877-87.
300. Givon U. Management of the spastic hip in cerebral palsy. Current Opinion in Pediatrics. 2017;29(1):65-9.
301. Webbe J, Sinha I, Gale C. Core outcome sets. Archives of Disease in Childhood-Education and Practice. 2018;103(3):163-6.
302. Pons C, Rémy-Néris O, Médée B, Brochard S. Validity and reliability of radiological methods to assess proximal hip geometry in children with cerebral palsy: a systematic review. Developmental Medicine & Child Neurology. 2013;55(12):1089-102.
303. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic reviews. 2015;4(1):1-9.
304. Hupe M. EndNote X9. Journal of Electronic Resources in Medical Libraries. 2019;16(3-4):117-9.
305. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Systematic reviews. 2016;5(1):1-10.
306. Chougule S, Dabis J, Petrie A, Daly K, Gelfer Y. Is head-shaft angle a valuable continuous risk factor for hip migration in cerebral palsy? J Child Orthop. 2016;10(6):651-6.
307. Analan PD, Yilmaz EE, Adam M, Leblebici B. The effect of physician experience on the measurement reliability of the Reimers' hip migration percentage in children with cerebral palsy. Journal of Physical Therapy Science. 2015;27(10):3255-8.
308. Demir N, Demirel M, Turna Ö, Yildizlar D, Demirbaş Ö, Sağlam Y. Effect of clinician's experience and expertise on the inter-and intra-observer reliability of hip migration index in children with cerebral palsy: A STROBE-compliant retrospective study. Medicine. 2021;100(10).
309. Krebs A, Strobl WM, Grill F. Neurogenic hip dislocation in cerebral palsy: quality of life and results after hip reconstruction. Journal of children's orthopaedics. 2008;2(2):125-31.
310. Chung MK, Zulkarnain A, Lee JB, Cho BC, Chung CY, Lee KM, et al. Functional status and amount of hip displacement independently affect acetabular

dysplasia in cerebral palsy. Developmental Medicine & Child Neurology. 2017;59(7):743-9.
311. Lee KM, Kang JY, Chung CY, Kwon DG, Lee SH, Choi IH, et al. Clinical relevance of valgus deformity of proximal femur in cerebral palsy. Journal of Pediatric Orthopaedics. 2010;30(7):720-5.
312. Hermanson M, Hägglund G, Riad J, Wagner P. Head-shaft angle is a risk factor for hip displacement in children with cerebral palsy. Acta orthopaedica. 2015;86(2):229-32.
313. Terjesen T, Horn J. The Femoral Head-Shaft Angle Is Not a Predictor of Hip Displacement in Children Under 5 Years With Cerebral Palsy: A Population-based Study of Children at GMFCS Levels III-V. J Pediatr Orthop. 2021;41(8):e659-e63.
314. van der List JPJ, Witbreuk MM, Buizer AI, van der Sluijs JA. The prognostic value of the head-shaft angle on hip displacement in children with cerebral palsy. Journal of Children's Orthopaedics. 2015;9(2):129-35.
315. Williamson PR, Altman DG, Bagley H, Barnes KL, Blazeby JM, Brookes ST, et al. The COMET handbook: version 1.0. Trials. 2017;18(3):1-50.
316. Andersen G, Hollung S, Jahnsen R, Elkjær S, Klevberg G. Cerebral palsy registry of Norway and cerebral palsy follow-up program yearly report for 2019. Vestfold Hospial Trust Tønsberg; 2019.
317. Dalkey N, Helmer O. An experimental application of the Delphi method to the use of experts. Management science. 1963;9(3):458-67.
318. CLARKE NP. Bristish Society for Children's Orthopaedic Surgery. Journal of bone and joint surgery British volume. 2008;90:520-3.
319. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. Journal of clinical epidemiology. 2011;64(4):383-94.
320. Lechler P, Frink M, Gulati A, Murray D, Renkawitz T, Bücking B, et al. The influence of hip rotation on femoral offset in plain radiographs. Acta orthopaedica. 2014;85(4):389-95.
321. Foroohar A, McCarthy JJ, Yucha D, Clarke S, Brey J. Head-shaft angle measurement in children with cerebral palsy. Journal of Pediatric Orthopaedics. 2009;29(3):248-50.
322. Gaston MS, Wordie SJ, Wagner P, Hägglund G, Robb JE. The CPUP Hip Score predicts displacement of the hip in children with cerebral palsy. The Bone & Joint Journal. 2022;104(5):640-4.
323. Kay RM, Jaki KA, Skaggs DL. The effect of femoral rotation on the projected femoral neck-shaft angle. Journal of Pediatric Orthopaedics. 2000;20(6):736-9.
324. Boese CK, Dargel J, Oppermann J, Eysel P, Scheyerer MJ, Bredow J, et al. The femoral neck-shaft angle on plain radiographs: a systematic review. Skeletal Radiology. 2016;45(1):19-28.
325. Hägglund G, Lauge-Pedersen H, Persson M. Radiographic threshold values for hip screening in cerebral palsy. Journal of children's orthopaedics. 2007;1(1):43-7.
326. DiFazio R, Shore B, Vessey JA, Miller PE, Snyder BD. Effect of hip reconstructive surgery on health-related quality of life of non-ambulatory children with cerebral palsy. JBJS. 2016;98(14):1190-8.

327. Spencer J, Sait M. The'true'acetabular index in children with cerebral palsy. Annals of the Royal College of Surgeons of England. 2004;86(5):371.
328. Hall DA, Smith H, Heffernan E, Fackrell K, Group COMiTIDRS. Recruiting and retaining participants in e-Delphi surveys for core outcome set development: evaluating the COMiT'ID study. PloS one. 2018;13(7):e0201378.
329. Olveres J, González G, Torres F, Moreno-Tagle JC, Carbajal-Degante E, Valencia-Rodríguez A, et al. What is new in computer vision and artificial intelligence in medical image analysis applications. Quantitative imaging in medicine and surgery. 2021;11(8):3830.
330. Lindner C, Thiagarajah S, Wilkinson J, Wallis G, Cootes T, Consortium a. Development of a fully automatic shape model matching (FASMM) system to derive statistical shape models from radiographs: application to the accurate capture and global representation of proximal femur shape. Osteoarthritis and cartilage. 2013;21(10):1537-44.
331. Lindner C, Wallis G, Cootes T. Increasing shape modelling accuracy by adjusting for subject positioning: an application to the analysis of radiographic proximal femur symmetry using data from the Osteoarthritis Initiative. Bone. 2014;61:64-70.
332. Lindner C, Cootes TF, editors. Fully automatic cephalometric evaluation using random forest regression-voting. IEEE International Symposium on Biomedical Imaging (ISBI) 2015–Grand Challenges in Dental X-ray Image Analysis–Automated Detection and Analysis for Diagnosis in Cephalometric X-ray Image; 2015.
333. Blanco NV, Cootes TF, Lindner C, Carmona IT, Carreira MJ, editors. Fully Automatic Teeth Segmentation in Adult OPG Images. International Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging; 2018: Springer.
334. Lindner C, Cootes T, editors. Automatic Extraction of Hand-Bone Shapes using Random Forest Regression-Voting in the Constrained Local Model Framework. International Symposium on Statistical Shape Models (SHAPE); 2014.
335. Lindner C, Thiagarajah S, Wilkinson JM, Panoutsopoulou K, Day-Williams AG, Consortium a, et al. Investigation of association between hip osteoarthritis susceptibility loci and radiographic proximal femur shape. Arthritis & Rheumatology. 2015;67(8):2076-84.
336. Lindner C, Cootes T, editors. Fully automated radiographic knee shape analysis of the OAI dataset: Is knee shape asymmetry an early indicator of unilateral knee OA? 9th International Workshop on Osteoarthritis Imaging; 2016.
337. Lauder J, Harris J, Layton B, Heire P, Sorani A, DeSancha M, et al. A fully automatic system to assess foot collapse on lateral weight-bearing foot radiographs: A pilot study. Computer Methods and Programs in Biomedicine. 2022;213:106507.
338. Gielis WP, Rayegan H, Arbabi V, Ahmadi Brooghani SY, Lindner C, Cootes TF, et al. Predicting the mechanical hip–knee–ankle angle accurately from standard knee radiographs: a cross-validation experiment in 100 patients. Acta orthopaedica. 2020;91(6):732-7.
339. Davison AK, Cootes TF, Perry DC, Luo W, Lindner C, editors. Perthes disease classification using shape and appearance modelling. International Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging; 2018: Springer.

340. (England) HRA, Social NIf, Care and Health Research (Wales) CSOS, Social Care and, (Wales). HR. Governance arrangements for
research ethics committees:
2018 edition. 2018.
341. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological methods. 1996;1(1):30.
342. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of chiropractic medicine. 2016;15(2):155-63.
343. Lindner C, Thiagarajah S, Wilkinson JM, Wallis GA, Cootes TF, Consortium a, editors. Accurate bone segmentation in 2D radiographs using fully automatic shape model matching based on regression-voting. International conference on medical image computing and computer-assisted intervention; 2013: Springer.
344. Carriero A, Jonkers I, Shefelbine SJ. Mechanobiological prediction of proximal femoral deformities in children with cerebral palsy. Computer methods in biomechanics and biomedical engineering. 2011;14(03):253-62.
345. Cootes TF, Taylor CJ, Cooper DH, Graham J. Active shape models-their training and application. Computer vision and image understanding. 1995;61(1):38-59.
346. Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. IEEE Transactions on pattern analysis and machine intelligence. 2001;23(6):681-5.

Appendices

Appendix 1: Systematic review search strategy for PubMed, SCOPUS and Web of Science.

Appendix 1			
Database	**Number**	**Query**	**Notes**
PubMed	#1	Cerebral pals*[tiab]	
	#2	Cerebral palsy[Mesh]	
	#3	Little's disease[tiab] OR Little Disease[tiab] OR Spastic Diplegi*[tiab]	
	#4	Little's disease[Mesh] OR Little Disease[Mesh] OR Spastic Diplegia[Mesh]	
	#5	#1 OR #2 OR #3 OR #4	Cerebral Palsy
	#6	Paediatric*[tiab] OR pediatric*[tiab] OR neonat*[tiab] OR infant*[tiab] OR child*[tiab] OR adolescen*[tiab] OR teen*[tiab] OR young adult*[tiab]	
	#7	Pediatrics[Mesh] OR Infant[Mesh] OR Child[Mesh] OR Adolescent[Mesh:NoExp] OR Young Adult[Mesh:NoExp]	
	#8	#6 OR #7	Paediatrics
	#9	hip[tiab] OR pelv*[tiab] OR acetab*[tiab] OR cotyloid[tiab] OR coxa*[tiab] OR ischi*[tiab]	
	#10	hip[Mesh] OR pelvic bones[Mesh] OR acetabulum[Mesh] OR acetabula[Mesh] OR acetabulas[Mesh] OR coxa[Mesh] OR coxas[Mesh] OR ischium[Mesh]	
	#11	#9 OR #10	Hip
	#12	radiograph*[tiab] OR radiol*[tiab] OR x-ray*[tiab] OR x ray*[tiab] OR xray*[tiab] OR X-Radiation*[tiab] OR XRadiation*[tiab] OR X Radiation*[tiab] OR roentgen ray*[tiab]	
	#13	radiography[Mesh] OR radiology[Mesh] OR x-ray[Mesh] OR x ray[Mesh] OR xray[Mesh] OR X-Radiation[Mesh] OR X Radiation[Mesh] OR roentgen ray[Mesh]	
	#14	#12 OR #13	X-ray

Appendix 1 (continued)			
Database	**Number**	**Query**	**Notes**
PubMed	#15	#5 AND #8 AND #11 AND #14	Cerebral Palsy AND Paediatrics AND Hip AND X-ray
SCOPUS	#1	TITLE-ABS-KEY("Cerebral pals*")	
	#2	TITLE-ABS-KEY("Little* disease" OR "Spastic Diplegi*")	
	#3	#1 OR #2	Cerebral Palsy
	#4	TITLE-ABS-KEY(Paediatric* OR pediatric* OR neonat* OR infant* OR child* OR adolescen* OR teen* OR "young adult*")	
	#5	TITLE-ABS-KEY(hip OR pelvi* OR acetabul* OR cotyloid OR coxa* OR ischi*)	
	#6	TITLE-ABS-KEY(radiograph* OR radiol* OR "x-ray*" OR "x ray*" OR xray* OR "X-Radiation*" OR XRadiation* OR "X Radiation*" OR "roentgen ray*")	
	#7	#3 AND #4 AND #5 AND #6	Cerebral Palsy AND Paediatrics AND Hip AND X-ray
Web of Science	#1	TS=("Cerebral pals*")	
	#2	TS=("Little* disease" OR "Spastic Diplegi*")	
	#3	#1 OR #2	Cerebral Palsy
	#4	TS=(Paediatric* OR pediatric* OR neonat* OR infant* OR child* OR adolescen* OR teen* OR "young adult*")	
	#5	TS=(hip OR pelvi* OR acetabul* OR cotyloid OR coxa* OR ischi*)	
	#6	TS=(radiograph* OR radiol* OR "x-ray*" OR "x ray*" OR xray* OR "X-Radiation*" OR XRadiation* OR "X Radiation*" OR "roentgen ray*")	
	#7	#3 AND #4 AND #5 AND #6	Cerebral Palsy AND Paediatrics AND Hip AND X-ray

Appendix 2: Studies that were not accessible

Appendix 2			
Title	**Author**	**Year**	**Journal**
Early Bony Hip Reconstructive Surgery for Hip Subluxation in Children With Severe Cerebral Palsy	Bean, B. K. and Baird, G. O. and Caskey, P. M. and Bronson, W. B. and McMulkin, M. L. and Tompkins, B. J.	2020	Orthopedics
Evaluation of adductor myotomy versus adductor transfer to ischiadic tuber in the treatment of spastic hip in cerebral palsy	Andrzej Borowski, Ewa Pogonowicz, Rafał Plebański, Marek Synder, Andrzej Grzegorzewski	2011	Ortopedia Traumatologia Rehabilitacja
Two-stage surgery in the treatment of spastic hip dislocation--comparison between early and late results of open reduction and derotation-varus femoral osteotomy combined with Dega pelvic osteotomy preceded by soft tissue release	Marek Jóźwiak and Aleksander Koch	2011	Ortopedia, Traumatologia Rehabilitacja
Soft tissue, varus derotation femoral and pelvic surgery in cerebral palsy children: A mid-term outcome study	Panou, A. and Testa, G. and Peccati, A. and Tsibidakis, H. and Portinaro, N. M.		Minerva Ortopedicae Traumatologica
The effect of obturator nerve block on hip lateralization in low functioning children with spastic cerebral palsy	Park, E. S. and Rha, D. W. and Lee, W. C. and Sim, E. G.	2014	Yonsei Med J
Outcome of Femoral Varus Derotational Osteotomy for the Spastic Hip Displacement: Implication for the Indication of Concomitant Pelvic Osteotomy	Park, H. and Abdel-Baki, S. W. and Park, K. B. and Park, B. K. and Rhee, I. and Hong, S. P. and Kim, H. W.	2020	Journal of Clinical Medicine

