

ADVANCED IMAGING IN ANKLE FRACTURES



**JASPER
PRIJS**

ADVANCED IMAGING IN ANKLE FRACTURES

Jasper Prijs

Colofon

Copyright 2025 © Jasper Prijs

All rights reserved. No parts of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or by any means without permission of the author.

Provided by thesis specialist Ridderprint, ridderprint.nl

Printing: Ridderprint

Layout and design: Indah Hijmans, persoonlijkproefschrift.nl

Cover design: Jasper Prijs



university of
 groningen



Flinders
University

Advanced Imaging in Ankle Fractures

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. J.M.A. Scherpen
and in accordance with
the decision by the College of Deans

and

to obtain the degree of PhD at
Flinders University, Australia
on the authority of the
Chancellor of Flinders University
Mr. John Hood

This thesis will be defended in public on
Monday 10 November 2025 at 12.45 hours

by

Jasper Prijs

born on 17 April 1996

Promotores

Prof. P.C. Jutte

Prof. R.L. Jaarsma

Prof. J.N. Doornberg

Co-promotores

Dr. F.F.A. IJpma

Assessment Committee

Prof. J.P.P.M. de Vries

Prof. M. Maas

Prof. P.M.A. van Ooijen

A/Prof. T. White

TABLE OF CONTENTS

General Introduction	7
Chapter 1 Artificial Intelligence and Computer Vision in Orthopaedic Trauma: the Why, How, and What	13
PART I Exploring Artificial Intelligence in Ankle Fractures	
Chapter 2 An Increasing Number of Convolutional Neural Networks for Fracture Recognition and Classification in Orthopaedics: are these Externally Validated and Ready for Clinical Application?— <i>a Systematic Review</i>	25
Chapter 3 Artificial Intelligence Fracture Recognition on Computed Tomography: Review of Literature and Recommendations— <i>a Systematic Review</i>	39
Chapter 4 Development and External Validation of Automated Detection, Classification, and Localization of Ankle Fractures: Inside the Black Box of a Convolutional Neural Network (CNN)	57
Chapter 5 External Validation of an Artificial Intelligence Multilabel Deep Learning Model Capable of Ankle Fracture Classification	77
Chapter 6 Artificial Intelligence and 3D-Guided Surgery in Orthopaedic Trauma: Why, How, and What— <i>a book chapter in Rockwood and Green</i>	101
PART II Advanced Imaging in Paediatric Ankle Fracture Trauma	
Chapter 7 Triplane Ankle Fracture Patterns in Paediatric Patients: Extent of Physeal Closure Does Not Dictate Pathoanatomy	125
Chapter 8 Understanding the Mechanism of Injury and Fracture Pattern of Paediatric Triplane Ankle Fractures versus Adult Trimalleolar Fractures	143
PART III Prospective Analysis of Ankle Fractures	
Chapter 9 Subtle Factors Associated with Outcome of Ankle Fractures: Do Not Forget Intra-Articular Loose Bodies	159
General Discussion	178
English Summary	184
Nederlandse Samenvatting	187
Appendix	193

Great Barrier Reef, Cairns, Queensland, Australia



GENERAL INTRODUCTION

GENERAL INTRODUCTION

The ankle joint consists of three bones: the tibia, which bears most of the weight; the fibula, which forms the lateral stability of the ankle fork; and the talus, which allows flexion and extension of the ankle and bears the weight transferred from the tibia to the foot. Most of the stability of the ankle is provided by its ligaments. Notable for this thesis are mainly the Anterior and Posterior Inferior Tibiofibular Ligaments (AITFL & PITFL)—that form the syndesmosis—and the medial deltoid ligament. The syndesmosis provides lateral stability, whereas the medial deltoid supports medial stability. The ankle is crucial for our mobility and vulnerable to injuries as it bears forces many times our body weight during regular walking and up to thirteen times during running.¹ Due to these factors, ankle fractures are one of the most common fractures presenting in the Emergency Department (ED).

Diagnostics for ankle fractures rely on imaging, conventionally with radiographs, and increasingly combined with Computed Tomography (CT) scans. The availability of large datasets of radiographs and CT scans allows for advanced imaging analyses using Artificial Intelligence (AI) applications and 3D technology. We hypothesized that advancements in imaging modalities can improve diagnostics, understanding, and treatment of ankle fractures. Hence, this thesis explores ankle fractures in adults and children, with the use of advanced imaging such as Convolutional Neural Networks (CNNs), fracture maps, and 3D reconstructions.

PART I – Exploring Artificial Intelligence in Ankle Fractures

Artificial Intelligence has surged and is increasingly affecting our daily lives, despite 'doom and gloom' by many. Despite its challenges, and since the start of this thesis, the field has seen incredible development. It began in 1946 with the pioneering work of Alan Turing, the godfather of computers, with the creation of the 'Turing machine'. A thought experiment that originated computers as we now know them. Without his efforts, many of the studies in this thesis would not have been possible. In 1983, Apple Inc. introduced the first computer with a graphical user interface (GUI)—versus a terminal with just lines of text— which made it possible to operate computers as we are doing today. GUIs have become the default way to communicate with computers. As part of this thesis, I moved to Adelaide, Australia, twice, in total for two years. Despite the innovative character of this thesis, when I first arrived in Adelaide, I encountered one of the last computers that was still operated using a terminal instead of a GUI in the Emergency Department of Flinders Medical Centre. Along with this thesis, that computer has advanced with technology to include a graphical interface and electronic patient filesystem. I believe that Artificial Intelligence will provide a similar advancement to Medicine, as the computer from Flinders Emergency Department brought to the ED workflow during the timeframe of this thesis.

Artificial Intelligence, and 'Computer Vision' enable computers to assess images and provide predictions about objects in them. Currently, Convolutional Neural Networks (CNNs)

can perform at the level of experienced surgeons and may in the future even surpass them. CNNs are designed to mirror a human neural network—our brain—and work in a similar fashion, where it is trained to recognize certain features, akin to our own learning. In **Chapter 1**, we explore the possibilities and drawbacks of this advanced technology in trauma. Part I expands on this introduction and provides a detailed overview and original research about Artificial Intelligence in Orthopaedic Trauma. Developing a Machine Learning (ML) model and testing it on your own data often results in good results; however, ideally, the model needs to perform regardless of its environment—a process known as external validation. In **Chapter 2**, we explore not only how many of the studies developing a machine learning model were externally validated, but also in which way they did this. As the field advances, models are increasingly developed to analyze Computed Tomography (CT) scans. As CT scans are ubiquitous in complex trauma, we evaluate the accuracy and benefits of ML models for this more advanced imaging modality in **Chapter 3**. For **Chapter 4**, we developed a CNN that detects, classifies, and segments lateral malleolar fractures. Apart from localizing fractures to enhance diagnostics and optimal treatment, classification systems are plagued by low inter-observer reliability. In an international collaboration with colleagues from Stockholm, Sweden, we externally validated a CNN that classifies ankle fractures according to the AO system² in **Chapter 5**. This first part is concluded by **Chapter 6**, a book chapter in Rockwood and Green Fractures in adults, which summarizes recent developments and describes future perspectives.