Appendix 3: Studies included in the systematic review

Appendix 3				
Study ID	**Title**	**Author (Year)**	**Journal**	**Location**
1	Hip pain in adolescents with cerebral palsy: a population-based longitudinal study	Larsen (2021)	Dev Med Child Neurol	Norway
2	Remodelling of femoral head deformity after hip reconstructive surgery in patients with cerebral palsy	Min (2021)	Bone Joint J	South Korea
3	Failure of Hip Reconstruction in Children With Cerebral Palsy: What Are the Risk Factors?	Minaie (2021)	J Pediatr Orthop	USA
4	Combined pelvic and femoral reconstruction in children with cerebral palsy	Alassaf (2018)	J Int Med Res	Canada
5	Recurrence of hip instability after reconstructive surgery in patients with cerebral palsy	Bayusentono (2014)	J Bone Joint Surg Am	South Korea
6	Hip-joint congruity after Dega osteotomy in patients with cerebral palsy: long-term results	Braatz (2016)	Int Orthop	Germany
7	Hip displacement in children with cerebra palsy in Scotland: a total population study	Bugler (2018)	Journal of Childrens Orthopaedics	Scotland (UK)
8	Results and complications of percutaneous pelvic osteotomy and intertrochanteric varus shortening osteotomy in 54 consecutively operated GMFCS level IV and V cerebral palsy patients	Canavese (2017)	Eur J Orthop Surg Traumatol	France and Switzerland
9	Hip Development After Selective Dorsal Rhizotomy in Patients with Cerebral Palsy	Chan (2013)	Journal of Orthopaedics, Trauma and Rehabilitation	China
10	Outcomes of Isolated Varus Derotational Osteotomy in Children With Cerebral Palsy Hip Dysplasia and Predictors of Resubluxation	Chang (2018)	J Pediatr Orthop	USA
11	Prevalence and predictive factors of hip displacement in children with cerebral palsy at paediatric institute, Kuala Lumpur hospital	Ching (2017)	Neurology Asia	Malaysia
12	Determinants of Hip and Femoral Deformities in Children With Spastic Cerebral Palsy	Cho (2018)	Ann Rehabil Med	South Korea

Appendix 3 (continued)				
Study ID	**Title**	**Author (Year)**	**Journal**	**Location**
13	Is head shaft angle a valuable continuous risk factor for hip migration in cerebral palsy?	Chougule (2016)	Journal of Children's Orthopaedics	England (UK)
14	The Impact of Spinal Fusion on Hip Displacement in Cerebral Palsy	Cobanoglu (2020)	Indian J Orthop	USA
15	Soft tissue surgery as an initial treatment for hip displacement in spastic cerebral palsy	Silva (2020)	Sicot-J	Brazil
16	Prevalence of hip dislocation among children with cerebral palsy in regions with and without a surveillance programme: a cross sectional study in Sweden and Norway	Elkamil (2011)	BMC Musculoskelet Disord	Sweden and Norway
17	Soft tissue release of the spastic hip by psoas-rectus transfer and adductor tenotomy for long-term functional improvement and prevention of hip dislocation	Heimkes (2011)	Journal of Pediatric Orthopaedics Part B	Germany
18	Head-shaft angle is a risk factor for hip displacement in children with cerebral palsy	Hermanson (2015a)	Acta Orthop	Sweden
19	Surgical management of hip subluxation and dislocation in children with cerebral palsy: isolated VDRO or combined surgery?	Huh (2011)	J Pediatr Orthop	USA
20	Five-year outcome of state-wide hip surveillance of children and adolescents with cerebral palsy	Kentish (2011)	Journal of Pediatric Rehabilitation Medicine	Austrailia
21	Hip displacement in relation to age and gross motor function in children with cerebral palsy	Larnert (2014)	Journal of Children's Orthopaedics	Sweden
22	The prognostic value of the head-shaft angle on hip displacement in children with cerebral palsy	List (2015a)	Journal of Children's Orthopaedics	Netherlands
23	Parameters of radiographic coxometry in reconstructive operations on the hip joint as part of multilevel surgical interventions in children with cerebral palsy	Tomov (2019)	Genij Ortopedii	Russia
24	Use of iliac crest allograft for Dega pelvic osteotomy in patients with cerebral palsy	Sung (2018a)	BMC Musculoskelet Disord	South Korea

Appendix 3 (continued)				
Study ID	**Title**	**Author (Year)**	**Journal**	**Location**
25	Hip pain is more frequent in severe hip displacement: a population-based study of 77 children with cerebral palsy	Ramstad (2016)	J Pediatr Orthop B	Norway
26	Hip displacement and dislocation in a total population of children with cerebral palsy in Scotland	Wordie (2020)	Bone Joint J	Scotland (UK)
27	The Femoral Head-Shaft Angle Is Not a Predictor of Hip Displacement in Children Under 5 Years With Cerebral Palsy: A Population-based Study of Children at GMFCS Levels III-V	Terjesen (2021)	J Pediatr Orthop	Norway
28	Fate of stable hips after prophylactic femoral varization osteotomy in patients with cerebral palsy	Sung (2018b)	BMC Musculoskelet Disord	South Korea
29	Avascular necrosis as a complication of the treatment of dislocation of the hip in children with cerebral palsy	Koch (2015)	Bone Joint J	Poland
30	The natural history of hip development in cerebral palsy	Terjesen (2012)	Dev Med Child Neurol	Norway
31	Hip displacement in children with cerebral palsy	Wordie (2021)	Bone Joint J	Scotland (UK)
32	Acetabular and femoral remodeling after varus derotational osteotomy in cerebral palsy: The effect of age and Gross Motor Function Classification Level	Shore (2016)	Journal of Pediatric Orthopaedics Part B	USA
33	Severe hip displacement reduces health-related quality of life in children with cerebral palsy	Ramstad (2017)	Acta Orthop	Norway
34	Patterns of hip migration in non-ambulant children with cerebral palsy: A prospective cohort study	Poirot (2020)	Annals of Physical and Rehabilitation Medicine	France
35	Outcome of Femoral Varus Derotational Osteotomy for the Spastic Hip Displacement: Implication for the Indication of Concomitant Pelvic Osteotomy	Park (2020)	Journal of Clinical Medicine	South Korea
36	Incidence and risk factors of hip joint pain in children with severe cerebral palsy	Jozwiak (2011)	Disability and Rehabilitation	Poland

Appendix 3 (continued)				
Study ID	**Title**	**Author (Year)**	**Journal**	**Location**
37	Prediction of hip displacement in children with cerebral palsy: development of the CPUP hip score	Hermanson (2015b)	Bone Joint J	Sweden
38	Pelvic obliquity and measurement of hip displacement in children with cerebral palsy	Hägglund (2018)	Acta Orthop	Sweden
39	Association between pelvic obliquity and scoliosis, hip displacement and asymmetric hip abduction in children with cerebral palsy: a cross-sectional registry study	Hägglund (2020)	BMC Musculoskelet Disord	Sweden
40	Utility of combined hip abduction angle for hip surveillance in children with cerebral palsy	Divecha (2011)	Indian Journal of Orthopaedics	India
41	Proximal femoral geometry before and after varus rotational osteotomy in children with cerebral palsy and neuromuscular hip dysplasia	Davids (2013)	J Pediatr Orthop	USA
42	Reliability of radiologic measures of hip displacement in a cohort of preschool-aged children with cerebral palsy	Craven (2014)	J Pediatr Orthop	Austrailia
43	Acetabular Remodeling After a Varus Derotational Osteotomy in Children With Cerebral Palsy	Chang (2016)	J Pediatr Orthop	USA
44	Proximal femoral osteotomy in children with cerebral palsy: the perspective of the trainee	Zhou (2017)	Journal of Childrens Orthopaedics	Austrailia
45	Clinical and radiographic results of multilevel surgical interventions for hip subluxation and dislocation in children with cerebral palsy	Tomov (2018)	Genij Ortopedii	Russia
46	The head–shaft angle of the hip in early childhood	List (2015b)	Bone Joint J	Netherlands
47	Assessment of hip displacement in children with cerebral palsy using machine learning approach	Pham (2021)	Med Biol Eng Comput	Canada

Appendix 4: Study details and patient characteristics

Appendix 4								
Study ID	**Study type**	**Patients (controls)**	**No. of hips**	**Age**	**Sex (% female)**	**Duration of follow up**	**Timepoint of assessment**	**Primary intervention**
1	prospective	67	128	Mean:14y7m SD:1y5m	42%	Not specified	latest radiograph of the pelvis and hip joints taken for CPOP. If the latest radiograph was taken before 2017, the respondent was asked to permit a new radiograph to be taken.	hip surveillance programme
2	retrospective	108	214	Mean:9.4y SD:3.2y	30%	5.2 years	Pre- + postoperative	HRS (hip reconstructive surgery) including FVDO
3	retrospective	179	291	Mean:7.8y SD:3.3y	43%	3.9±2.1 years	Preoperative	HRS (hip reconstructive surgery) including femoral osteotomy, both a femoral osteotomy and a concurrent acetabular osteotomy and isolated acetabular osteotomy
4	retrospective	71	85	Mean:8.4y SD:3.2y	52%	6.6 ± 3.1 years	Pre- + postoperative (including follow up)	VDRO combined with modified Dega osteotomy
5	retrospective	76	144	Mean:8.5y SD:2.3y	25%	4.9 years (SD, 2.4 years; range, 1.0 to 9.8 years)	Pre- + postoperative (including follow up - at least 2 years)	reconstructive surgery (femoral varus derotational osteotomy alone or combined with a modified Dega pelvic osteotomy)

Appendix 4 (continued)								
Study ID	**Study type**	**Patients (controls)**	**No. of hips**	**Age**	**Sex (% female)**	**Duration of follow up**	**Timepoint of assessment**	**Primary intervention**
6	retrospective	72	72	Median:7.6y Range:4.7–16.3y	38%	7.7 years (4.9 to 11.8)	Pre- + postoperative (including follow up)	single-event multilevel surgery (SEMLS) in combination with hip reconstruction by using a periacetabular osteotomy as described by Dega concerning post-operative remodeling and plasticity of the femoral head post-operatively
7	cross-sectional	1171	Not specified	Mean:7.88y	Not specified	N/A	first hip radiograph as part of the CPIPS programme	hip surveillance programme
8	retrospective	54	64	Mean:9.1y SD:3.3y Range:4.0–16.5y	37%	43.9 ± 19.5 months (range 3–72).	Pre- + postoperative (including follow up)	simultaneous soft tissue release, VDRSO, and PPO
9	retrospective	53	Not specified	Mean:7.9y SD:2.2y	43%	5.3 years with minimal follow up of 12 months.	Pre- + postoperative (including follow up)	bilateral selective dorsal rhizotomy
10	retrospective	91	179	Mean:4.6y SD:1.6y Range:2.4-10.6y	36%	5.4 (1.03-10.20)	Pre- + postoperative (most recent preoperative radiograph and all available postoperative anteroposterior pelvis radiographs.)	isolated **femoral** varus derotational osteotomy (VDRO)

Appendix 4 (continued)								
Study ID	**Study type**	**Patients (controls)**	**No. of hips**	**Age**	**Sex (% female)**	**Duration of follow up**	**Timepoint of assessment**	**Primary intervention**
11	cross-sectional	75	Not specified	Median:7.7y IQR:6.5y	44%	N/A	Not specified	hip surveillance programme
12	retrospective	57	Not specified	Mean:9.3y SD:1.8y Range:7–14y	46%	68.4 months (SD=22.0; range, 60–124 months). Duration between P/Ex and imaging study (mo)	Preoperative (between baseline physical exam and pre-operative evaluation for ortho paedic surgery)	Nil
13	retrospective	100 (103)	Not specified	Mean:8.8y SD:4.3y Range:3–18y	Not specified	7.5 (range 5–10) years	taken in A&E during a 6-month period	hip surveillance programme
14	retrospective	50 Group 1: 19 Group 2: 23 Group 3: 8	100	Group 1 Mean:12y SD:2y Range:8–15y Group 2 Mean:12y SD:1.7y Range:8–15y Group 3 Mean:12y SD:2y Range:8–12y Overall range:7.5–15.0y	56%	Group 1: 54 ± 30 (6–129) months of follow-up Group 2: 45 ± 32 (range 4–98) months of follow-up Group 3: 50 ± 39 (range 3–129) months of follow-up	pre- + post- spinal fusion surgery	spinal fusion with pelvic fixation

Appendix 4 (continued)								
Study ID	**Study type**	**Patients (controls)**	**No. of hips**	**Age**	**Sex (% female)**	**Duration of follow up**	**Timepoint of assessment**	**Primary intervention**
15	cross-sectional	93	Not specified	Mean:6.8y SD:3y Range:2.4–12y	43%	10.3 years (SD = 6) years	Pre- + postoperative	soft tissue surgery as the first treatment for hip displacement
16	cross-sectional	255 (119 Norwegian + 136 Swedish)	Not specified	Sweden Mean:5.7y SD:2.3y Norway Mean:7.6y SD:2.9y	45.6% in Sweden + 38.7% in Norway	N/A	most recent or preoperative	hip surveillance programme
17	retrospective	71	140	Mean:7y Range:3-12y	41%	12.8 years (1.0/27.0)	The measurement was taken twice; the first one was taken as close in time as possible before surgery (up to 3 months before) and the second radiograph on the day of the last reassessment.	Soft tissue release of the spastic hip by psoas–rectus transfer and adductor tenotomy
18	retrospective	145	Not specified	Mean:3.5y Range:0.6-9.7y	50%	Not specified	first radiograph in CPUP + a follow-up period of 5 years or until development of MP > 40% of either hip within 5 years	Hip surveillance programme

Appendix 4 (continued)								
Study ID	**Study type**	**Patients (controls)**	**No. of hips**	**Age**	**Sex (% female)**	**Duration of follow up**	**Timepoint of assessment**	**Primary intervention**
19	retrospective	75	116	Mean:7.0y Range:2.1-12.1y	40%	4.6 years (range, 2.0 to 10.7 y)	Pre- + postoperative (including follow up)	isolated varus derotational osteotomy (VDRO) + VDRO combined with open hip reduction and/or pelvic osteotomy
20	prospective	1115	Not specified	Not specified	42%	1.2 years (range 1 month –5 +8yrs)	During hip surveillance program (multiple timepoints - not specified)	Hip surveillance programme
21	retrospective	353	Not specified	Not specified	Not specified	Not specified ?	Before 3 years of age + follow up between 2-7 years	Hip surveillance programme
22	retrospective	50	100	Not specified	30%	Not specified ?	age of two years (12–32 months; T1), age of four years (36–60 months; T2) and age of seven years (72–96 months; T3)	Nil
23	prospective	124	Not specified	Mean:7.01y SD:2.47y Range:3–13y	Not specified	At least 30 months	Pre- + postoperative (including follow up - annually)	reconstructive surgery for hip dislocation with also simultaneous surgical interventions for: contractures of the knee joints and/or contractures of the ankle joints and foot deformities

Appendix 4 (continued)								
Study ID	**Study type**	**Patients (controls)**	**No. of hips**	**Age**	**Sex (% female)**	**Duration of follow up**	**Timepoint of assessment**	**Primary intervention**
24	retrospective	110	150	Mean:8.7y SD:2.4y Range:2.8-13.8y	38%	2.9 ± 2.6 (1.0 to 12.0)	Pre- + postoperative (including follow up)	Dega pelvic osteotomy using iliac crest allograft
25	prospective	77	154	Mean:9.5y SD:1.6y Range:7–12y	38%	Not specified	radiograph taken nearest to the time that the questionnaire (pain assessment) was answered	Nil
26	retrospective	1,171	Not specified	Mean:7.9y Range:2-16y	Not specified	N/A	Pre- + post-CPIPS	hip surveillance programme
27	retrospective	101	Not specified	Mean:2.4y Range:0.8-4.9y	40%	4.3 years (range, 0.9 to 11.8 y)	At diagnosis + at the last follow-up or last preoperative radiograph	hip surveillance programme
28	retrospective	119	224	Mean:8.9y SD:2.7y Range:2.8 to 16.5y	34%	3.3 ± 2.7 (1 to 11.9)	Pre- + postoperative (including follow up - at least two follow-up evaluations)	hip reconstructive surgery including FVO (femoral varization osteotomy)
29	retrospective	81	115	Mean:9y Range:3.5-13.8y	49%	5.5 years (1.6 to 15.1)	Pre- + postoperative (including follow up - one year post-operatively and at final review)	open reduction of the hip

Appendix 4 (continued)								
Study ID	**Study type**	**Patients (controls)**	**No. of hips**	**Age**	**Sex (% female)**	**Duration of follow up**	**Timepoint of assessment**	**Primary intervention**
30	prospective	335	Not specified	Mean:3y Range:6m–7y 11m	44%	2 years 9 months (range 6mo–7y 3mo)	CPOP: Initial radiograph (shortly after diagnosis, preferably at the age of 1 year, in children with pronounced spasticity OR for all other children, a radiograph at the age of 2 years) + last follow up radiograph (until operative treatment for hip displacement or until the most recent radiograph in those who had not undergone hip surgery)	hip surveillance programme
31	retrospective	239	346	Mean:11.6y Range:3-18y	38%	6.5 (2 to 14.8)	birth through to the date of analysis	hip surveillance programme
32	retrospective	55	102	Median:6.5y Range:3.2–15.6y	54%	7.4 years (range 3–11 years)	Pre- + postoperative (including follow up - within the first year (postoperative), and then at ~ 1-year follow-up intervals for a minimum of 2 years and a maximum of 11 years)	isolated varus derotational osteotomy (VDRO)
33	prospective	67	Not specified	Mean:9y Range:7–12y	40%	Not specified	radiograph taken nearest to the time the questionnaire was answered. The mean length of time between radiograph and questionnaire was 5.4 (0–25) months, and no surgery was performed during this interval.	hip surveillance programme
34	prospective	235	Not specified	Mean:6y4m Range:2y 4m-10y11m)	55%	median follow-up of 2.7 years (range 0.4–6.3; mean 2.6)	baseline and at each annual visit	Nil

Appendix 4 (continued)								
Study ID	**Study type**	**Patients (controls)**	**No. of hips**	**Age**	**Sex (% female)**	**Duration of follow up**	**Timepoint of assessment**	**Primary intervention**
35	retrospective	72	144	Mean:6.2y Range:3.2-12.2y	33%	7.0 (2.0 to 16.0) years	Preoperative, postoperative, and final follow-up radiographs	bilateral VDROs without concomitant pelvic osteotomy
36	cross-sectional	73	99	Mean10.8y Range:4.0–18.0y	42%	N/A	Not specified	physiotherapy (abduction treatment and horse-back riding therapy)
37	prospective	145	Not specified	Mean:3.5y Range:0.6-9.7y	50%	Not specified (followed up until hip displacement (MP > 40%) occurred (group 1) or for five years without hip displacement (group 2))	CPUP: first radiographic examination (MP & HSA) + MP was then measured prospectively once a year according to the CPUP schedule until hip displacement (MP > 40%) occurred (group 1) or for five years without hip displacement (group 2).	hip surveillance programme
38	cross-sectional	268	Not specified	Not specified (children < 18 years)	44%	N/A	CPUP: First pelvic radiograph in CPUP during study period	hip surveillance programme
39	cross-sectional	337	Not specified	Not specified	45%	N/A	CPUP: First pelvic radiograph in CPUP during study period	hip surveillance programme
40	cross-sectional	103	206	Mean:5.03y Range:2–11y	53%	N/A	Not specified	hip surveillance programme

Appendix 4 (continued)								
Study ID	**Study type**	**Patients (controls)**	**No. of hips**	**Age**	**Sex (% female)**	**Duration of follow up**	**Timepoint of assessment**	**Primary intervention**
41	retrospective	75	137	Mean:7y SD:2y8m Range:3y2m-17y5m	44%	5 years and 6 months (range, 1 to 12 y and 7 mo)	preoperative, postoperative, and follow-up	Varus Rotational Osteotomy
42	cross-sectional	133	Not specified	Median:35.6m Range:30.5-36.4m	36%	Not specified (18, 24, 30, 36, 48, and 60 months)	During hip surveillance (at 18, 24, 30, 36, and 48 months)	hip surveillance programme
43	retrospective	87 (917)	174	Mean:4.6y SD:1.6y Range:2.4-10.6 y	Not specified	5.1 ± 2.2 years (range, 1.1 to 9.9 y)	NSA (intraoperatively), ADR (preoperative and postoperative - but may not be relevant)	isolated varus derotation osteotomy
44	prospective	90	180	Mean:7y11m Range:4y3m-13y 9m	47%	mean 28 months; 21 to 40	Pre- + postoperative (including follow up - three, six and 12 weeks and at six and 12 months following surgery and yearly thereafter)	Proximal femoral osteotomy
45	Retrospective	50	Not specified	Mean:6.2y SD:1.37y	Not specified	2.8 ± 1.7 years (mean)	Pre- + postoperative	Surgery
46	Retrospective	50	100	Mean: T1:24m T2:49m T3:84m	30%	Not specified	Different age intervals	None
47	retrospective	122	Not specified	Not specified Range:4–10y	Not specified	Not specified	Not specified	None

Appendix 5: Measurements reported per article with definitions

Appendix 5			
Study ID	**Measurement**	**Verbatim written definition**	**Visual aid**
1	RMP[1]	Not specified	No
2	RMP	MP was calculated by dividing the width of the femoral head lateral to the Perkin's line by the total width of the femoral head	Yes
	NSA[2]	NSA was defined as the angle between a line passing through the centre of the femoral shaft and another line connecting the femoral head centre and the midpoint of the femoral neck	No
	MHR[3]	Concentric circles were drawn at the centre of the femoral head with the larger circle outlining the outer cortex of the femoral head and the inner circle outlining the innermost cortex of the head. The ratio between the radii of the two circles was calculated as the sphericity of the femoral head	Yes
3	RMP	Migration percentage (MP) measured as MP = (A/B) ×100.]	Yes
	AcI[4]	AI measured as the angle formed by Hilgenreiner's line and a line draw from the lateral triradiate cartilage to the lateral acetabular margin	Yes
	NSA	NSA measured as the angle formed by a line bisecting the femoral head (crossing the epiphysis at 90 degrees) and a line formed along the axis of the femoral shaft.	Yes
4	RMP	Not specified	No
	CEA[5]	Not specified	No
	AA[6]	Not specified	No
5	NSA	On the left hip, the neck-shaft angle (NSA) was defined as the angle between a line passing through the center of the femoral shaft and another line connecting the femoral head center and the midpoint of the femoral neck. The femoral head center was the center of the largest best-fitting circle inside the femoral head.	Yes
	HSA[7]	Not specified	No
	RMP	On the right hip, the migration percentage (MP) was calculated by dividing the width of the femoral head lateral to Perkin's line (A) by the total width of the femoral head (B).]	Yes
6	RMP	Not specified	No
	CEA	Not specified	No

Appendix 5 (continued)			
Study ID	**Measurement**	**Verbatim written definition**	**Visual aid**
7	RMP	Not specified	No
8	RMP	Not specified	No
	AA	Not specified	No
9	CEA	Not specified	Yes
	RMP	Not specified	Yes
10	Shenton's line	Not specified	No
	RMP	Not specified	No
11	RMP	MP is obtained by identifying Hilgenreiner's line (H) and Perkin's line (P) and then measuring the proportion (%) of capital epiphysis that has migrated beyond Perkin's line laterally (A/B x 100).]	Yes
	AcI	AI as the angle between the slope of the acetabulum and Hilgenreiner's line	Yes
12	RMP	Measurement of migration percentage (MP). MP=B/A×100. Hilgenreiner's line and Perkins line are marked as 'H' and 'P'. MP is the proportion (%) of the capital epiphysis that appears to lie outside the acetabulum.	Yes
	NSA	Measurement of the femoral neck and shaft angle (FNS). 'a' is the FNS measurement, performed in standard anterior-posterior X-rays of the proximal femur or pelvis, which was generated by the intersection angle between the femoral neck axis and femoral shaft axis.	Yes
13	RMP	Not specified	Yes
	HSA	Not specified	Yes
14	RMP	migration index (x/y × 100)	Yes
	AA	acetabular angle (narrow angle between yellow lines)	Yes
15	RMP	Not specified	Yes
	AcI	Not specified	Yes
	HSA	Not specified	Yes
16	RMP	Not specified	No

Appendix 5 (continued)			
Study ID	**Measurement**	**Verbatim written definition**	**Visual aid**
17	RMP	Not specified	No
	AcI	Not specified	No
	NSA	Not specified	No
18	RMP	migration percentage (MP), calculated as b/c × 100, on the left hip	Yes
	HSA	HSA: The HSA is measured by drawing a line midway through the femoral shaft and then drawing another line perpendicular to the proximal femoral physis through the center of the proximal femoral epiphysis	Yes
19	RMP	Not specified	No
	AcI	Not specified	No
	AA	Not specified	No
	NSA	Not specified	No
	CEA	Not specified	No
20	RMP	Not specified	No
21	RMP	Measurement of Migration Percentage (MP). MP = A/B 9 100. On the right hip with a "Gothic arch" formation of the lateral margin, the midpoint of the arch is used as reference point	Yes
22	RMP	the migration percentage (MP) is measured by a Hilgenreiner's line (H) and three perpendicular lines. The MP is measured by A/B x 100 %	Yes
	HSA	head-shaft angle (C) by measuring the medial angle between a line perpendicular to the proximal femoral epiphysis and a line through the middle of the femoral shaft.	Yes
23	RMP	Not specified	No
	AcI	Not specified	No
	ADR[8]	Not specified	No
	NSA	Not specified	No
	CEA	Not specified	No

Appendix 5 (continued)			
Study ID	**Measurement**	**Verbatim written definition**	**Visual aid**
24	NSA	neck-shaft angle (NSA) was defined as the angle between a line passing through the center of the femoral shaft and another line connecting the center of the femoral head and the midpoint of the femoral neck. The center of the femoral head was the center of the largest best-fitting circle inside the femoral head	Yes
	RMP	Migration percentage (MP) was calculated by dividing the width of the femoral head lateral to Perkin's line (a) by the total width of the femoral head (b)	Yes
	AcI	Acetabular index (AI) was defined as the angle between the acetabular roof and the Hilgenreiner's line	Yes
25	RMP	MP is the percentage of the femoral head lateral to the acetabulum (lateral to Perkins' line), measured parallel to Hilgenreiner's line.	No
26	RMP	Not specified	No
27	RMP	MP is the percentage of the femoral head lateral to the acetabulum (lateral to Perkins' line). + Measurement of the migration percentage (MP) is shown in the left hip (MP = a/b×100)	Yes
	HSA	The head-shaft angle (HSA) is the medial angle between a line perpendicular to the proximal femoral physis and a line through the middle of the femoral shaft	Yes
28	NSA	NSA was defined as the angle between a line passing through the center of the femoral shaft and another line connecting the femoral head center and the midpoint of the femoral neck. The femoral head center was the center of the largest best-fitting circle inside the femoral head.	Yes
	HSA	HSA was defined as the angle between a line passing through the center of the femoral shaft and another line perpendicular to the proximal femoral physis passing through the center of the proximal femoral epiphysis	Yes
	RMP	RMP was calculated by dividing the width of the femoral head lateral to Perkin's line (A) by the total width of the femoral head	Yes
29	RMP	Not specified	No
	AcI	Not specified	No
	ESA[9]	Not specified	No
	PFA[10]	Not specified	No

Appendix 5 (continued)			
Study ID	**Measurement**	**Verbatim written definition**	**Visual aid**
30	RMP	RMP the percentage of the femoral head lateral to the acetabulum (lateral to Perkins' line), measured parallel to Hilgenreiner's line.	Yes
	AcI	AI as the slope of the acetabular roof, which is the angle between the acetabular roof and Hilgenreiner's line	Yes
31	RMP	Not specified	No
32	NSA	Not specified	No
	AcI	Not specified	No
	RMP	Not specified	No
	CEA	Not specified	No
33	RMP	RMP as the percentage of the femoral head lateral to the acetabulum (lateral to Perkins' line), measured parallel to Hilgenreiner's line.	No
34	RMP	Not specified	Yes
35	RMP	Not specified	No
	AcI	Not specified	No
	NSA	Not specified	No
	HSA	Not specified	No
36	RMP	Not specified	No
37	RMP	Hilgenreiner's and Perkins lines; b: is the horizontal distance that the femoral head has translated lateral to Perkins line, c: is the horizontal measurement of the femoral head medial to Perkins line. The MP is b/c × 100	Yes
	HSA	HSA was measured by the angle intersecting two lines; one passing through the proximal mid diaphyseal line of the femoral shaft and a second perpendicular to the proximal femoral physis	Yes
38	RMP	Not specified	Yes
	PAMP[11]	Not specified	Yes
39	RMP	Not specified	No
40	RMP	Not specified	No

Appendix 5 (continued)			
Study ID	**Measurement**	**Verbatim written definition**	**Visual aid**
41	NSA	The femoral shaft and neck axes were drawn on the radiographic film, and the NSA was measured as the angle subtended by the intersection of these 2 axes. To define the shaft axis 2 lines were drawn perpendicular to the diaphysis of the femur. A line connecting the midpoints of these 2 lines was then drawn to represent the shaft axis. The neck axis was defined in a similar manner utilising 2 lines drawn through the proximal and distal margins of the middle third of the femoral neck. The head axis was determined by drawing a line perpendicular to the proximal femoral physis. The HSA was calculated from the relation between the head and shaft axes	Yes
	HSA		Yes
	MeI[12]	The MeI was determined by dividing the distance between the proximal physis and the lateral margin of the greater trochanter on the neck axis line (described above for NSA) by the distance between the proximal physis and the intersection with the shaft axis line (described above for NSA) on the neck axis line. The greater the MeI, the greater the medialisation of the femoral shaft.	Yes
42	RMP	Not specified	No
	HEA[13]	HEA as the acute angle between a line drawn parallel to and through the proximal femoral epiphysis and Hilgenreiner line (HL).	Yes
	AcI	Not specified	No
	NSA	Not specified	No
43	NSA	Not specified	No
	ADR	The acetabular depth ratio (ADR), where A is the depth and B is the width. The width B is measured as the distance from the inferior teardrop to the lateral edge of the sourcil and the depth A is the perpendicular distance from the midpoint of B to the deepest point of the acetabular roof. ADR = A/BÂ100. This hip has an ADR of 24	Yes
44	RMP	Not specified	No
	NSA	Not specified	No
45	RMP	Not specified	No
	AcI	Not specified	No
	ADR	Not specified	No
	NSA	Not specified	No
	CEA	Not specified	No

Appendix 5 (continued)			
Study ID	**Measurement**	**Verbatim written definition**	**Visual aid**
46	HSA	The head–shaft angle (HSA) is the medial angle between a line perpendicular to the proximal femoral epiphysis and a line through the middle of the femoral shaft.	Yes
47	RMP	Migration percentage MP is defined as the ratio of the femoral head migrated beyond the acetabular edge (a) to the total width of the femoral head (b).	Yes

[1]RMP: Reimers' migration percentage

[2]NSA: Femoral neck-shaft angle

[3]MHR: Mose hip ratio

[4]AcI: Acetabular index

[5]CEA: Centre edge angle of Wiberg

[6]AA: Acetabular angle or Sharp's angle

[7]HSA: Femoral head-shaft angle

[8]ADR: Acetabular Depth Ratio

[9]ESA: Epiphyseal shaft angle

[10]PFA: Pelvic femoral angle

[11]PAMP: Pelvic adjusted migration percentage

[12]MeI: Medialization index

[13]HEA: Hilgenreiner epiphyseal angle

Appendix 6: Round one survey

Delphi survey R1: A core measurement set for hip migration in cerebral palsy

Dear participant,

Thank you for taking part in this survey exercise. This project aims to identify the most important radiographic measurements used to determine the early signs of hip disease amongst children with cerebral palsy. The core measurements identified from this Delphi survey will be used to develop software which will measure hip migration automatically using artificial intelligence.