PART II – Advanced Imaging in Paediatric Ankle Fractures

Paediatric trauma is often different from adult trauma, due to the presence of unique anatomy in the developing skeleton. In the ankle specifically, differences are the relative size of ligaments and the presence of growth plates (physes). Injuries involving the physis are referred to as 'transitional' fractures and occur mostly between the ages of 12 and 15, as this is the period where the physis transitions from open to closed. Tillaux and triplane fractures are transitional injuries involving the ankle. Triplane fractures are complex three-dimensional injuries that extend into three planes, namely the sagittal (epiphysis), axial (physis), and coronal (metaphysis). Triplane fractures occur in children where closure of the physis has started, and the fracture pattern is hypothesized to be determined by the extent of closure. This classic teaching that fracture pattern is based on age and respective extent of physeal closure is based on studies from the 1960s. Part II challenges these assumptions regarding pathophysiology and fracture pattern of paediatric ankle fractures by using advanced imaging techniques. **Chapter 7** re-evaluates the theory from the 1960s that the physis is the primary determinant of the fracture pattern of triplane fractures. Moreover, much is still unknown about the exact trauma mechanism and optimal reduction strategies for triplane fractures. **Chapter 8** explores the similarities and differences between the adult trimalleolar and paediatric triplane fractures to facilitate improved understanding and treatment of these complex injuries.

PART III – Prospective Analysis of Ankle Fractures

Compared to PART II, where we explored the trauma mechanisms of injuries that are scarcely studied, trauma mechanisms of adult ankle fractures are well described. Prospective collection of data is crucial to understand which factors are associated with clinical outcome after ankle fractures. The challenge in adult ankle trauma is described in the following paradox: despite the great experience of surgeons in handling these high-volume injuries, the outcomes following operative fixation remain surprisingly suboptimal. While surgeons exhibit confidence and familiarity with treating these common fractures, patients often grapple with short-term and long-term consequences such as diminished mobility, chronic pain, and early posttraumatic osteoarthritis. It is still not fully elucidated why specific characteristics, such as type II Haraguchi posterior malleolar fractures, fracture severity, trimalleolar involvement, and Body Mass Index (BMI) are touted to be significant contributors to diminished clinical outcomes. These factors alone have not provided a comprehensive explanation for the variability in outcomes and treatment for these fractures. Consequently, the reliance on surgeons' experience persists, highlighting a gap where evidence-based decision-making could play a pivotal role. This prompts a closer examination of potentially 'subtle' surgical and pathoanatomic factors—such as the quality of postoperative reduction, presence of loose bodies, or malpositioned osteosynthesis materials. Understanding these prognostic factors that influence outcomes is paramount for effectively informing patients post-trauma and refining surgical decision-making processes. We explored the effect of these factors on post-operative outcomes in **Chapter 9**.

REFERENCES

1. Burdett RG. Forces predicted at the ankle during running. *Med Sci Sports Exerc.* 1982;14(4):308-16. doi: 10.1249/00005768-198204000-00010. PMID: 7132650.
2. Meinberg EG, Agel J, Roberts CS, Karam MD, Kellam JF. Fracture and Dislocation Classification Compendium-2018. *J Orthop Trauma.* 2018 Jan;32 Suppl 1:S1-S170. doi: 10.1097/BOT.0000000000001063. PMID: 29256945.



Artificial Intelligence and Computer Vision in Orthopaedic Trauma: the why, how, and what

J. Prijs

Z. Liao

S. Ashkani-Esfahani

J. Olczak

M. Gordon

P. Jayakumar

P.C. Jutte

R.L. Jaarsma

F.F.A. IJpma

J.N. Doornberg

on behalf of the Machine Learning Consortium

ABSTRACT

Artificial intelligence (AI) is, in essence, the concept of 'computer thinking', encompassing methods that train computers to perform and learn from executing certain tasks, called machine learning, and methods to build intricate computer models that both learn and adapt, called convolutional neural networks. Computer vision is a function of AI by which machine learning and convolutional neural networks can be applied to enable computers to capture, analyze, and interpret information from clinical images and visual inputs. This annotation summarizes key considerations and future perspectives concerning computer vision, questioning the need for this technology (the 'why'), the current applications (the 'what'), and the approach to unlocking its full potential (the 'how').



WHY AI AND COMPUTER VISION?

In orthopaedic surgery, we have been shown time and time again that “surgeons agree mostly with themselves, but not so much with each other”.¹⁻⁸ Daniel Kahneman coined this form of human bias “WYSIATI”: What You See Is All There Is.⁹ Our field is rife with unsatisfactory levels of interobserver reliability in the recognition and classification of fractures among surgeons. The issue of reliability covers trauma from injuries involving the upper^{2-4,7} and lower limb.^{5,6,8} Advances in the power of hardware and computing, the development of more accurate imaging techniques, and improvements in the capabilities of software by using computer vision, promise to increase the speed and accuracy of diagnosis and overcome concerns about reliability for the evaluation of images in trauma.^{10,11} The widely used convolutional neural networks have several characteristic features and merits. Compared with conventional machine-learning methods such as decision tree, random forest, boosting, and support vector machines, which are typically used to solve problems in machine-learning on top of structured data, the convolutional filtering operations in a convolutional neural network can respond to local patterns in features of input which are spatially and temporally correlated. These consume fewer computational resources compared with a matrix multiplication process and hence are predominantly used in the processing of images and videos. Convolutional neural networks also include a type of deep neural network in which the structure of the model can be easily modified by adding or removing a layer. There is a rich literature involving the structure of convolutional neural networks in the computer vision domain, providing a good methodological basis for the analysis of medical images. Finally, the number of learnable parameters in a convolutional neural network is at a scale of millions to billions, and the optimization of the parameters of models is often favourable when dealing with a massive amount of data.

Rather than a replacement for human interpretation, we believe that the attraction of computer vision in the practice of trauma surgery lies in augmenting the diagnostic capabilities of surgeons and musculoskeletal radiologists, reducing bias and variation, minimizing error and mismanagement, and ultimately buying time to focus on our patients and delivering optimal care.^{10,12,13}

HOW DOES COMPUTER VISION WORK?

AI algorithms are now incorporated into many digital products, from smartphones to automated vehicles. The data generated through use of these devices serve as a perpetual source of information for further computer learning and improvement. In orthopaedic surgery, AI is being used in the development of advanced models of prediction as well as automated methods for the diagnosis and classification of different conditions. Models which predict the stratification of risk using machine learning now go beyond conventional statistics identifying

non-linear relationships between individual characteristics and outcomes.^{14,15} For instance, models have been used to predict same-day discharge and assess balance and prosthetic alignment during total knee arthroplasty.^{16,17} Computer vision has been evaluated in the detection and classification of fractures using radiographs and CT scans.^{18,19} In other specialties, clinicians are using this technology to interpret images such as mammograms, fundoscopies for papilloedema, and CT scans for the identification of intracerebral haemorrhage.^{20–22} There has been a considerable increase in the number of studies aiming to improve clinical decision-making through the analysis of large databases using AI and computer vision.^{18,19,23} The next phase should focus on prospective clinical evaluation, the maturation of techniques, and expansion of work to gain external validity in geographical areas and populations, in order to consolidate accuracy, reliability, and transferability while minimizing bias.¹⁹ Kunze et al²⁴ and others have emphasized these factors and the need for improvement in the regulations and standards for taxonomy, the quality of data, critical appraisal, and reporting.^{25–28}

WHAT ARE WE DOING WITH AI AND COMPUTER VISION?

Appreciating the fundamental differences in 'learning'—the process of absorbing information to increase knowledge, skills, and capabilities, and applying this intelligence across a variety of different contexts—between humans and AI-powered machines can help us improve our understanding of the technology behind computer vision. Humans use the brain's computational power, memory, and innate ability to learn from direct experience or to be trained by others. We are also taught to explain how and why we came to certain conclusions about the things we have learned and interpret, and write out the mathematics (or 'logic') so that it can be understood and validated by others. In contrast, machines driven by AI rely on the provision of data and the respective outcomes into the system to build current and future logic, and understand how outcomes might be inferred. A trained machine-learning model is highly convolutional, encapsulating millions of numerical parameters that collectively contribute to any decision it makes. Therefore, it is beyond our human capacity to fully explain why a model came to a certain conclusion, as the decision could be based on either a pattern that makes sense (clinically) or on a pattern with apparent association to the decision (i.e. A model may learn to recognize sheep by learning the texture of grass, as sheep are always found on grass).

Increasing the number of labels and observers is the most common way to deal with inadvertent human interobserver variation and mistakes. However, what we are teaching the computer is the majority-voted decision, which is usually the best available truth but unfortunately not error-free. If we want the computer to learn beyond what is given (i.e. Information based on our understanding such as of classifications) it needs to act with the task (environment) and trial-and-error actions, where the process is in many ways similar to the evolutionary process. For example, the alphago Zero chess player made by Google AI was



created by allowing AI players to play against each other.²⁹ This was different from the original alphago,³⁰ which learned from human moves. After a huge number of games, the AI players start to invent moves. As the computer can play so much quicker than a human, it may cover or surpass the entirety of games played throughout human history and thus generate a huge amount of data, which is key to an excellent model. In order to generate enormous datasets and create models that outperform us, it is essential that we collaborate, not only nationally but globally. However, it is also essential to consider ethical issues. For example, what if a dataset of 100,000 images is lost? Even though these images were anonymized, it would still lead to headlines and have an enormous effect on the future collection of these datasets. In addition to ethical considerations, laws between countries about sharing data between institutions, each with their own protocols and mandates, often significantly impair collaborations.

In computer vision-based analysis of orthopaedic images, the input can include any form of digital data, most often radiographs and CT scans. Medical images are usually stored in the Digital Imaging and Communications in Medicine (DICOM) format as this contains substantial, often unnecessary, and sometimes incorrect, information about the patients and the study, the data are converted into more generic formats such as Portable Network Graphics (PNG, lossless) or Joint Photographic Experts Group (JPEG, lossy compression) files to minimize redundancy and increase efficiency. These data and converted formats are then split into training and test sets in a 60:40 or 80:20 ratio. Within the training set, a separate set of images is selected or stochastically sampled, often using n-fold cross-validation, to develop the validation set. This is then used to optimize the performance of the training set without compromising the objectivity of the test set, which is then finally used to evaluate performance. In other words, one is not directly training the model to fit the test set as a strategy to avoid overfitting the model. Thus, the computer model can effectively perform the designated task, not only on the images it has seen before, but on the images it has yet to see. This characteristic is termed 'generalization'. The computer can reach human-level performance, or even outperform humans in certain tasks, but limitations in the ways of validating decisions can lower the reliability of medical AI systems, making the use of applied AI in medicine challenging.

PITFALLS AND WHAT TO LOOK OUT FOR WHEN APPRAISING MANUSCRIPTS DEALING WITH CONVOLUTIONAL NEURAL NETWORKS FOR FRACTURES

There is a healthy reservation or resistance towards using AI in diagnostics and medical decision-making, and anyone who has had AI take the wheel can attest that the deviation from the normal situation is challenging. However, as we gain more experience with the applications of AI, it will become easier to understand and navigate through these situations. Even though computers,

given the 'artificial' intelligence, might be able to perform certain tasks better than humans, they do not possess common sense and are therefore always 'stupid' or cold as robots. The main weakness of convolutional neural networks is the fact that their quality relies heavily on the database upon which they were trained. One cannot expect such a network to recognize fractures or pathology it has not seen before, even though they may be similar to what it already 'knows'. Therefore, the utmost care must be taken when choosing the data that are used for training, testing, and validation, either internally or externally. External validation is a crucial step in the validation of a model on new data from a different geographical location, as this could expose possible biases and performance weaknesses.¹⁹ Many AI models in orthopaedic surgery have not undergone external validation.¹⁹ however, assisting clinicians with AI-based solutions has some important strengths, including consistent predictions, no mental fatigue, no inherent bias, and analysis in just a few seconds. It can reach the level of an experienced clinician and is therefore able to provide continual top-level expertise effortlessly.³¹⁻³⁶

FUTURE PERSPECTIVES

Regardless of the challenges in the past, present, and future, there has been a rapid development of AI and a surge of practical applications in day-to-day life. We enjoy the use of voice assistance to turn on the lights, dictate a message, or as reminders. We believe the future of medicine will enjoy similar quality-of-life improvements, with significant effects on the lives of our patients. Would it not bring comfort to patients and doctors to be able to make informed decisions together, based on the patient's specific medical characteristics, and to focus on the patients who require close monitoring, and spend one's time where it is the most efficient? We do not believe that AI will replace doctors, but will instead reduce the burdens on us and allow us to spend our time more efficiently with our patients.

In order to achieve these goals, we need to overcome one of the most difficult challenges yet: the relative shortage of quality data in a single hospital. We need to rise above isolated models that are developed, tested, and applied clinically in one centre, and thus are not applicable elsewhere. Only together can we create large enough databases to predict the conditions that matter, such as patient-specific outcomes based on individual characteristics, the risks of postoperative infection, hardware failure, morbidities, and mortality.

TAKE HOME MESSAGE

Artificial intelligence has seen a surge of applications; however, only together can the orthopaedic community create large databases so we can train models that are globally applicable and with a greater ability to predict the conditions that matter.