This survey is round one of a Delphi process designed to obtain the personal opinions of UK paediatric orthopaedic surgeons relating to the key issue. The initial measurement list has been derived from a systematic review of the literature from 2011 onwards.

The Delphi process will request your views on three different instances:

Round one: Participants will score a list of measurements and enter any additional measurements of importance that is not currently listed.

Round two: Participants will be presented with data from round one. The listed measurements from round one will then be rescored, with any new measurements added to the list.

Consensus meeting: The steering group of the Delphi process will perform a consensus setting exercise to form a final core measurement.

Your participation and responses will be confidential and anonymous for the entirety of the study. Data from this study will not be disclosed to any outside party or other participants of the Delphi process.

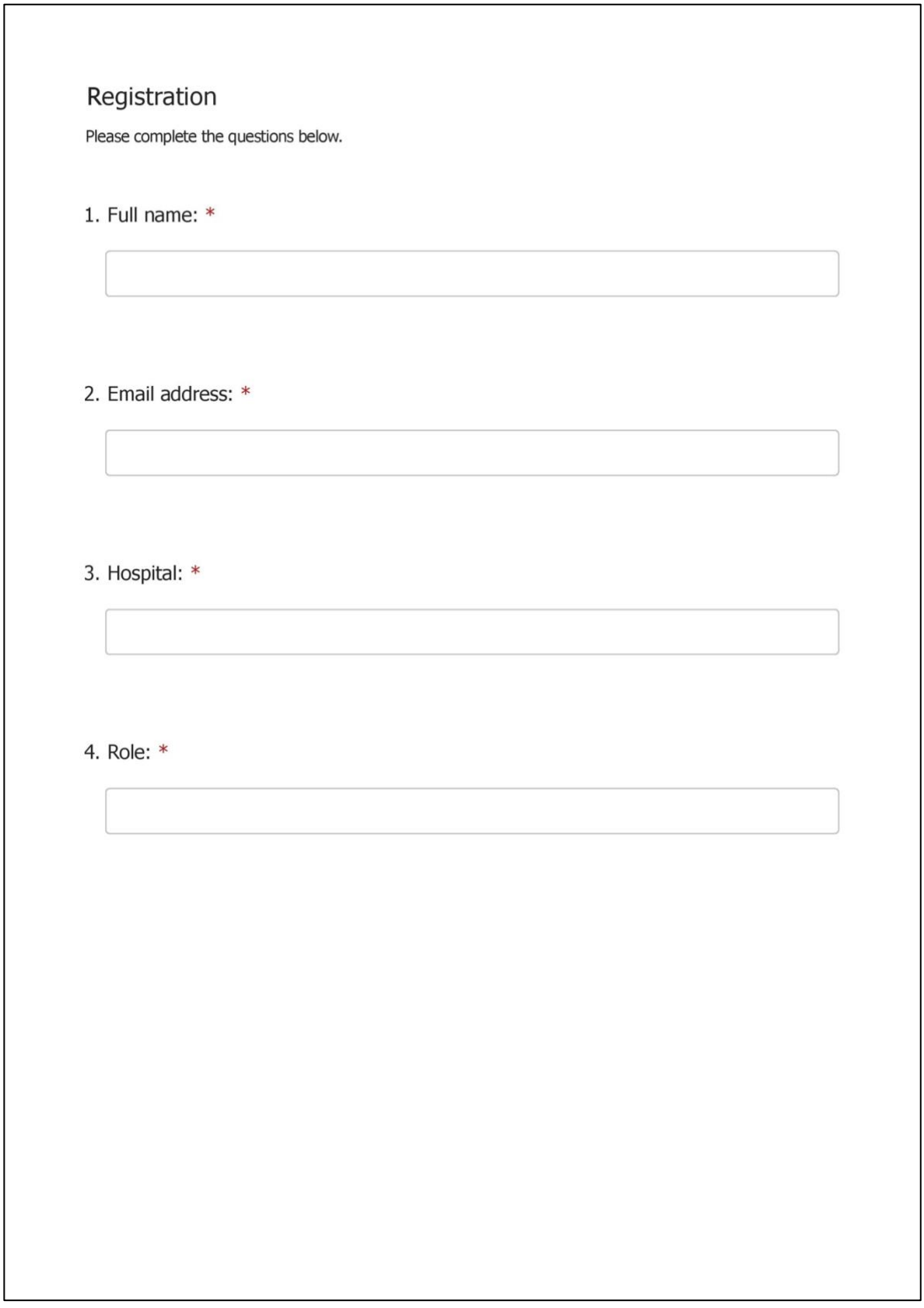
If you have any questions, please contact Josiah at hlpjosep@liverpool.ac.uk

Thank you for your assistance,

Prince Josiah Joseph

* Required

Appendix 6 (continued)



Registration

Please complete the questions below.

1. Full name: *

2. Email address: *

3. Hospital: *

4. Role: *

Appendix 6 (continued)

Scoring of core measurement set

Below is an initial list of measurements that have been identified from a systematic review looking at the radiographic measurements of hips in children with cerebral palsy. Please score all the measurements to the best of your knowledge.

A 9 point Likert scale will be used to determine the importance of each measurement.
Scores of 1–3 will be considered 'not important'.
Scores of 4–6 will be considered 'important but not critical'.
Scores of 7–9 will be considered 'critically important'.

5. Reimer's migration percentage (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

6. Pelvic obliquity (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

7. Femoral head-shaft angle HSA (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

8. Femoral neck-shaft angle (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

9. Mose hip ratio (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

Appendix 6 (continued)

10. Acetabular index (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

11. Centre edge angle (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

12. Sharp's angle or Acetabular angle (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

13. Shenton's line (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

14. Acetabular depth ratio (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

15. Epiphyseal shaft angle (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

Appendix 6 (continued)

16. Hilgenreiner epiphyseal angle (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

17. Pelvic femoral angle (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

18. Pelvic adjusted migration percentage (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

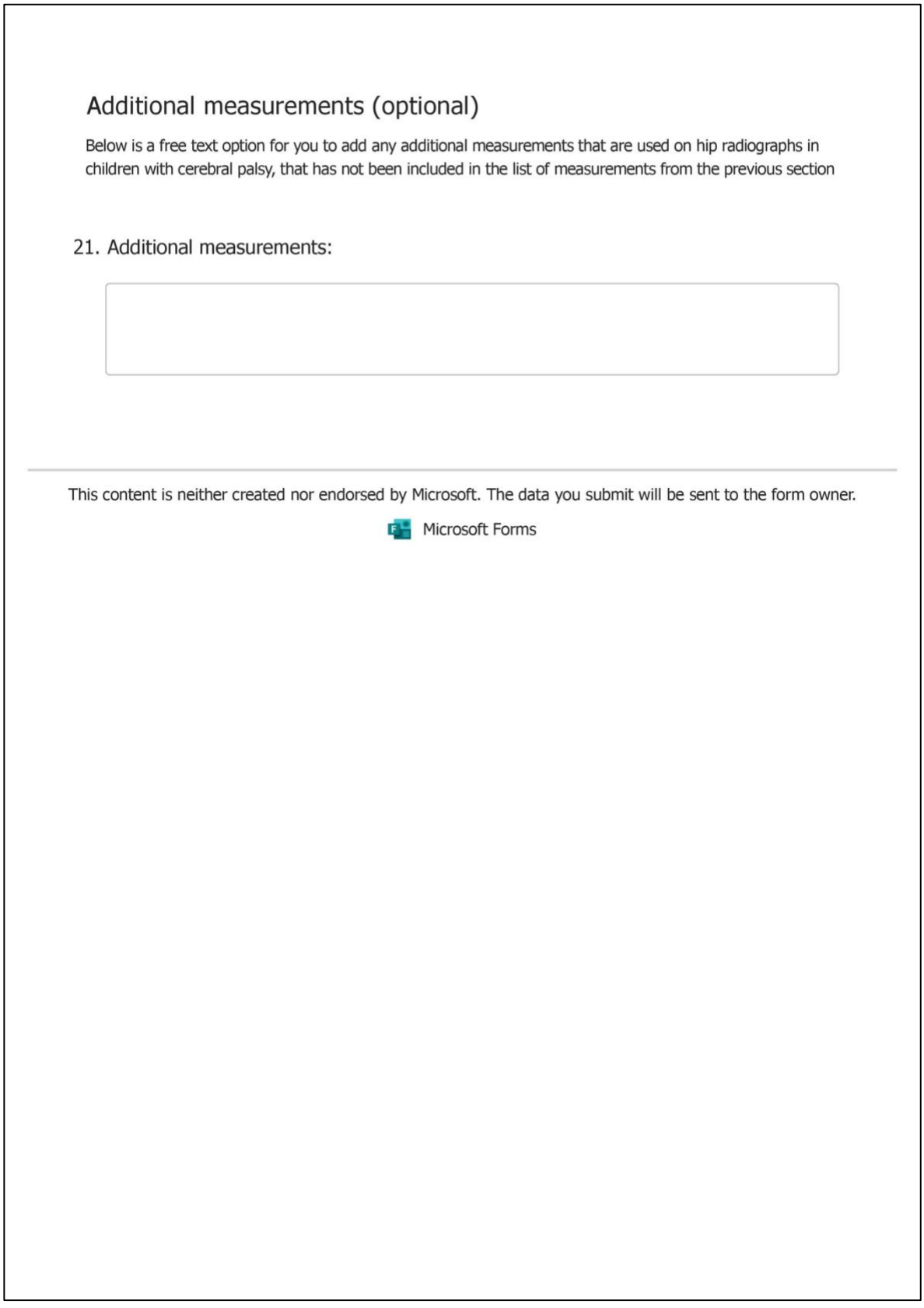
19. Medialization index (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

20. Do you want to add any additional measurements to this list? *

- Yes
- No

Appendix 6 (continued)



Additional measurements (optional)

Below is a free text option for you to add any additional measurements that are used on hip radiographs in children with cerebral palsy, that has not been included in the list of measurements from the previous section

21. Additional measurements:

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.

Microsoft Forms

Appendix 7: Delphi Round Two Survey

Delphi survey Round 2: A core measurement set for hip migration in cerebral palsy

Dear participant,

Thank you for taking part in this survey exercise. This project aims to identify the most important radiographic measurements used to determine the early signs of hip disease amongst children with cerebral palsy. The core measurements identified from this Delphi survey will be used to develop software which will measure hip migration automatically using artificial intelligence.

This survey is round two of the Delphi process designed to obtain the personal opinions of UK paediatric orthopaedic surgeons relating to the key issue.

The Delphi process will request your views on three different instances:

Round one: Participants will score a list of measurements and enter any additional measurements of importance that is not currently listed.

Round two: Participants will be presented with data from round one. The listed measurements from round one will then be rescored, with any new measurements added to the list.

Consensus meeting: The steering group of the Delphi process will perform a consensus setting exercise to form a final core measurement.

Your participation and responses will be confidential and anonymous for the entirety of the study. Data from this study will not be disclosed to any outside party or other participants of the Delphi process.

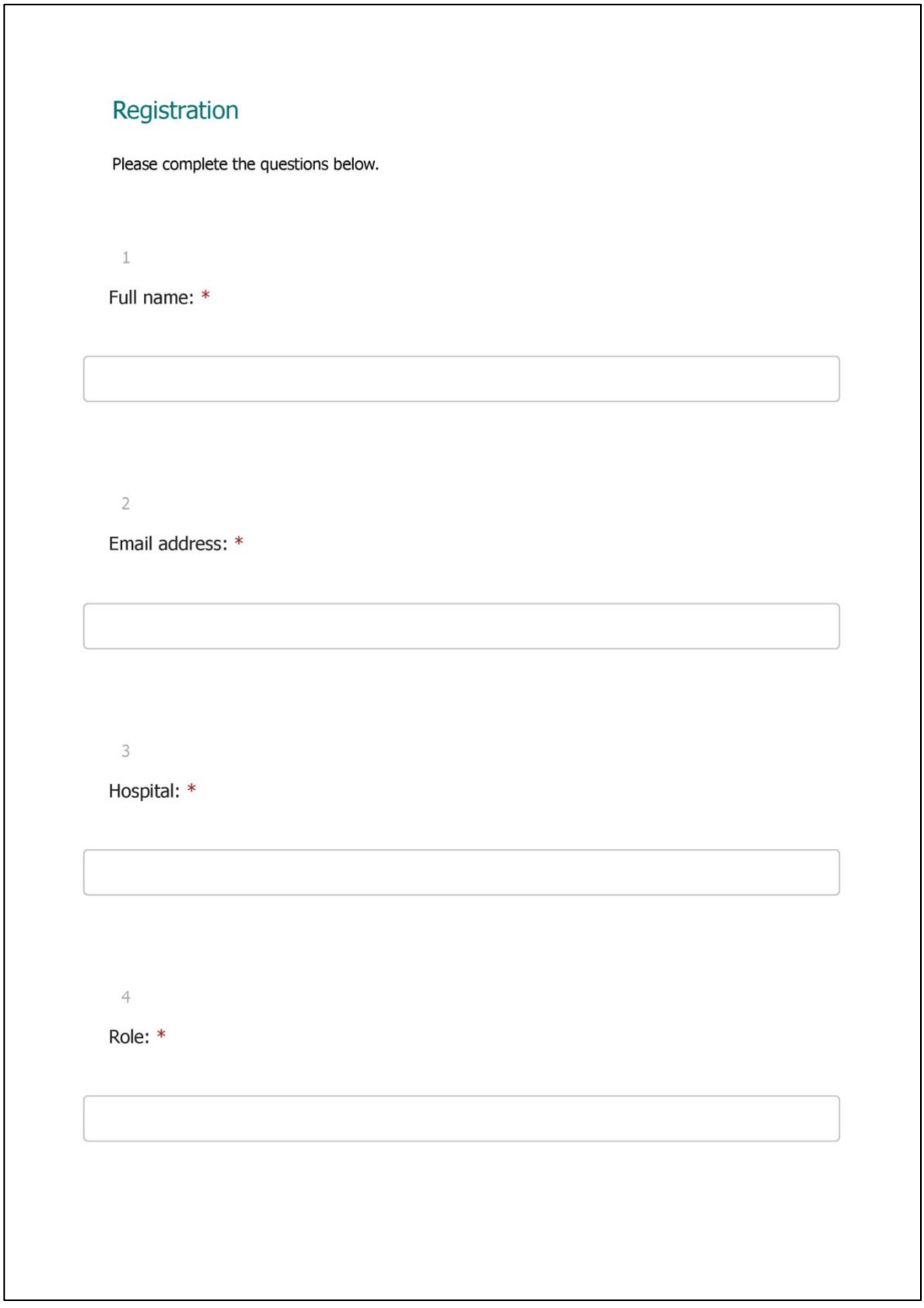
If you have any questions, please contact Josiah at hlpjosep@liverpool.ac.uk

Thank you for your assistance,

Prince Josiah Joseph

* Required

Appendix 7 (continued)



Registration

Please complete the questions below.