REFERENCES

1. Becker SJE, Bruinsma WE, Guitton TG, et al. Interobserver agreement of the Eaton-Glickel classification for trapeziometacarpal and scaphotrapezial arthrosis. *J Hand Surg Am.* 2016;41(4):532-540.
2. Beks RB, Drikkoningen T, Claessen F, Guitton TG, Ring D, Variation G. Interobserver variability of the diagnosis of scaphoid proximal pole fractures. *J Wrist Surg.* 2018;7(4):350-354.
3. Bruinsma WE, Guitton TG, Warner JJP, Ring D, Science of Variation Group. Interobserver reliability of classification and characterization of proximal humeral fractures: a comparison of two and three-dimensional CT. *J Bone Joint Surg Am.* 2013;95-A(17):1600-1604.
4. Doornberg JN, Guitton TG, Ring D, Variation G. Diagnosis of elbow fracture patterns on radiographs: interobserver reliability and diagnostic accuracy. *Clin Orthop Relat Res.* 2013;471(4):1373-1378.
5. Mellema JJ, Doornberg JN, Molenaars RJ, Ring D, Kloen P, Traumatplatform Study Collaborative & Science of Variation Group. Tibial plateau fracture characteristics: reliability and diagnostic accuracy. *J Orthop Trauma.* 2016;30(5):e144-51.
6. Mellema JJ, Doornberg JN, Molenaars RJ, Ring D, Kloen P, Traumatplatform Study Collaborative & Science of Variation Group. Interobserver reliability of the Schatzker and Luo classification systems for tibial plateau fractures. *Injury.* 2016;47(4):944-949.
7. Doornberg J, Lindenhovius A, Kloen P, van Dijk CN, Zurakowski D, Ring D. Two and three-dimensional computed tomography for the classification and management of distal humeral fractures. Evaluation of reliability and diagnostic accuracy. *J Bone Joint Surg Am.* 2006;88-A(8):1795-1801.
8. Doornberg JN, Rademakers MV, van den Bekerom MP, et al. Twodimensional and three-dimensional computed tomography for the classification and characterisation of tibial plateau fractures. *Injury.* 2011;42(12):1416-1425.
9. Kahneman D. *Thinking, fast and slow.* 1st ed. New York, New York, USA: Farrar, Straus and Giroux, 2011.
10. Ring D. How Artificial Intelligence May Improve Compassion in Orthopaedic Surgery. *AAOS Now;* November 1, 2019. <https://www.aaos.org/aaosnow/2019/nov/commentary/commentary04/> (date last accessed 29 June 2022).
11. Ashkani-Esfahani S, Mojahed Yazdi R, Bhimani R, et al. Assessment of ankle fractures using deep learning algorithms and convolutional neural network. *Orthopedics.* Jul 282020.
12. Oosterhoff JHF, Doornberg JN, Machine Learning Consortium. Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of Gartner's hype cycle. *EFORT Open Rev.* 2020;5(10):593-603.
13. Ashkani-Esfahani S, Mojahed-Yazdi R. How artificial intelligence improves orthopaedic practice. *EC Orthop.* 2021;12(7):84-87.
14. Hendrickx LAM, Sobol GL, Langerhuizen DWG, et al. A machine learning algorithm to predict the probability of (occult) posterior malleolar fractures associated with tibial shaft fractures to guide "malleolus first" fixation. *J Orthop Trauma.* 2020;34(3):131-138.
15. Oosterhoff JHF, Karhade AV, Oberai T, Franco-Garcia E, Doornberg JN, Schwab JH. Prediction of postoperative delirium in geriatric hip fracture patients: a clinical prediction model using machine learning algorithms. *Geriatr Orthop Surg Rehabil.* 2021;12:21514593211062276.
16. Wei C, Quan T, Wang KY, et al. Artificial neural network prediction of sameday discharge following primary total knee arthroplasty based on preoperative and intraoperative variables. *Bone Joint J.* 2021;103-B(8):1358-1366.
17. Verstraete MA, Moore RE, Roche M, Conditt MA. The application of machine learning to balance a total knee arthroplasty. *Bone Jt Open.* 2020;1(6):236-244.



18. Langerhuizen DWG, Janssen SJ, Mallee WH, et al. What are the applications and limitations of artificial intelligence for fracture detection and classification in orthopaedic trauma imaging? A systematic review. *Clin Orthop Relat Res*. 2019;477(11):2482-2491.
19. Oliveira E Carmo L, van den Merkhof A, Olczak J, et al. An increasing number of convolutional neural networks for fracture recognition and classification in orthopaedics: are these externally validated and ready for clinical application? *Bone Jt Open*. 2021;2(10):879-885.
20. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet*. 2018;392(10162):2388-2396.
21. Milea D, Najjar RP, Zubo J, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med*. 2020;382(18):1687-1695.
22. Lotter W, Diab AR, Haslam B, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat Med*. 2021;27(2):244-249.
23. McDonnell JM, Evans SR, McCarthy L, et al. The diagnostic and prognostic value of artificial intelligence and artificial neural networks in spinal surgery: a narrative review. *Bone Joint J*. 2021;103-B(9):1442-1448.
24. Kunze KN, Orr M, Krebs V, Bhandari M, Piuze NS. Potential benefits, unintended consequences, and future roles of artificial intelligence in orthopaedic surgery research: a call to emphasize data quality and indications. *Bone Jt Open*. 2022;3(1):93-97.
25. Farrow L, Zhong M, Ashcroft GP, Anderson L, Meek RMD. Interpretation and reporting of predictive or diagnostic machine-learning research in trauma & orthopaedics. *Bone Joint J*. 2021;103-B(12):1754-1758.
26. Olczak J, Pavlopoulos J, Prijs J, et al. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. *Acta Orthop*. 2021;92(5):513-525.
27. Bayliss L, Jones LD. The role of artificial intelligence and machine learning in predicting orthopaedic outcomes. *Bone Joint J*. 2019;101-B(12):1476-1478.
28. Jones LD, Golan D, Hanna SA, Ramachandran M. Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern? *Bone Joint Res*. 2018;7(3):223-225.
29. Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. *Nature*. 2017;550(7676):354-359.
30. Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*. 2016;529(7587):484-489.
31. Yamada Y, Maki S, Kishida S, et al. Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: ensemble decision-making with antero-posterior and lateral radiographs. *Acta Orthop*. 2020;91(6):699-704.
32. Chung SW, Han SS, Lee JW, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop*. 2018;89(4):468-473.
33. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. *Skeletal Radiol*. 2019;48(2):239-244.
34. Mawatari T, Hayashida Y, Katsuragawa S, et al. The effect of deep convolutional neural networks on radiologists' performance in the detection of hip fractures on digital pelvic radiographs. *Eur J Radiol*. 2020;130:109188.
35. Blüthgen C, Becker AS, Vittoria de Martini I, Meier A, Martini K, Frauenfelder T. Detection and localization of distal radius fractures: Deep learning system versus radiologists. *Eur J Radiol*. 2020;126:108925.

36. Olczak J, Emilson F, Razavian A, Antonsson T, Stark A, Gordon M. Ankle fracture classification using deep learning: automating detailed AO Foundation/ Orthopedic Trauma Association (AO/OTA) 2018 malleolar fracture identification reaches a high degree of correct classification. *Acta Orthop.* 2021;92(1):102-108.



PART I

**Exploring Artificial Intelligence
in Ankle Fractures**

Kernot Range Freecamp, Curtin Springs, Northern Territory, Australia



**An Increasing Number of Convolutional
Neural Networks for Fracture Recognition
and Classification in Orthopaedics**

are these externally validated and ready for clinical application?

L. Oliveira e Carmo

A. van den Merkhof

J. Olczak

M. Gordon

P. C. Jutte

R. L. Jaarsma

F. F. A. IJpma

J. N. Doornberg

J. Prijs

Machine Learning Consortium

ABSTRACT

Aims

The number of convolutional neural networks (CNN) available for fracture detection and classification is rapidly increasing. External validation of a CNN on a temporally separate (separated by time) or geographically separate (separated by location) dataset is crucial to assess generalizability of the CNN before application to clinical practice in other institutions. We aimed to answer the following questions: are current CNNs for fracture recognition externally valid?; which methods are applied for external validation (EV)?; and, what are reported performances of the EV sets compared to the internal validation (IV) sets of these CNNs?

Methods

The PubMed and Embase databases were systematically searched from January 2010 to October 2020 according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement. The type of EV, characteristics of the external dataset, and diagnostic performance characteristics on the IV and EV datasets were collected and compared. Quality assessment was conducted using a seven-item checklist based on a modified Methodologic Index for Non-Randomized Studies instrument (MINORS).

Results

Out of 1,349 studies, 36 reported development of a CNN for fracture detection and/or classification. Of these, only four (11%) reported a form of EV. One study used temporal EV, one conducted both temporal and geographical EV, and two used geographical EV. When comparing the CNN's performance on the IV set versus the EV set, the following were found: AUCs of 0.967 (IV) versus 0.975 (EV), 0.976 (IV) versus 0.985 to 0.992 (EV), 0.93 to 0.96 (IV) versus 0.80 to 0.89 (EV), and F1-scores of 0.856 to 0.863 (IV) versus 0.757 to 0.840 (EV).

Conclusion

The number of externally validated CNNs in orthopaedic trauma for fracture recognition is still scarce. This greatly limits the potential for transfer of these CNNs from the developing institute to another hospital to achieve similar diagnostic performance. We recommend the use of geographical EV and statements such as the Consolidated Standards of Reporting Trials–Artificial Intelligence (CONSORT-AI), the Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence (SPIRIT-AI) and the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis–Machine Learning (TRIPOD-ML) to critically appraise performance of CNNs and improve methodological rigor, quality of future models, and facilitate eventual implementation in clinical practice.

INTRODUCTION

An increase in the use of artificial intelligence (AI), particularly convolutional neural networks (CNNs, which mimic human visual cortex neurons), has been observed in medical imaging.¹⁻⁴ CNNs are able to process enormous volumes of data that surpass the pace of human observations, and in the field of orthopaedic trauma, CNNs have been reported to perform at the level of experienced orthopaedic surgeons and radiologists in detection and classification of distal radius, hip, proximal humerus, pelvis, and femur fractures.⁵⁻¹¹

Performance of CNNs is evaluated using unseen data from the same initial longitudinal dataset used for training the CNN, called the test set or internal validation (IV) set. However, characteristics of these data are identical (i.e. same hospital and time period) to those used for model development. Algorithms generally perform poorly when externally validated with datasets from different institutions.¹²⁻¹⁵ For example, in automated recognition of distal radius fractures, Blüthgen et al⁶ reported decreased performance using the external validation (EV) set, while performance was excellent using the IV set. To explore weaknesses and generalizability of CNNs, two techniques can be used: geographical (separated by location) or temporal (separated by time) validation (Figure 1).¹⁶ Arguably only the former truly represents EV that allows transfer of locally trained CNNs to applications in other hospitals.¹⁷ Hence, geographical EV is considered the most stringent test of a model's performance and an important step before clinical implementation.¹⁷⁻¹⁹

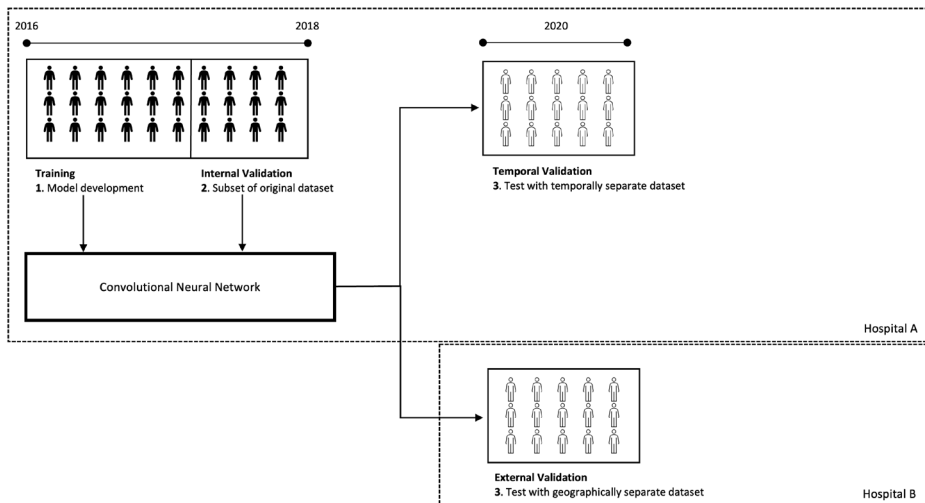


Figure 1. Training, Internal Validation, Temporal Validation and External Validation Datasets

Therefore, we aimed to answer the following: are CNNs for fracture recognition externally valid?; what are current methods applied for EV of CNNs for fracture recognition in the field of orthopaedic trauma?; and what are the reported performances of EV compared to the IV? To our knowledge, this is the first study to evaluate current applications of EV of CNNs used in orthopaedics for fracture detection and/or classification.

METHODS

A literature search according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement²⁰ (Figure 2) was conducted in the PubMed and EMBASE libraries for articles published between January 2010 and October 2020. The protocol was registered on PROSPERO (CRD42020216478) prior to screening the articles. Together with a medical librarian a search strategy was formulated (Supplementary Material).

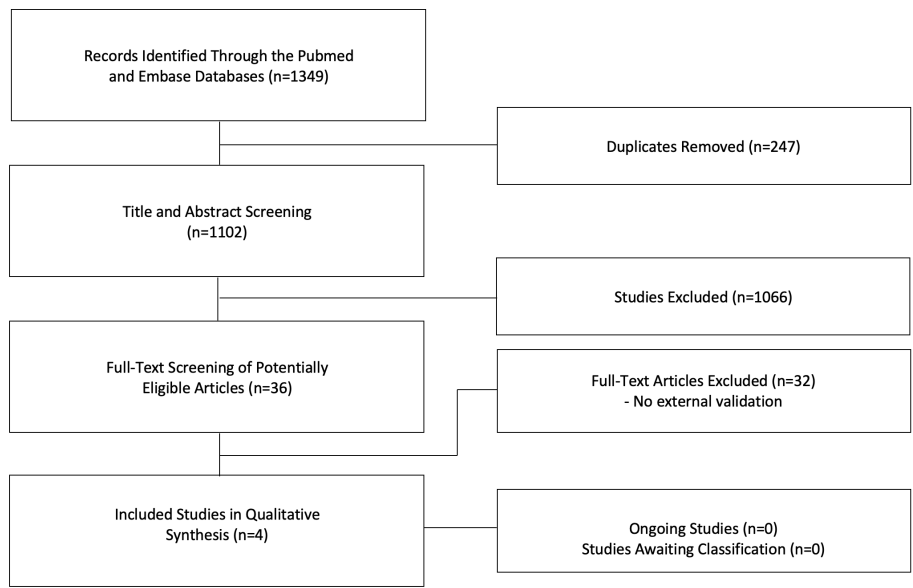


Figure II. PRISMA Flowchart depicting the study selection during screening and inclusion of articles for a search period from 2010 to October 2020

Two reviewers (LOEC, AVDM) independently screened the titles and abstracts of the retrieved articles. They subsequently performed the full-text screening to check eligibility of articles with predetermined inclusion criteria. Disagreements between reviewers were solved by consulting a third reviewer (JP). Due to ambiguous and unclear reporting of 'external validation' in articles found during the preliminary searches, all articles that reported the use of a CNN in

orthopaedic trauma were selected for full-text review. From the full-text review, only articles that evaluated their CNN on a separate dataset—geographically or temporally—from that used during the CNN development (the “external validation”) were included.

Inclusion criteria were journal articles reporting the use of a CNN in orthopaedic trauma including a form of EV, studies published after 2010, and written in English, Dutch, French, Portuguese, or Spanish. Exclusion criteria were the use of a CNN outside of an orthopaedic trauma setting, studies evaluating robot-assisted surgery techniques, studies with mixed cohort without clear subgroup reporting, review articles, letters to the editor, meeting abstracts, technique papers, and animal and cadaveric studies.

The search strategy yielded a total of 1,349 articles. After removal of duplicates, a total of 1,102 articles were screened. Overall, 36 studies reported the use of a CNN for fracture detection and/or classification and were selected for full-text review. Of these, four studies reported a form of EV (Figure 2). Additionally, no new studies were identified after manually screening the reference lists of included studies.