1

Full name: *

2

Email address: *

3

Hospital: *

4

Role: *

Appendix 7 (continued)

Measurements that have reached consensus in Round 1

Below are the measurements that have reached consensus in Round 1.

For a measurement to be included, the vast majority (> 70%) must agree that the measurement in question is 'critically important' with only a minority (< 15%) deeming it to be 'not important'.

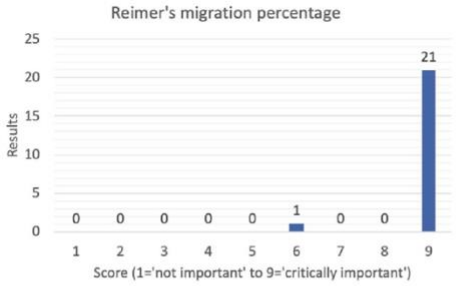
For a measurement to be excluded, the vast majority (> 70%) must agree that the measurement in question is 'not important', with only a small minority (< 15%) considering it to be 'critically important'.

No action is required in this section of the survey, however there is an optional comments section

5

Reimer's migration percentage

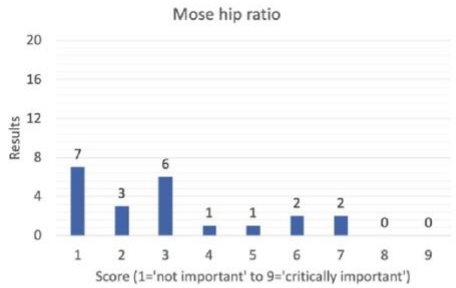
Voted in with 95% of participants deeming RMP as 'critically important' and 0% deeming it 'not important'.



6

Mose Hip Ratio

Voted out with 73% of participants deeming Mose hip ratio as 'not important' and 9.1% of participants deeming it as 'critically important'.



Appendix 7 (continued)

Scoring of measurements with results from Round 1

Below is the list of measurements along with the results from the 1st round of the Delphi process.

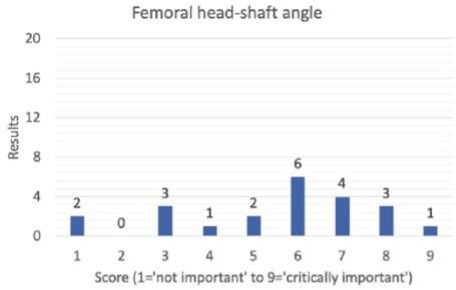
A total of 22 participants took part in the previous round.

Please score all the measurements to the best of your knowledge after reviewing the results of Round 1.

A 9 point Likert scale will be used to determine the importance of each measurement.
Scores of 1–3 will be considered 'not important'.
Scores of 4–6 will be considered 'important but not critical'.
Scores of 7–9 will be considered 'critically important'.

7

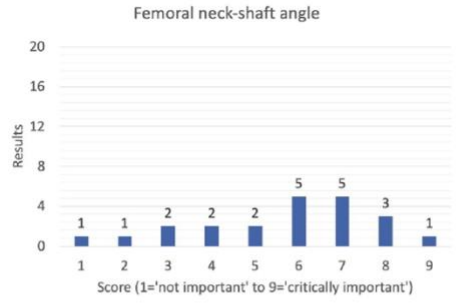
Femoral head-shaft angle HSA (1='not important' to 9='critically important') *



1 2 3 4 5 6 7 8 9

8

Femoral neck-shaft angle (1='not important' to 9='critically important') *

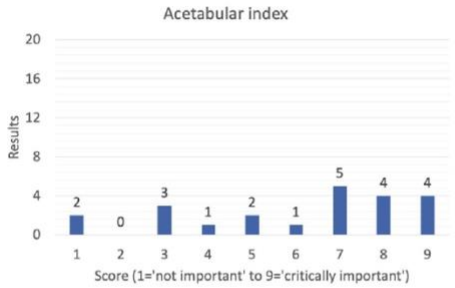


1 2 3 4 5 6 7 8 9

Appendix 7 (continued)

9

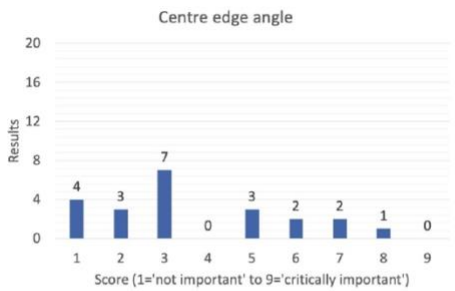
Acetabular index (1='not important' to 9='critically important') *



1 2 3 4 5 6 7 8 9

10

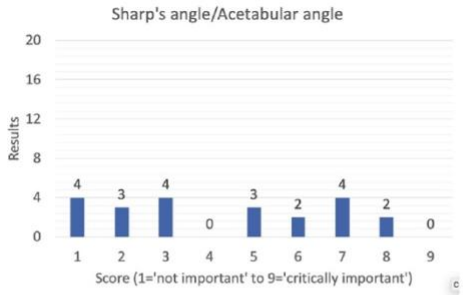
Centre edge angle (1='not important' to 9='critically important') *



1 2 3 4 5 6 7 8 9

11

Sharp's angle or Acetabular angle (1='not important' to 9='critically important') *

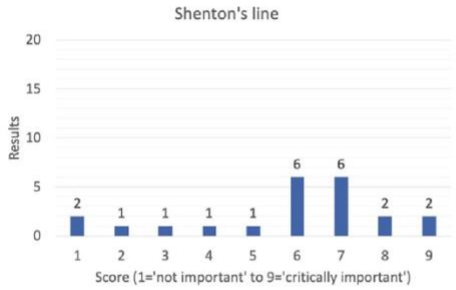


1 2 3 4 5 6 7 8 9

Appendix 7 (continued)

12

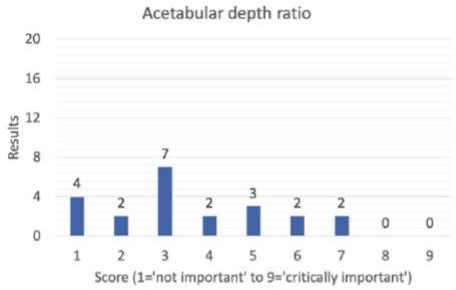
Shenton's line (1='not important' to 9='critically important') *



1 2 3 4 5 6 7 8 9

13

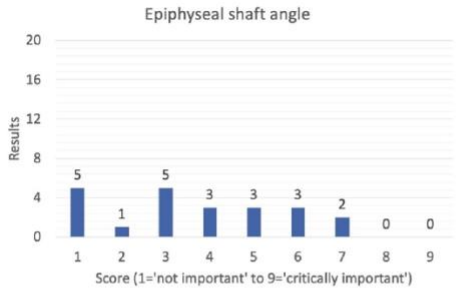
Acetabular depth ratio (1='not important' to 9='critically important') *



1 2 3 4 5 6 7 8 9

14

Epiphyseal shaft angle (1='not important' to 9='critically important') *

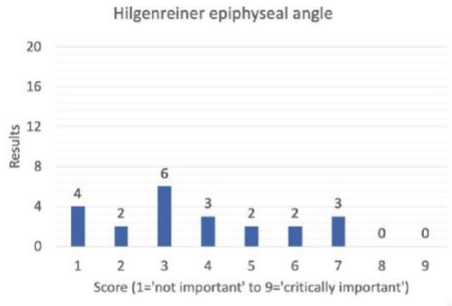


1 2 3 4 5 6 7 8 9

Appendix 7 (continued)

15

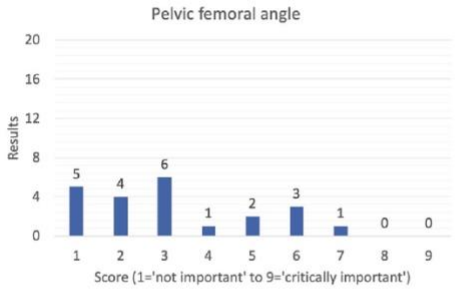
Hilgenreiner epiphyseal angle (1='not important' to 9='critically important') *



1 2 3 4 5 6 7 8 9

16

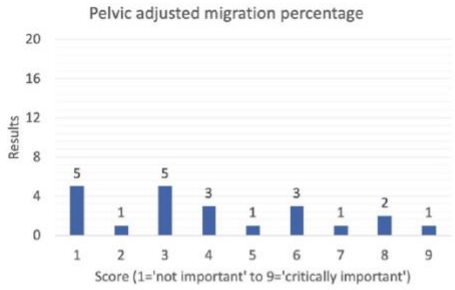
Pelvic femoral angle (1='not important' to 9='critically important') *



1 2 3 4 5 6 7 8 9

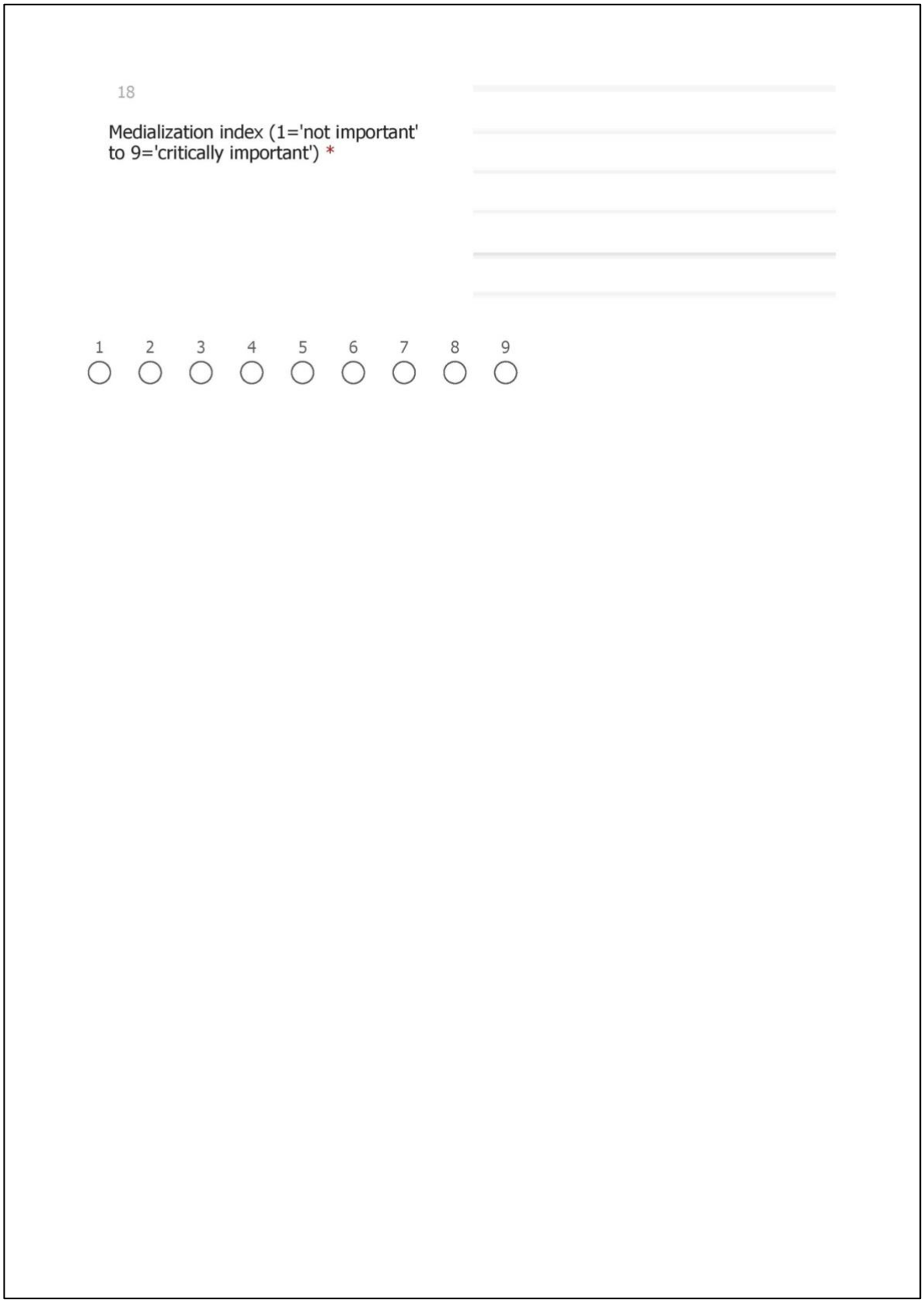
17

Pelvic adjusted migration percentage (1='not important' to 9='critically important') *



1 2 3 4 5 6 7 8 9

Appendix 7 (continued)



Appendix 7 (continued)

Scoring of new measurements

Below is a list a new measurements that have been suggested by partcipants from round one of the Delphi process.

Please score all the measurements to the best of your knowledge.

A 9 point Likert scale will be used to determine the importance of each measurement.
Scores of 1–3 will be considered 'not important'.
Scores of 4–6 will be considered 'important but not critical'.
Scores of 7–9 will be considered 'critically important'.