Quality assessment was performed by two independent reviewers (LOEC, AVDM). Disagreement was solved through discussion with a third reviewer (JP). The Quality Assessment of Diagnostic Accuracy Studies (QUADAS) criteria, a tool designed for the assessment of published diagnostic studies in systematic reviews, was not used because it was previously considered difficult to apply in machine-learning studies.^{21,22} Due to lack of suitable tools assessing risk of bias for machine learning studies, we modified the Methodologic Index for Non-Randomized Studies (MINORS) instrument, commonly used to assess the quality of cohorts or case-control studies.²³ The modified MINORS included the following items: disclosure, study aim, input features, ground truth, EV method, EV dataset, and performance metric. Screening and full-text review were conducted using Covidence (Veritas Health Innovation, Australia). Standardized forms were used to extract and record data (Excel v. 16.21; Microsoft, USA).

Outcome measures

To answer the primary research question, EV was defined as verification of model performance on a separate dataset, geographically or temporally, from that used for model development. To answer the secondary research question, type and characteristics of the EV set (dataset used, number of images, location and date of collection) were collected from the included articles. To answer the tertiary research question, performances of the CNN on the IV and EV datasets were collected and compared. All four studies were used to answer both secondary and tertiary research questions.

The following items were collected from all included studies: authors, year of publication, input feature (e.g. radiographs), radiological views if applicable (e.g. anteroposterior (AP)), anatomical location, output classes, ground truth label assignment, CNN model used, size, source and date of the initial dataset used for development, performance on IV set (e.g. area



under the curve (AUC)), method of EV (temporal or geographical), size, source and date of EV set, and performance on the EV set (e.g. AUC).

Three studies^{6,24,25} reported the area under the receiving operating characteristics curve (AUC-ROC) to evaluate IV and EV performance. The AUC is a common metric to report CNN performance,²⁶ where a value of 1.0 indicates perfect discriminatory performance, whereas 0.5 indicates a prediction equal to that of chance. One study used F1-score to evaluate model performance.²⁷ The F1-score (scored between 0 and 1) is a harmonic mean of precision (positive predictive value) and recall (sensitivity), where it requires both to be high for the F1-score to be high.

EV dataset characteristics and CNN features

All studies addressed AI models for fracture detection. In addition, one also used localization of fractures on images.⁶ Zhou et al²⁷ addressed both fracture detection and classification. The CNNs detected fractures on a single anatomical location like the wrist,^{6,24} elbow,²⁵ or ribs.²⁷ Input features of three studies^{6,24,25} were conventional radiographs; one study used CT scans.²⁷ All four studies reported the use of IV, with sets ranging from 98 CT scans²⁷ to 3,500 radiographs.²⁴

Quality appraisal

All studies reported disclosure. Study aim was clearly stated in all included studies, thereby reducing the possibility of outcome bias. All four studies clearly described the size, time, and location of collection of the EV dataset used, how the performance of the AI model was determined, and the ground truth (the reference standards used in AI). Out of the four studies included, three studies clearly stated the EV method used.^{6,24,25} One study used external data to improve model robustness and generalizability, however this was done before internally validating the model performance on the test set.²⁷ The inclusion and exclusion criteria for input features were clearly described in three studies.^{6,25,27} However, for one of the studies it was unclear which eligibility criteria were used for included radiographs.²⁴

Statistical analysis

Performance metrics used in each study were described, as well as its values for fracture detection and classification tasks. The values were given for both IV and EV set whenever available. Descriptive statistics such as size of the EV, training, and IV set were reported.

Table I. Method of External Validation, Characteristics of Datasets and Performance

Author, Year	Anatomical Location	AI Models Used	Input Feature and Imaging Direction	Output Classes	Ground Truth Label Assignment	Performance Metric
Lindsey et al.(23), 2018	Wrist *	CNN	Xray ; AP, lat	2	1 or 2 orthopedic surgeons	AUC
Choi et al.(7), 2020	Elbow / Distal humerus	CNN	X-ray; AP, lat	2	2 pediatric radiologists	AUC
Bluthgen et al.(5), 2020	Wrist / Distal radius	DLS†	Xray; AP, lat, combined	2	2 radiology residents, reports, available CTs	AUC
Zhou et al.(46), 2020	Ribs	CNN	CT; (NA)	3	2 musculo-skeletal radiologists, 2 senior radiologists, thoracic surgeon	F1 score

Table I. Method of External Validation, Characteristics of Datasets and Performance (cont.)

Author, Year	Performance of Internal Validation	Performance of External Validation	Size Training Set	Size Internal Validation Set	Type of External Validation	Size External Validation Set
Lindsey et al.(23), 2018	0.967	0.975	31490	Set 1: 3500 Set 2: 1400	Temporal	1400
Choi et al.(7), 2020	0.976	Temporal: 0.985; Geographical: 0.992	1012	Not Performed	Temporal and Geographical	Temporal: 258 Geographical: 95
Bluthgen et al.(5), 2020	Model 1: 0.93 Model 2: 0.96	Model 1: 0.80 Model 2: 0.89	524	100	Geographical	100
Zhou et al.(46), 2020	0.863 0.856	0.840 0.811 0.757	876	30	Geographical	173



RESULTS

To answer the primary research question, which CNNs for fracture recognition are externally valid and thus available for transfer from the developer to another hospital: four of 36 (11%) studies to date reported the use of EV (Table I).

To answer the second research question (which methods of EV for fracture recognition CNNs are currently used in the field of orthopaedic trauma), the following methodologies were identified (Table I).

CNNs deployed by Lindsey et al²⁴ were trained and internally validated on 31,490 and 3,500 respective radiographs between September 2000 and March 2016, and temporal EV performed with 1,400 radiographs from July to September 2016 from the same hospital. No geographical EV was applied.

Choi et al²⁵ conducted both temporal and geographical EV and used 258 patients for their temporal EV, which were collected between January and December 2018, and 95 patients collected at another hospital for their geographical EV. The CNN was trained and internally validated on 1,012 and 257 radiographs from their institution collected between January 2013 and December 2017.

The final two studies used geographical EV exclusively. Zhou et al²⁷ reported the use of a total of 75 patients for the geographical EV from three different respective hospitals with the original model trained and internally validated on 876 and 98 patients respectively, while Blüthgen et al⁶ randomly selected 100 patients from the MURA dataset²⁸ with the index CNN trained and internally validated on 166 and 42 patients from the authors' local institution.

Performance of CNN on EV compared to test set

To answer the third study question on performance of CNNs for fracture recognition on test set versus EV, this systematic review yielded four studies.

Comparing the CNNs' performance on the IV versus EV set, the following values are found: AUC of 0.967 vs 0.975 for distal radius fracture recognition,²⁴ AUC of 0.976 versus 0.985 (temporal) and 0.992 (geographical) for paediatric supracondylar fracture recognition,²⁵ AUC of 0.93 to 0.96 versus 0.80 to 0.89 for recognition of distal radius fractures,⁶ as well as an F1-score of 0.856 to 0.863 versus 0.757 to 0.840 for rib fracture recognition and classification on thoracic CT scans.²⁷

Lindsey et al²⁴ reported slightly improved performance (AUC 0.967 vs 0.975) upon temporal EV. Choi et al²⁵ reported an increase of the AUC when geographically externally validated, a decrease of 10% accuracy detecting normal elbows, and an increase of 5% accuracy in detecting fractures, whereas the temporal EV set accuracy performed similarly to the IV set. Blüthgen et al⁶ report a decrease in performance, for which the decrease in AP view of the distal radius is statistically significant ($p = 0.008$ to 0.021); however, calculating p-values in comparing AUCs has limited value. In Zhou et al,²⁷ a decreased F1-value is reported for the geographical EV sets.

DISCUSSION

There has been a significant increase in the use of CNNs in the field of orthopaedics the past few years.^{1–11} Papers tout promising results, however careful evaluation of performance and clinical utility of CNNs is warranted. EV is one of the crucial steps to secure generalizability of CNNs developed to detect pathoanatomy,^{19,29–31} prior to implementation into clinical practice. As many studies in our field now claim to have developed CNNs that perform at least on par with radiologists and orthopaedic surgeons,^{6,8,9,24,25} we aimed to review if these CNNs for fracture recognition are indeed externally valid and thus ready for clinical application; and secondly which methods of EV were used. Just four out of 36 full-text reviewed studies report any form of EV, and three applied and tested their algorithm to a geographically different dataset. None of the current CNNs have been prospectively tested in clinical practice.

This study has several strengths and weaknesses: first, an appropriate risk of bias assessment tool currently does not exist for studies reporting the use of a CNN, therefore we modified the MINORS tool. Second, although a broad search strategy encompassing two large databases was used, potentially relevant publications or algorithms developed for commercial purposes might have been missed. Third, comparability of the diagnostic performance characteristics between studies is limited as studies developed CNNs recognizing different types of fractures, however this factor did not affect answering our research questions.

Although EV of CNNs for fracture recognition is scarce in orthopaedic trauma, authors of four included studies did stress the importance of EV.^{6,24,25,27} They discussed the use of EV in evaluating CNNs, to discover generalizability and real-world performance. Indeed, EV evaluates the performance of CNNs in a different clinical context, a crucial step prior to implementation in clinical practice.¹⁸ In other fields of medicine, this step is believed to be paramount before translation to clinical practice.³⁰ EV is considered the sequel to IV in evaluating a model, as it addresses transportability, rather than reproducibility.³² The effect of factors, such as differences in demographics, operator-dependent radiological variances (for example, angle, rotation, and radiation dosage when performing a radiograph or CT), and brand and quality of radiograph machines on performance of the CNN need to be evaluated before one can transport any CNN to another institution.^{18,33} This is highlighted by Raisuddin et al,³⁴ who advocate for in-depth analyses of artificial intelligence models, as reported in their paper where their model had great performance on radiographs from the general population, but significantly reduced performance on cases that were deemed hard for diagnosis by clinicians.

In general, true model performance as tested via EV is lower than the performance assessed with the dataset used for model development.^{13,35,36} In this review for fracture detection and classification, studies conducting temporal EV reported similar or slightly improved performance compared to the IV set.^{24,25} In contrast, studies using a geographically split dataset reported a decrease in performance with use of EV,^{6,25,27} indicating the superiority of geographical over temporal validation. Blüthgen et al⁶ explains that the decrease in



performance observed indicates that the “variance” in images differed significantly between the IV and geographical EV sets, emphasizing the importance of geographical EV.

Not only variances in data, but also variation in labelling, can lead to varying performance: label noise can severely impact performance of CNNs,³⁷ and radiology reports are often based on only one observer.³⁸ In addition, these reports can have a variety of expertise and accuracy depending on who interprets the images.²⁴ Data labelling performed by a single expert carries significant risk of developing a biased CNN, catered to the opinion of one observer. Expert consensus can also be used, based on the assumption that agreement implies accuracy.³⁹ Nonetheless, limited availability of qualified experts to provide accurate image labels is a challenging problem when developing CNNs.³⁸ Although the input of experts—especially regarding evaluation of model predictions—is imperative to ensure clinical accuracy and relevance, reference standards such as follow-up imaging and surgical confirmation are considered the most accurate method to train CNNs.³⁸ However, these are not always available, especially in simple fractures.

Although the importance of and need for EV is highlighted by many studies,^{18,40-45} this review shows that EV of fracture recognition CNNs remains scarce. In addition, there is a lack of uniformity in the method of conducting and reporting of EV, such as defining ground truth. We therefore recommend readers to be cautious in interpreting performance when evaluation is limited to an internal or temporal validation set—as performance may vary when encountering data with different characteristics—and ideally geographical EV should be used to assess ‘true’ performance and generalizability. In addition, we advise the development and use of standardized methodology such as the recently published statements like the Clinical Artificial Intelligence Research (CAIR) checklist,⁴⁶ Standard Protocol Items: Recommendations for Interventional Trials—Artificial Intelligence (SPIRIT-AI),⁴⁷ and CONSolidated Standard for Reporting Trials—Artificial Intelligence (CONSORT-AI).⁴⁸ Several announced statements are still in development, like the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis—Machine Learning (TRIPOD-ML)⁴ and the Standards for Reporting Diagnostic Accuracy –Artificial Intelligence (STARD-AI). Using these standardized statements will improve methodological rigor, quality of future models, and facilitate eventual implementation in clinical practice.

Take home message

We recommend readers to be cautious in interpreting performance when evaluation is limited to an internal or temporal validation set—as performance may vary when encountering data with different characteristics—and ideally geographical external validation should be used to assess ‘true’ performance and generalizability.

REFERENCES

1. Topol, E.J., High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 2019. 25(1): p. 44-56.
2. Choy, G., O. Khalilzadeh, M. Michalski, S. Do, A.E. Samir, O.S. Pinykh, J.R. Geis, P.V. Pandharipande, J.A. Brink, and K.J. Dreyer, Current applications and future impact of machine learning in radiology. *Radiology*, 2018. 288(2): p. 318-328.
3. Liu, X., S.C. Rivera, L. Faes, L.F. Di Ruffano, C. Yau, P.A. Keane¹⁰, H. Ashrafi¹¹, A. Darzi¹¹, S.J. Vollmer, and J. Deeks, Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat. Med.*, 2019. 25: p. 1467-1468.
4. Collins, G.S. and K.G. Moons, Reporting of artificial intelligence prediction models. *The Lancet*, 2019. 393(10181): p. 1577-1579.
5. Adams, M., W. Chen, D. Holcdorf, M.W. McCusker, P.D.L. Howe, and F. Gaillard, Computer vs human: Deep learning versus perceptual training for the detection of neck of femur fractures. *Journal of Medical Imaging and Radiation Oncology*, 2019. 63(1): p. 27-32.
6. Blüthgen, C., A.S. Becker, I. Vittoria de Martini, A. Meier, K. Martini, and T. Frauenfelder, Detection and localization of distal radius fractures: Deep learning system versus radiologists. *European Journal of Radiology*, 2020. 126.
7. Urakawa, T., Y. Tanaka, S. Goto, H. Matsuzawa, K. Watanabe, and N. Endo, Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. *Skeletal Radiol*, 2019. 48(2): p. 239-244.
8. Chung, S.W., S.S. Han, J.W. Lee, K.S. Oh, N.R. Kim, J.P. Yoon, J.Y. Kim, S.H. Moon, J. Kwon, H.J. Lee, Y.M. Noh, and Y. Kim, Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthopaedica*, 2018. 89(4): p. 468-473.
9. Tomita, N., Y.Y. Cheung, and S. Hassanpour, Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Comput Biol Med*, 2018. 98: p. 8-15.
10. Yamada, Y., S. Maki, S. Kishida, H. Nagai, J. Arima, N. Yamakawa, Y. Iijima, Y. Shiko, Y. Kawasaki, T. Kotani, Y. Shiga, K. Inage, S. Orita, Y. Eguchi, H. Takahashi, T. Yamashita, S. Minami, and S. Ohtori, Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: ensemble decision-making with antero-posterior and lateral radiographs. *Acta Orthop*, 2020: p. 1-6.
11. Kalmet, P.H.S., S. Sanduleanu, S. Primakov, G. Wu, A. Jochems, T. Refaee, A. Ibrahim, L.V. Hulst, P. Lambin, and M. Poeze, Deep learning in fracture detection: a narrative review. *Acta Orthopaedica*, 2020. 91(2): p. 215-220.
12. Bongers, M.E., Q.C. Thio, A.V. Karhade, M.L. Stor, K.A. Raskin, S.A.L. Calderon, T.F. DeLaney, M.L. Ferrone, and J.H. Schwab, Does the SORG algorithm predict 5-year survival in patients with chondrosarcoma? An external validation. A Publication of The Association of Bone and Joint Surgeons® | *CORR*®, 2019. 477(10): p. 2296-2303.
13. Siontis, G.C., I. Tzoulaki, P.J. Castaldi, and J.P. Ioannidis, External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of clinical epidemiology*, 2015. 68(1): p. 25-34.
14. Liu, K.-L., T. Wu, P.-T. Chen, Y.M. Tsai, H. Roth, M.-S. Wu, W.-C. Liao, and W. Wang, Deep learning to distinguish pancreatic cancer tissue from non-cancerous pancreatic tissue: a retrospective study with cross-racial external validation. *The Lancet Digital Health*, 2020. 2(6): p. e303-e313.
15. Gertych, A., Z. Swiderska-Chadaj, Z. Ma, N. Ing, T. Markiewicz, S. Cierniak, H. Salemi, S. Guzman, A.E. Walts, and B.S. Knudsen, Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Scientific reports*, 2019. 9(1): p. 1-12.
16. Steyerberg, E.W., *Clinical prediction models*. 2019: Springer.