19

Sourcil (Tönnis) angle (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

20

Femoral head shape/congruency (1='not important' to 9='critically important') *

1 2 3 4 5 6 7 8 9

Appendix 8: Delphi Round One Responses

Appendix 8																
Participant ID	**Time (min:sec)**	**RMP**	**HSA**	**NSA**	**MHR**	**AcI**	**CEA**	**AA**	**SL**	**ADR**	**ESA**	**HEA**	**PFA**	**PAMP**	**MeI**	**Additional Measurements**
1	01:04	9	1	7	1	1	1	1	1	1	1	1	1	1	1	Nil
2	08:28	9	7	7	3	7	5	3	6	3	5	3	3	3	3	Nil
3	03:15	9	5	5	3	7	7	7	6	3	3	3	3	4	3	Nil
4	02:48	9	6	8	1	4	2	2	7	6	3	3	3	3	3	Nil
5	01:28	9	1	4	1	1	1	1	5	1	1	1	1	1	1	Nil
6	03:08	9	5	5	5	7	5	5	9	5	5	5	5	9	5	Nil
7	01:20	9	8	7	6	8	6	7	8	7	6	7	6	8	6	Nil
8	05:20	9	6	4	2	5	3	5	6	4	4	4	4	4	3	Nil
9	01:40	9	6	6	4	9	8	7	9	7	3	6	3	4	4	Nil
10	01:34	9	7	9	7	8	6	6	8	6	5	7	5	5	5	Nil
11	01:27	6	3	6	3	7	3	7	2	2	3	3	7	8	6	Nil
12	03:54	9	3	7	7	9	7	6	7	4	4	4	6	6	4	Nil
13	02:43	9	8	8	6	9	5	8	6	5	6	6	6	7	5	Nil
14	02:04	9	6	7	1	3	1	1	1	1	1	1	1	1	1	Nil
15	1192:27	9	7	6	1	8	3	5	7	3	1	7	1	1	1	Sourcil (Tönnis) angle Femoral head shape / congruency (not strictly a quantitative measurement)
16	16:11	9	6	2	2	3	2	3	7	3	1	2	2	6	2	Nil
17	05:08	9	7	6	1	5	1	1	4	1	4	1	1	1	1	Nil
18	02:48	9	6	3	3	9	3	8	7	5	7	5	3	3	3	Nil
19	02:30	9	9	3	3	3	3	3	3	3	3	3	3	3	3	Nil
20	03:19	9	3	8	1	8	3	3	6	2	2	2	2	6	2	Nil
21	01:53	9	8	6	3	6	2	2	7	3	7	3	2	3	2	Nil
22	01:54	9	4	1	2	7	3	2	6	3	6	4	2	2	2	Nil

Appendix 9: Delphi Round Two Responses

Appendix 9															
Participant ID	**Time (min:sec)**	**HSA**	**NSA**	**AcI**	**CEA**	**AA**	**SL**	**ADR**	**ESA**	**HEA**	**PFA**	**PAMP**	**MeI**	**STA**	**FHS**
1	03:26	1	9	1	1	1	5	1	1	1	1	1	1	1	1
2	05:20	6	7	6	3	3	7	3	5	3	2	5	3	1	7
3	04:42	5	5	6	4	5	7	5	5	4	5	4	3	5	7
4	02:52	6	7	7	3	3	7	3	3	3	3	3	3	3	7
5	02:42	3	7	3	3	1	7	1	1	1	1	1	1	2	8
6	05:40	6	7	7	3	3	7	3	1	3	1	5	3	5	9
7	02:34	7	7	9	7	7	8	3	3	3	3	3	3	7	8
8	02:49	7	7	4	2	2	5	3	2	3	2	2	3	3	5
10	01:35	3	7	8	5	6	7	3	1	1	1	2	2	5	7
11	04:19	5	8	8	5	5	2	7	7	8	2	3	4	5	8
12	02:43	5	8	9	7	5	6	5	5	5	5	8	5	7	9
13	03:51	7	7	8	5	7	8	5	5	5	7	7	4	4	7
14	04:56	6	7	2	2	2	2	2	2	2	1	2	2	2	8
15	11:06	7	8	8	3	6	9	3	3	7	2	3	1	8	8
16	05:16	7	4	7	1	3	7	3	1	1	1	5	1	3	5
17	05:53	7	6	7	2	1	6	2	3	2	1	3	2	2	6
18	03:21	6	3	8	2	2	8	1	1	3	3	1	1	2	6
19	06:57	2	9	1	1	1	1	1	1	1	1	1	1	1	1
20	05:05	3	6	7	3	2	6	2	3	3	2	3	3	3	6
21	04:54	8	5	7	2	3	7	2	8	5	2	5	1	2	6
22	02:52	2	1	7	1	1	4	1	4	3	1	1	1	1	1

Appendix 10: RMP values generated by each of the clinicians and the AI model for 450 randomly selected images from the training dataset. The 'manual' measurement is the mean of the measurements recorded by five clinicians for each image.

Appendix 10							
Image ID	Clinician 1	Clinician 2	Clinician 3	Clinician 4	Clinician 5	Manual	AI
1	12	10	11	8	17	11.6	15.1404
2	20	19	19	18	22	19.6	19.643
3	15	13	14	14	15	14.2	16.246
4	16	0	7	6	15	8.8	13.7143
5	17	18	18	15	20	17.6	10.0682
6	56	57	58	55	52	55.6	56.9104
7	52	42	50	45	45	46.8	51.3589
8	12	10	13	8	17	12	18.4969
9	41	40	39	39	45	40.8	51.1729
10	26	25	24	26	28	25.8	30.7776
11	19	12	11	10	14	13.2	13.8835
12	22	26	20	23	24	23	22.1916
13	12	8	7	10	18	11	12.2161
14	65	63	64	69	56	63.4	59.8752
15	33	34	33	32	33	33	33.7978
16	85	94	91	89	95	90.8	86.0601
17	100	87	99	88	82	91.2	100
18	11	8	11	12	13	11	9.592
19	78	100	100	58	72	81.6	92.9574
20	13	8	6	13	13	10.6	14.7477
21	54	54	58	54	41	52.2	52.1079
22	16	7	7	7	10	9.4	15.5345
23	94	94	94	94	95	94.2	94.909
24	11	8	12	7	16	10.8	9.97121
25	43	44	46	48	44	45	48.3322
26	34	30	32	30	23	29.8	29.3286
27	10	0	7	8	15	8	8.66488
28	19	2	14	7	9	10.2	0.55829
29	42	39	40	37	37	39	35.4541
30	39	37	37	29	40	36.4	43.9202
31	83	83	81	81	74	80.4	86.9551
32	100	95	99	95	99	97.6	99.5701
33	60	77	77	78	80	74.4	80.6679
34	17	14	11	12	15	13.8	16.6417

Appendix 10 (continued)							
Image ID	Clinician 1	Clinician 2	Clinician 3	Clinician 4	Clinician 5	Manual	AI
35	24	25	23	23	27	24.4	29.9589
36	38	34	46	25	49	38.4	41.9377
37	62	79	56	20	10	45.4	32.2581
38	53	42	56	45	40	47.2	49.2717
39	100	100	100	100	100	100	100
40	100	0	62	95	0	51.4	81.1082
41	10	2	16	10	29	13.4	16.7362
42	81	79	79	74	75	77.6	80.5523
43	30	27	27	25	29	27.6	29.6751
44	22	24	20	22	25	22.6	34.9067
45	25	22	25	31	34	27.4	33.633
46	13	5	17	8	13	11.2	8.76511
47	40	38	38	37	39	38.4	38.9373
48	18	10	13	9	10	12	12.1403
49	19	18	21	19	21	19.6	24.201
50	23	23	20	20	22	21.6	28.9348
51	32	27	26	25	26	27.2	28.6615
52	71	64	65	62	59	64.2	65.061
53	24	20	18	17	19	19.6	21.4921
54	25	8	14	12	24	16.6	14.9124
55	32	29	29	30	32	30.4	37.4037
56	0	0	0	0	0	0	29.6948
57	92	92	100	92	98	94.8	100
58	33	31	33	30	32	31.8	32.9242
59	29	17	28	16	29	23.8	23.973
60	95	96	90	87	83	90.2	100
61	15	14	14	11	15	13.8	15.1789
62	21	17	18	14	23	18.6	15.8523
63	25	0	2	0	2	5.8	1.01129
64	32	32	31	29	35	31.8	31.3489
65	20	16	15	12	17	16	16.605
66	25	20	23	18	24	22	36.9026
67	31	26	25	24	25	26.2	29.9798
68	0	0	14	0	0	2.8	0
69	55	55	57	53	50	54	54.846
70	62	42	45	56	57	52.4	34.8218
71	25	26	25	23	28	25.4	24.3794
72	100	56	91	54	81	76.4	65.4855

Appendix 10 (continued)							
Image ID	Clinician 1	Clinician 2	Clinician 3	Clinician 4	Clinician 5	Manual	AI
73	14	8	10	8	19	11.8	18.5568
74	48	49	48	48	48	48.2	53.3842
75	30	33	31	31	31	31.2	28.1076
76	68	69	67	70	74	69.6	71.5029
77	16	12	14	15	16	14.6	17.4906
78	70	66	67	67	52	64.4	63.3347
79	68	66	72	68	73	69.4	74.8457
80	39	35	36	36	31	35.4	34.9216
81	100	49	44	49	64	61.2	75.3812
82	20	10	16	16	31	18.6	39.1422
83	62	57	58	58	53	57.6	57.1107
84	29	26	26	26	23	26	32.4564
85	19	14	14	9	19	15	22.2325
86	31	28	26	28	28	28.2	33.4915
87	76	75	74	77	100	80.4	78.7379
88	51	49	54	50	55	51.8	53.1993
89	54	43	54	44	55	50	48.5607
90	35	20	21	19	21	23.2	20.5482
91	72	68	67	69	80	71.2	73.1748
92	26	24	22	23	33	25.6	24.3886
93	20	15	14	13	19	16.2	14.1656
94	34	22	24	23	32	27	26.809
95	100	100	100	100	100	100	100
96	16	9	11	11	12	11.8	10.1184
97	18	13	7	11	20	13.8	9.94869
98	46	33	30	29	31	33.8	30.4005
99	15	5	7	10	12	9.8	0
100	41	35	37	37	38	37.6	30.2937
101	41	34	45	39	40	39.8	45.5559
102	29	25	30	25	25	26.8	28.877
103	72	71	69	68	60	68	61.331
104	24	19	20	14	27	20.8	26.7228
105	84	82	83	80	80	81.8	86.6634
106	0	0	0	0	3	0.6	0
107	0	0	2	0	22	4.8	0.12165
108	100	87	91	91	100	93.8	69.3409
109	23	22	22	21	22	22	34.7766
110	24	21	22	20	25	22.4	17.5631

Appendix 10 (continued)							
Image ID	Clinician 1	Clinician 2	Clinician 3	Clinician 4	Clinician 5	Manual	AI
111	13	0	7	4	18	8.4	7.86095
112	11	12	12	14	20	13.8	23.2307
113	22	17	18	25	21	20.6	23.2241
114	100	100	100	100	100	100	100
115	29	22	22	18	22	22.6	27.2211
116	1	2	0	4	2	1.8	1.34587
117	11	7	8	8	8	8.4	0
118	5	0	0	2	5	2.4	0
119	22	21	22	22	20	21.4	43.3276
120	28	17	19	19	24	21.4	22.6169
121	19	16	18	18	20	18.2	24.0284
122	0	0	0	0	0	0	0
123	13	7	9	7	18	10.8	19.4031
124	38	33	37	27	32	33.4	57.1012
125	23	21	22	20	27	22.6	31.7515
126	23	10	14	14	24	17	22.9571
127	17	19	17	17	20	18	25.0707
128	17	14	15	15	17	15.6	34.5858
129	12	10	16	13	24	15	19.1415
130	31	24	31	27	33	29.2	33.1954
131	9	5	5	10	21	10	6.82425
132	17	10	16	15	29	17.4	16.6277
133	13	8	14	11	13	11.8	10.1178
134	24	25	22	22	28	24.2	24.6775
135	38	39	40	43	39	39.8	39.909
136	72	69	71	69	73	70.8	80.5849
137	19	17	20	16	20	18.4	17.5973
138	23	14	17	21	21	19.2	24.6357
139	0	0	0	1	12	2.6	0.86285
140	20	21	20	16	27	20.8	20.0569
141	22	23	25	22	29	24.2	24.9507
142	19	9	16	6	11	12.2	9.38745
143	0	22	22	22	25	18.2	0
144	76	76	67	66	69	70.8	67.7822
145	14	14	25	16	25	18.8	20.6045
146	31	46	38	32	36	36.6	35.8336
147	100	98	96	97	97	97.6	100
148	15	11	11	9	13	11.8	11.9322

Appendix 10 (continued)							
Image ID	Clinician 1	Clinician 2	Clinician 3	Clinician 4	Clinician 5	Manual	AI
149	24	21	22	21	26	22.8	22.8452
150	11	5	5	6	14	8.2	11.5267
151	74	73	71	74	75	73.4	72.0905
152	27	0	0	5	18	10	11.2618
153	19	14	19	13	18	16.6	21.1597
154	18	11	14	14	16	14.6	16.9379
155	28	23	24	18	24	23.4	18.9346
156	70	44	61	38	67	56	21.6328
157	38	30	35	26	34	32.6	24.9099
158	93	94	90	91	71	87.8	89.5634
159	100	100	100	100	100	100	100
160	14	8	8	12	10	10.4	12.4226
161	0	0	0	11	3	2.8	1.60097
162	36	34	34	34	33	34.2	39.9526
163	0	0	0	0	1	0.2	1.83402
164	36	36	34	34	37	35.4	10.3591
165	79	81	81	82	87	82	100
166	26	30	29	28	28	28.2	30.6492
167	19	22	21	19	22	20.6	33.3286
168	20	21	17	19	22	19.8	27.2762
169	96	86	87	86	82	87.4	84.693
170	36	50	38	35	35	38.8	36.0883
171	28	21	26	25	24	24.8	22.5423
172	0	0	0	4	5	1.8	0
173	15	12	16	13	16	14.4	16.7571
174	63	64	63	63	65	63.6	66.7714
175	100	100	100	100	100	100	70.2846
176	10	3	6	3	5	5.4	9.48024
177	31	31	32	31	32	31.4	32.7754
178	37	32	34	31	35	33.8	39.0436
179	10	5	7	8	12	8.4	1.12666
180	15	17	16	18	26	18.4	23.1217
181	11	0	0	0	7	3.6	0.71434
182	13	16	18	15	26	17.6	23.257
183	12	11	11	15	18	13.4	16.4727
184	38	0	25	24	26	22.6	23.9426
185	24	27	27	28	27	26.6	22.7825
186	21	18	19	17	19	18.8	17.6844

Appendix 10 (continued)							
Image ID	Clinician 1	Clinician 2	Clinician 3	Clinician 4	Clinician 5	Manual	AI
187	19	22	17	17	24	19.8	16.6498
188	21	17	14	17	18	17.4	20.6118
189	18	5	12	12	11	11.6	13.7869
190	77	81	82	80	81	80.2	82.4704
191	27	20	21	23	22	22.6	36.4503
192	21	19	25	22	26	22.6	27.4751
193	47	45	47	45	39	44.6	33.345
194	12	11	12	13	21	13.8	13.4727
195	23	11	17	10	18	15.8	20.2164
196	0	0	23	20	18	12.2	18.5729
197	23	21	23	23	22	22.4	23.7162
198	56	50	57	50	47	52	51.871
199	10	7	11	10	10	9.6	12.7641
200	58	58	58	56	57	57.4	55.357
201	3	6	6	5	6	5.2	4.88928
202	39	35	37	35	38	36.8	37.2158
203	18	7	18	19	20	16.4	8.41785
204	28	31	31	29	34	30.6	36.6434
205	21	23	24	24	29	24.2	26.1572
206	22	24	24	21	20	22.2	22.0257
207	6	90	100	92	96	76.8	100
208	83	89	82	78	81	82.6	100
209	13	12	6	10	9	10	9.3899
210	29	27	30	24	26	27.2	29.9651
211	23	23	26	25	25	24.4	27.595
212	49	54	50	50	51	50.8	50.5331
213	22	25	25	24	25	24.2	37.65
214	24	21	27	23	24	23.8	26.0618
215	25	20	22	21	23	22.2	31.8022
216	16	15	17	15	29	18.4	20.606
217	62	29	29	39	33	38.4	18.3683
218	81	81	76	75	75	77.6	79.1237
219	20	5	5	16	16	12.4	0
220	27	28	31	30	28	28.8	35.4987
221	41	40	36	41	33	38.2	39.9774
222	76	80	78	78	78	78	85.5284
223	14	5	8	6	6	7.8	12.5044
224	36	26	34	32	36	32.8	54.0935

Appendix 10 (continued)							
Image ID	Clinician 1	Clinician 2	Clinician 3	Clinician 4	Clinician 5	Manual	AI
225	6	7	7	7	14	8.2	23.0904
226	26	27	27	21	24	25	19.6775
227	0	0	0	0	8	1.6	0
228	37	26	33	27	32	31	33.0579
229	15	11	14	13	17	14	10.2739
230	26	22	25	26	29	25.6	28.0881
231	39	43	43	39	46	42	41.8559
232	11	2	4	5	1	4.6	10.0103
233	37	30	32	33	34	33.2	34.1853
234	0	0	0	0	0	0	0
235	20	12	16	16	14	15.6	12.1326
236	18	18	11	21	25	18.6	45.3084
237	18	14	19	20	19	18	21.2733
238	41	42	41	41	47	42.4	49.2319
239	22	22	21	23	24	22.4	26.3037
240	7	4	9	6	11	7.4	6.00244
241	19	20	20	19	21	19.8	30.0129
242	23	19	19	17	21	19.8	31.5756
243	46	36	40	43	55	44	50.5914
244	15	17	18	18	23	18.2	20.3076
245	71	79	72	69	65	71.2	100
246	29	22	30	29	32	28.4	39.1704
247	0	1	0	7	10	3.6	13.132
248	57	52	77	51	52	57.8	54.2188
249	21	14	19	15	19	17.6	19.7637
250	39	37	37	39	45	39.4	42.018
251	44	40	42	43	47	43.2	47.9853
252	14	11	6	9	19	11.8	14.1402
253	17	14	19	19	17	17.2	18.2555
254	63	63	66	49	69	62	0
255	6	0	8	3	4	4.2	3.85129
256	92	92	91	88	87	90	90.2851
257	41	34	30	30	26	32.2	33.4217
258	13	7	11	11	11	10.6	16.4304
259	24	3	2	0	8	7.4	0.90308
260	24	20	23	22	24	22.6	23.8062
261	20	18	19	18	19	18.8	21.9772
262	100	93	98	100	97	97.6	90.3108

Appendix 10 (continued)							
Image ID	Clinician 1	Clinician 2	Clinician 3	Clinician 4	Clinician 5	Manual	AI
263	0	0	0	0	0	0	0
264	31	18	18	22	21	22	41.1059
265	7	3	6	6	9	6.2	11.0864
266	0	0	0	0	0	0	0
267	36	14	30	14	32	25.2	33.6077
268	13	1	13	0	0	5.4	2.94593
269	22	16	16	17	25	19.2	17.1307
270	17	11	17	16	20	16.2	42.9657
271	11	0	6	0	4	4.2	1.70083
272	19	21	25	20	22	21.4	26.1172
273	18	15	16	16	17	16.4	22.5221
274	59	45	39	42	41	45.2	47.1271
275	34	26	30	30	26	29.2	32.6319
276	43	21	24	16	17	24.2	19.3309
277	100	100	100	100	100	100	100
278	19	14	14	13	14	14.8	15.0715
279	27	22	27	26	28	26	35.0576
280	24	18	21	18	23	20.8	26.6594
281	49	49	46	49	48	48.2	47.3735
282	53	55	57	56	50	54.2	54.6188
283	10	0	0	0	1	2.2	0.84879
284	30	25	25	24	25	25.8	29.9709
285	100	100	100	100	92	98.4	100
286	23	18	22	23	24	22	27.3124
287	36	33	35	37	35	35.2	36.5684
288	25	24	25	24	24	24.4	38.9383
289	51	44	48	46	36	45	51.1323
290	27	28	31	29	31	29.2	29.3734
291	32	26	26	26	26	27.2	38.2534
292	17	16	16	19	19	17.4	7.60493
293	60	59	60	61	62	60.4	68.2421
294	26	23	22	22	22	23	22.6775
295	23	11	16	26	15	18.2	23.3345
296	27	11	16	8	9	14.2	8.87414
297	25	25	25	24	24	24.6	30.5736
298	28	27	28	30	32	29	27.4182
299	16	12	12	14	13	13.4	31.3785
300	27	25	25	25	18	24	26.9776

Appendix 10 (continued)							
Image ID	Clinician 1	Clinician 2	Clinician 3	Clinician 4	Clinician 5	Manual	AI
301	15	17	20	16	17	17	14.1888
302	54	45	53	43	44	47.8	0
303	28	20	19	19	15	20.2	19.627
304	52	50	52	50	50	50.8	42.5127
305	24	22	24	23	24	23.4	26.8572
306	25	18	22	17	21	20.6	17.5759
307	12	1	1	4	2	4	7.75526
308	11	1	4	6	6	5.6	8.06986
309	13	9	7	10	12	10.2	13.2371
310	17	9	14	10	14	12.8	3.89806
311	11	0	4	0	1	3.2	0
312	5	0	2	6	4	3.4	0
313	43	40	16	27	39	33	22.4886
314	10	7	8	7	9	8.2	2.37765
315	14	13	14	14	15	14	16.8922
316	21	22	22	20	19	20.8	25.9281
317	100	100	100	100	100	100	100
318	28	29	37	28	28	30	34.8235
319	31	29	29	28	27	28.8	33.3293
320	77	82	77	79	79	78.8	83.0331
321	100	99	100	100	99	99.6	0
322	70	72	72	73	74	72.2	80.6687
323	31	29	28	29	32	29.8	33.0331
324	36	37	36	33	32	34.8	32.7627
325	16	18	16	17	15	16.4	27.6267
326	3	1	5	3	3	3	8.72994
327	23	19	19	13	17	18.2	15.7901
328	23	17	24	18	21	20.6	10.5761
329	37	12	13	11	10	16.6	22.0773
330	24	20	19	21	22	21.2	24.3439
331	15	12	13	12	12	12.8	20.3251
332	0	0	0	0	0	0	1.38181
333	100	100	100	100	100	100	100
334	30	29	34	31	29	30.6	32.8635
335	22	19	22	20	26	21.8	24.1505
336	25	23	25	27	24	24.8	33.2677
337	37	36	42	33	43	38.2	40.9447
338	68	54	71	55	71	63.8	56.6552

Appendix 10 (continued)							
Image ID	Clinician 1	Clinician 2	Clinician 3	Clinician 4	Clinician 5	Manual	AI
339	24	24	31	27	29	27	31.6345
340	63	71	71	73	67	69	100
341	40	38	37	37	41	38.6	87.8709
342	16	12	14	13	13	13.6	19.6755
343	34	44	40	31	39	37.6	37.1364
344	43	39	45	39	41	41.4	42.237
345	13	14	16	13	20	15.2	21.3892
346	30	30	29	29	27	29	31.5562
347	12	6	8	6	8	8	9.34866
348	32	0	0	0	33	13	54.2606
349	15	13	13	11	15	13.4	11.6779
350	27	27	27	30	27	27.6	32.2995
351	32	30	31	31	34	31.6	37.7682
352	17	21	20	17	19	18.8	29.4867
353	23	25	23	20	21	22.4	23.2299
354	28	24	27	25	29	26.6	42.1719
355	12	11	12	11	11	11.4	9.09496
356	49	50	54	52	59	52.8	59.7987
357	0	0	0	0	0	0	36.668
358	30	33	30	30	30	30.6	31.219
359	14	13	13	13	17	14	23.6339
360	0	0	0	0	0	0	0
361	35	32	32	33	32	32.8	36.7112
362	21	24	21	19	24	21.8	27.2574
363	37	35	35	34	36	35.4	48.7566
364	8	9	8	9	10	8.8	12.9725
365	24	23	28	23	25	24.6	26.2454
366	2	2	0	2	7	2.6	6.06522
367	11	10	10	5	13	9.8	13.6797
368	27	14	14	11	14	16	10.1501
369	26	27	25	26	28	26.4	27.8278
370	46	45	45	37	38	42.2	42.6926
371	15	13	15	13	16	14.4	18.0504
372	76	79	77	77	82	78.2	86.5901
373	51	46	44	41	40	44.4	45.8339
374	31	25	31	26	28	28.2	35.9209
375	22	24	23	23	22	22.8	20.8564
376	92	96	97	96	99	96	100

Appendix 10 (continued)							
Image ID	Clinician 1	Clinician 2	Clinician 3	Clinician 4	Clinician 5	Manual	AI
377	17	13	12	10	15	13.4	32.5982
378	100	95	85	84	94	91.6	0
379	100	100	100	100	88	97.6	100
380	10	5	8	6	8	7.4	6.18104
381	42	59	41	41	58	48.2	59.4743
382	29	32	32	30	28	30.2	25.9578
383	23	23	22	20	24	22.4	36.1543
384	71	63	63	67	61	65	62.8971
385	56	57	58	57	52	56	61.4929
386	59	56	62	59	56	58.4	64.5424
387	0	0	0	0	0	0	2.13934
388	15	16	18	15	17	16.2	17.7718
389	9	0	0	0	12	4.2	15.8478
390	80	0	0	0	0	16	0
391	70	73	72	71	59	69	73.9973
392	0	0	0	0	11	2.2	0
393	18	18	19	17	22	18.8	18.645
394	8	4	6	7	6	6.2	4.95016
395	38	35	35	34	38	36	48.3615
396	17	16	18	18	22	18.2	23.9376
397	19	5	7	6	18	11	10.1523
398	4	5	7	8	9	6.6	6.8958
399	23	14	23	0	7	13.4	9.85073
400	27	24	27	26	27	26.2	33.3172
401	27	25	24	26	22	24.8	31.2551
402	13	13	17	12	16	14.2	16.2968
403	0	0	0	0	0	0	0
404	30	41	29	41	42	36.6	38.0337
405	16	21	29	25	26	23.4	26.8589
406	8	4	7	5	8	6.4	10.1635
407	6	0	10	6	2	4.8	7.8898
408	16	13	15	16	18	15.6	12.6255
409	46	49	46	46	39	45.2	57.2022
410	15	14	19	14	21	16.6	19.166
411	26	18	24	18	28	22.8	21.8422
412	11	12	22	13	18	15.2	21.319
413	25	12	17	13	18	17	23.7864
414	8	6	10	8	12	8.8	16.245

Appendix 10 (continued)							
Image ID	Clinician 1	Clinician 2	Clinician 3	Clinician 4	Clinician 5	Manual	AI
415	9	11	11	13	12	11.2	17.5508
416	13	13	16	15	17	14.8	18.1738
417	33	23	31	26	31	28.8	29.4578
418	24	23	23	25	28	24.6	33.2147
419	22	23	24	23	24	23.2	25.9586
420	12	12	13	12	9	11.6	19.7927
421	23	22	22	23	22	22.4	27.6152
422	20	5	12	9	42	17.6	89.0446
423	13	12	9	10	12	11.2	1.26136
424	42	40	41	38	38	39.8	48.4493
425	78	76	72	79	63	73.6	75.2975
426	20	18	18	21	22	19.8	21.2423
427	13	8	10	8	14	10.6	15.094
428	12	10	9	5	7	8.6	2.20573
429	19	20	21	20	21	20.2	25.4772
430	8	5	10	8	12	8.6	12.7766
431	31	31	31	31	32	31.2	34.2777
432	23	19	20	18	20	20	25.2756
433	29	27	29	24	30	27.8	30.8596
434	29	27	27	29	29	28.2	34.1785
435	16	15	14	16	17	15.6	20.5195
436	32	33	32	32	32	32.2	33.4537
437	6	0	4	3	5	3.6	0
438	28	25	31	28	30	28.4	35.1116
439	29	27	29	30	30	29	28.6925
440	7	0	6	7	6	5.2	11.1275
441	13	12	11	12	13	12.2	20.0801
442	23	23	23	23	46	27.6	28.3819
443	33	34	35	34	42	35.6	41.7584
444	100	100	100	100	100	100	100
445	14	13	13	12	10	12.4	15.8576
446	12	5	12	9	14	10.4	10.591
447	13	13	14	13	12	13	23.3781
448	9	6	10	7	13	9	12.2612
449	21	25	21	16	29	22.4	25.9707
450	18	20	19	21	20	19.6	27.525

Appendix 11: Descriptive statistics for point-to-point distances (mm) for the pelvis annotations between AI model and AI Trainers.

Appendix 11							
Point	Mean	Standard error	Standard deviation	Median	90%	95%	99%
0	4.01	0.265	5.61	2.16	9.47	15.4	27.3
1	3.09	0.198	4.2	1.77	7.11	10.4	22.3
2	2.56	0.16	3.39	1.48	5.52	9.39	16.4
3	2.54	0.189	4.00	1.24	5.52	10.8	20.8
4	2.21	0.163	3.46	1.1	4.98	8.80	18.3
5	2.05	0.13	2.75	1.19	4.19	8.18	15.5
6	2.16	0.118	2.51	1.38	4.54	7.03	13.6
7	2.38	0.121	2.55	1.52	5.42	7.39	12.4
8	2.72	0.131	2.78	1.78	6.07	8.28	15.0
9	3.07	0.15	3.18	1.92	7.61	9.44	16.1
10	2.60	0.128	2.7	1.63	6.42	8.00	13.0
11	2.27	0.119	2.53	1.22	6.40	7.76	11.6
12	2.38	0.135	2.87	1.27	6.35	8.79	12.7
13	2.34	0.123	2.62	1.37	5.83	8.31	12.8
14	2.55	0.123	2.61	1.68	6.04	9.05	11.8
15	2.68	0.138	2.91	1.67	6.43	8.47	13.2
16	2.73	0.167	3.53	1.5	6.49	9.47	18.4
17	3.07	0.203	4.29	1.5	7.32	11.7	22.5
18	2.77	0.149	3.15	1.81	6.02	8.42	15.1
19	2.60	0.129	2.73	1.72	5.31	7.93	14.7
20	2.52	0.129	2.74	1.67	5.14	7.71	16.2
21	2.58	0.145	3.06	1.57	5.46	8.61	15.5
22	2.70	0.15	3.18	1.63	6.25	8.33	17.2
23	2.41	0.134	2.83	1.4	5.55	7.79	14.8
24	2.31	0.129	2.74	1.42	5.59	7.88	13.9
25	2.37	0.128	2.71	1.42	5.31	7.59	14.0
26	1.77	0.155	3.29	0.861	3.59	7.28	17.7
27	1.79	0.124	2.62	1.06	3.37	6.52	16.0
28	2.05	0.126	2.67	1.21	4.15	6.44	15.8
29	2.08	0.148	3.15	1.06	4.05	8.22	16.5
30	3.81	0.244	5.17	2.17	9.29	14.2	24.4
31	3.02	0.184	3.9	1.79	6.82	10.4	16.8
32	2.37	0.157	3.32	1.44	4.66	7.82	15.9
33	2.29	0.188	3.99	1.11	4.93	8.60	19.9
34	2.00	0.172	3.64	1.03	4.03	6.83	19.0
35	1.97	0.159	3.36	1.13	3.96	5.41	17.4
36	2.12	0.155	3.28	1.31	4.24	5.58	14.9
37	2.42	0.157	3.32	1.62	5.07	6.08	14.8
38	2.74	0.158	3.34	1.86	5.62	6.74	13.3
39	3.12	0.169	3.57	2.07	6.59	8.32	14.8
40	2.72	0.158	3.35	1.76	5.63	7.65	12.2
41	2.41	0.154	3.27	1.38	5.36	7.29	11.2

Appendix 11 (continued)							
Point	Mean	Standard error	Standard deviation	Median	90%	95%	99%
42	2.43	0.159	3.38	1.31	5.81	8.05	12.2
43	2.26	0.145	3.07	1.37	5.21	7.45	11.5
44	2.40	0.141	2.99	1.56	5.16	7.60	11.9
45	2.55	0.149	3.16	1.56	5.26	8.08	14.4
46	2.71	0.168	3.57	1.49	5.92	8.55	18.5
47	3.00	0.194	4.11	1.39	7.03	10.6	23.1
48	2.69	0.162	3.43	1.81	5.06	7.81	14.6
49	2.59	0.161	3.42	1.67	5.18	7.66	13.7
50	2.63	0.167	3.53	1.58	5.53	7.88	15.1
51	2.66	0.169	3.57	1.59	5.71	7.85	14.9
52	2.81	0.174	3.69	1.7	6.38	8.01	14.0
53	2.57	0.169	3.58	1.44	5.62	7.42	14.4
54	2.47	0.167	3.53	1.36	5.74	7.18	13.9
55	2.62	0.169	3.59	1.48	5.98	7.76	14.0
56	1.63	0.156	3.30	0.8	2.99	6.01	13.5
57	1.77	0.142	3.02	1.05	3.18	5.90	16.1
58	2.18	0.159	3.38	1.26	3.86	6.36	18.9
59	2.33	0.177	3.74	1.2	4.41	8.17	20.4

Appendix 12: Descriptive statistics for point-to-point distances (mm) for the proximal femur annotations between AI model and AI Trainers.

Appendix 12							
Point	Mean	Standard error	Standard deviation	Median	90%	95%	99%
0	11.2	0.615	13.0	6.01	28.4	35.6	54.7
1	9.96	0.553	11.7	5.36	24.7	30.6	51.9
2	8.98	0.496	10.5	5.02	20.6	27.8	52.3
3	7.95	0.451	9.55	4.89	16.9	23.7	50.7
4	6.73	0.417	8.84	4.29	13.2	19.3	49.4
5	5.41	0.394	8.35	3.39	9.91	15.6	49.0
6	4.39	0.383	8.11	2.4	7.83	12.2	48.4
7	4.66	0.385	8.15	2.78	8.05	11.3	45.5
8	4.29	0.382	8.09	2.36	6.87	9.55	47.4
9	4.11	0.378	8.01	2.22	6.34	10.5	48.0
10	4.03	0.376	7.96	1.9	7.27	12.2	46.5
11	4.07	0.374	7.93	1.69	9.00	13.9	45.5
12	4.30	0.374	7.93	1.75	9.79	15.9	44.3
13	4.12	0.412	8.72	1.41	8.52	15.0	52.5
14	4.01	0.413	8.75	1.43	8.45	14.8	52.1
15	3.94	0.417	8.83	1.48	7.56	13.9	54.4
16	4.00	0.428	9.06	1.54	7.15	13.2	57.7
17	4.19	0.441	9.34	1.75	7.03	13.9	60.3
18	4.43	0.46	9.75	1.83	6.98	13.3	65.7
19	4.55	0.481	10.2	1.98	7.14	12.9	67.2
20	4.56	0.502	10.6	1.9	7.07	13.7	68.1
21	4.51	0.526	11.2	1.65	7.41	14.1	67.2
22	4.62	0.554	11.7	1.5	7.64	14.1	68.6
23	4.94	0.581	12.3	1.41	8.08	17.0	72.3
24	4.42	0.552	11.7	1.35	6.66	14.4	66.3
25	4.81	0.522	11.1	2.11	7.88	13.6	63.5
26	5.60	0.505	10.7	2.76	10.4	15.7	61.3
27	6.64	0.504	10.7	3.53	14.0	19.8	58.8
28	7.81	0.514	10.9	4.52	16.6	24.4	58.2
29	9.06	0.534	11.3	5.63	20.2	29.3	58.9
30	10.4	0.565	12.0	6.48	23.3	33.8	60.1
31	4.63	0.375	7.95	2.2	9.52	16.9	43.3
32	4.57	0.378	8.00	2.06	8.58	17.4	46.1
33	4.34	0.38	8.06	1.91	8.38	15.1	46.9
34	4.10	0.387	8.2	1.64	7.97	14.2	47.0
35	3.86	0.394	8.34	1.38	6.98	13.5	48.9
36	3.67	0.399	8.46	1.2	6.59	12.4	51.9
37	3.63	0.408	8.64	1.31	6.59	11.9	53.4
38	3.64	0.409	8.67	1.32	6.27	12.3	56.1
39	3.75	0.411	8.7	1.37	7.00	11.3	53.8
40	3.97	0.419	8.89	1.52	8.08	13.1	51.8
41	4.12	0.419	8.88	1.58	8.08	15.8	52.1

Appendix 13: Descriptive statistics for point-to-point distances (mm) between Trainer 1a and Trainer 2.

Appendix 13							
Point	Mean	Standard Error	Standard Deviation	Median	90%	95%	99%
0	4.45	0.742	5.20	2.45	11.9	16.3	21.0
1	3.94	0.648	4.53	2.03	10.8	13.9	18.1
2	3.55	0.557	3.90	1.78	9.63	11.9	15.1
3	3.18	0.487	3.41	1.76	8.69	9.93	12.3
4	2.89	0.436	3.05	1.57	7.79	9.13	11.1
5	2.63	0.39	2.73	1.66	6.83	9.72	10.3
6	2.40	0.385	2.70	1.34	5.94	9.01	10.3
7	2.60	0.416	2.91	1.45	5.83	9.79	11.7
8	2.48	0.411	2.88	1.24	5.31	9.24	11.5
9	2.44	0.467	3.27	1.26	6.43	8.98	14.5
10	2.39	0.533	3.73	1.35	6.21	8.43	17.6
11	2.33	0.585	4.09	1.15	4.79	9.07	19.5
12	2.41	0.64	4.48	1.15	3.06	10.8	21.4
13	1.34	0.159	1.12	1.02	2.92	3.35	4.76
14	1.18	0.15	1.05	0.886	2.56	2.93	4.40
15	1.13	0.148	1.03	0.725	2.35	3.10	4.41
16	1.14	0.174	1.22	0.757	2.11	3.66	5.41
17	1.35	0.208	1.46	0.79	2.36	4.20	6.64
18	1.61	0.255	1.79	1.09	3.23	4.40	7.96
19	1.84	0.305	2.13	1.21	3.87	5.18	9.58
20	1.95	0.337	2.36	1.19	4.52	5.01	10.8
21	1.85	0.357	2.50	0.958	4.47	4.78	10.9
22	1.90	0.383	2.68	1.11	3.82	4.81	12.0
23	1.94	0.414	2.9	1.18	3.43	4.55	12.9
24	1.87	0.34	2.38	1.14	3.19	4.65	11.1
25	1.99	0.312	2.18	1.04	4.27	5.84	9.67
26	2.31	0.338	2.36	1.33	5.17	6.92	9.80
27	2.67	0.42	2.94	1.57	5.99	8.96	12.3
28	3.17	0.517	3.62	1.95	7.29	11.2	14.7
29	3.76	0.622	4.35	2.18	8.87	13.5	17.2
30	4.38	0.735	5.15	2.59	11.9	15.7	19.7
31	2.47	0.703	4.92	0.878	4.17	11.6	23.5
32	2.4	0.7	4.90	0.902	3.59	10.8	23.8
33	1.95	0.582	4.07	0.581	2.88	10.0	19.2
34	1.72	0.478	3.35	0.591	3.34	7.55	16.1
35	1.51	0.388	2.72	0.688	2.55	6.27	13.0
36	1.31	0.313	2.19	0.632	2.55	4.80	10.7
37	1.15	0.257	1.80	0.635	2.19	3.95	8.88
38	1.17	0.206	1.44	0.749	1.94	3.48	7.23
39	1.18	0.169	1.18	0.820	1.93	3.75	5.66
40	1.25	0.168	1.18	0.819	2.80	3.23	5.52
41	1.30	0.176	1.23	0.946	2.38	3.29	5.97

Appendix 14: Descriptive statistics for point-to-point distances (mm) between Trainer 1b and Trainer 2.

Appendix 14							
Point	Mean	Standard Error	Standard Deviation	Median	90%	95%	99%
0	3.61	0.599	4.19	1.9	8.09	12.6	17.5
1	3.12	0.522	3.65	1.6	8.08	10.8	15.5
2	2.73	0.452	3.17	1.72	7.59	9.13	13.4
3	2.44	0.394	2.76	1.58	6.16	8.62	11.4
4	2.19	0.355	2.48	1.23	5.44	7.66	10.3
5	2.14	0.328	2.3	1.3	4.44	7.30	9.48
6	2.04	0.313	2.19	1.32	4.63	5.78	9.78
7	1.79	0.324	2.27	1.04	3.78	5.82	10.5
8	1.74	0.3	2.1	1.05	4.59	5.76	9.37
9	1.75	0.364	2.55	0.895	4.02	7.04	11.4
10	1.80	0.446	3.12	0.865	3.23	6.05	15.4
11	1.91	0.515	3.6	0.98	2.83	5.72	18.2
12	2.19	0.583	4.08	1.06	2.97	6.07	20.9
13	1.41	0.183	1.28	1.02	2.90	3.41	6.02
14	1.37	0.185	1.3	1.03	2.70	3.10	6.32
15	1.33	0.192	1.35	0.914	2.52	2.94	6.38
16	1.35	0.215	1.51	0.883	2.72	3.80	6.72
17	1.60	0.27	1.89	1.04	3.17	4.70	8.89
18	1.85	0.304	2.13	1.16	3.83	5.57	9.85
19	2.13	0.333	2.33	1.32	4.35	6.47	10.7
20	2.19	0.354	2.48	1.46	4.73	6.40	11.1
21	2.02	0.353	2.47	1.46	3.93	4.91	11.2
22	1.80	0.369	2.58	1.07	3.22	3.57	11.3
23	1.60	0.391	2.74	0.8	2.92	3.54	11.7
24	1.47	0.324	2.27	0.832	3.14	3.50	10.5
25	1.70	0.271	1.9	1.31	2.81	4.54	9.30
26	1.99	0.276	1.93	1.31	4.11	5.88	8.71
27	2.28	0.341	2.39	1.55	4.53	7.09	10.7
28	2.72	0.421	2.95	1.59	5.59	9.04	12.7
29	3.23	0.51	3.57	1.92	7.66	10.8	14.8
30	3.78	0.605	4.24	2.28	11.0	12.8	17.0
31	2.31	0.645	4.52	0.965	3.54	6.55	23.0
32	2.12	0.645	4.51	0.933	2.82	6.01	23.1
33	1.74	0.533	3.73	0.668	2.63	5.12	19.0
34	1.48	0.432	3.02	0.663	2.54	4.28	15.4
35	1.26	0.346	2.42	0.598	2.32	3.70	12.3
36	1.12	0.284	1.99	0.569	2.13	3.13	10.1
37	1.05	0.224	1.57	0.58	1.86	3.02	7.69
38	1.02	0.17	1.19	0.66	1.91	2.43	5.86
39	1.04	0.136	0.955	0.733	1.68	3.21	4.55
40	1.20	0.159	1.11	0.861	2.44	3.90	4.90
41	1.38	0.206	1.44	0.896	2.71	4.55	6.74

Appendix 15: Descriptive statistics for point-to-point distances (mm) between Trainer 1a and Trainer 1b.

Appendix 15							
Point	Mean	Standard Error	Standard Deviation	Median	90%	95%	99%
0	2.82	0.512	3.58	1.28	8.08	9.60	14.7
1	2.58	0.44	3.08	1.3	6.96	8.18	13.3
2	2.46	0.376	2.63	1.49	5.55	6.75	11.9
3	2.38	0.336	2.35	1.6	5.20	6.52	10.4
4	2.32	0.328	2.29	1.64	5.12	7.97	9.56
5	2.17	0.345	2.42	1.41	4.78	6.00	11.4
6	2.07	0.367	2.57	1.07	5.44	6.03	11.5
7	2.17	0.415	2.9	0.84	5.68	6.82	12.9
8	1.85	0.346	2.42	0.842	4.47	5.61	10.8
9	1.56	0.285	2	0.817	4.35	5.57	8.59
10	1.34	0.244	1.71	0.682	3.29	5.59	7.23
11	1.20	0.227	1.59	0.657	2.65	4.53	7.12
12	1.24	0.235	1.65	0.91	2.50	3.11	8.12
13	0.831	0.116	0.811	0.6	1.50	1.89	3.80
14	0.708	0.103	0.718	0.546	1.04	1.83	3.50
15	0.674	0.0916	0.641	0.553	1.10	1.72	3.19
16	0.777	0.0922	0.645	0.608	1.53	2.16	2.97
17	1.05	0.129	0.904	0.799	2.12	2.75	4.06
18	1.30	0.191	1.34	0.838	2.67	4,00	5.95
19	1.50	0.21	1.47	1.08	2.73	3.90	7.23
20	1.51	0.188	1.31	1.23	2.89	3.37	6.30
21	1.48	0.165	1.15	1.16	3.23	3.69	4.54
22	1.37	0.16	1.12	1.1	3.04	3.50	4.54
23	1.26	0.171	1.2	0.804	3.16	3.38	4.77
24	1.15	0.155	1.08	0.819	2.99	3.22	4.46
25	1.23	0.169	1.18	0.813	2.64	3.18	4.82
26	1.44	0.206	1.44	0.859	3.31	3.67	6.01
27	1.72	0.26	1.82	1.08	4.38	4.91	7.69
28	2.05	0.32	2.24	1.32	5.47	6.51	9.37
29	2.41	0.383	2.68	1.62	6.55	8.11	11.1
30	2.81	0.446	3.12	1.92	7.64	9.72	12.7
31	1.14	0.24	1.68	0.561	2.52	3.48	7.79
32	1.28	0.261	1.83	0.704	2.59	3.69	9.06
33	1.02	0.222	1.56	0.427	2.15	2.92	7.69
34	0.947	0.214	1.5	0.439	2.06	2.62	7.47
35	0.911	0.248	1.73	0.449	1.35	2.23	8.47
36	0.804	0.246	1.72	0.388	1.20	2.04	8.01
37	0.642	0.179	1.26	0.362	0.88	1.63	5.80
38	0.565	0.153	1.07	0.259	1.05	1.56	4.63
39	0.654	0.118	0.827	0.391	1.27	1.79	3.91
40	0.717	0.105	0.733	0.475	1.54	2.26	3.28
41	0.794	0.105	0.737	0.572	1.52	1.94	3.27