17. König, I.R., J. Malley, C. Weimar, H.C. Diener, and A. Ziegler, Practical experiences on the necessity of external validation. *Statistics in medicine*, 2007. 26(30): p. 5499-5511.
18. Oosterhoff, J.H., J.N. Doornberg, and M.L. Consortium, Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of Gartner's hype cycle. *EFORT Open Reviews*, 2020. 5(10): p. 593-603.
19. Ho, S.Y., K. Phua, L. Wong, and W.W.B. Goh, Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. *Patterns*, 2020. 1(8): p. 100129.
20. Moher, D., A. Liberati, J. Tetzlaff, D.G. Altman, and P. Group, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*, 2009. 6(7): p. e1000097.
21. Whiting, P.F., A.W. Rutjes, M.E. Westwood, S. Mallett, J.J. Deeks, J.B. Reitsma, M.M. Leeflang, J.A. Sterne, and P.M. Bossuyt, QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine*, 2011. 155(8): p. 529-536.
22. Pellegrini, E., L. Ballerini, M.d.C.V. Hernandez, F.M. Chappell, V. González-Castro, D. Anblagan, S. Danso, S. Muñoz-Maniega, D. Job, and C. Pernet, Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 2018. 10: p. 519-535.
23. Slim, K., E. Nini, D. Forestier, F. Kwiatkowski, Y. Panis, and J. Chipponi, Methodological index for non-randomized studies (MINORS): development and validation of a new instrument. *ANZ journal of surgery*, 2003. 73(9): p. 712-716.
24. Lindsey, R., A. Daluiski, S. Chopra, A. Lachapelle, M. Mozer, S. Sicular, D. Hanel, M. Gardner, A. Gupta, R. Hotchkiss, and H. Potter, Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences*, 2018. 115(45): p. 11591-11596.
25. Choi, J.W., Y.J. Cho, S. Lee, J. Lee, S. Lee, Y.H. Choi, J.-E. Cheon, and J.Y. Ha, Using a Dual-Input Convolutional Neural Network for Automated Detection of Pediatric Supracondylar Fracture on Conventional Radiography. *Investigative Radiology*, 2020. 55(2): p. 101-110.
26. Langerhuizen, D.W., S.J. Janssen, W.H. Mallee, M.P. van den Bekerom, D. Ring, G.M. Kerkhoffs, R.L. Jaarsma, and J.N. Doornberg, 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): p. 2482-2491.
27. Zhou, Q.-Q., J. Wang, W. Tang, Z.-C. Hu, Z.-Y. Xia, X.-S. Li, R. Zhang, X. Yin, B. Zhang, and H. Zhang, Automatic Detection and Classification of Rib Fractures on Thoracic CT Using Convolutional Neural Network: Accuracy and Feasibility. *Korean Journal of Radiology*, 2020. 21(7): p. 869-879.
28. Rajpurkar, P., J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, and R.L. Ball, Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*, 2017.
29. Baldwin, D.R., J. Gustafson, L. Pickup, C. Arteta, P. Novotny, J. Declerck, T. Kadir, C. Figueiras, A. Sterba, A. Exell, V. Potesil, P. Holland, H. Spence, A. Clubley, E. O'Dowd, M. Clark, V. Ashford-Turner, M.E. Callister, and F.V. Gleeson, External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. *Thorax*, 2020. 75(4): p. 306-312.
30. Milea, D., R.P. Najjar, J. Zubo, D. Ting, C. Vasseneix, X. Xu, M. Aghsaei Fard, P. Fonseca, K. Vanikieti, W.A. Lagreze, C. La Morgia, C.Y. Cheung, S. Hamann, C. Chiquet, N. Sanda, H. Yang, L.J. Mejico, M.B. Rougier, R. Kho, T. Thi Ha Chau, S. Singhal, P. Gohier, C. Clermont-Vignal, C.Y. Cheng, J.B. Jonas, P. Yu-Wai-Man, C.L. Fraser, J.J. Chen, S. Ambika, N.R. Miller, Y. Liu, N.J. Newman, T.Y. Wong, V. Biousse, and B. Group, Artificial Intelligence to Detect Papilledema from Ocular Fundus Photographs. *N Engl J Med*, 2020. 382(18): p. 1687-1695.
31. Nam, J.G., S. Park, E.J. Hwang, J.H. Lee, K.N. Jin, K.Y. Lim, T.H. Vu, J.H. Sohn, S. Hwang, J.M. Goo, and C.M. Park, Development and Validation of Deep Learning-based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. *Radiology*, 2019. 290(1): p. 218-228.

32. Justice, A.C., K.E. Covinsky, and J.A. Berlin, Assessing the generalizability of prognostic information. *Annals of internal medicine*, 1999. 130(6): p. 515-524.
33. Park, S.H. and K. Han, Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, 2018. 286(3): p. 800-809.
34. Raisuddin, A.M., E. Vaattovaara, M. Nevalainen, M. Nikki, E. Jarvenpaa, K. Makkonen, P. Pinola, T. Palsio, A. Niemensivu, O. Tervonen, and A. Tiulpin, Critical evaluation of deep neural networks for wrist fracture detection. *Sci Rep*, 2021. 11(1): p. 6006.
35. Moons, K.G., A.P. Kengne, D.E. Grobbee, P. Royston, Y. Vergouwe, D.G. Altman, and M. Woodward, Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*, 2012. 98(9): p. 691-698.
36. Zendel, O., M. Murschitz, M. Humenberger, and W. Herzner, How good is my test data? Introducing safety analysis for computer vision. *International Journal of Computer Vision*, 2017. 125(1-3): p. 95-109.
37. Karimi, D., H. Dou, S.K. Warfield, and A. Gholipour, Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 2020. 65: p. 101759.
38. England, J.R. and P.M. Cheng, Artificial intelligence for medical image analysis: a guide for authors and reviewers. *American Journal of Roentgenology*, 2019. 212(3): p. 513-519.
39. Kundel, H.L. and M. Polansky, Measurement of observer agreement. *Radiology*, 2003. 228(2): p. 303-308.
40. Weikert, T., L.A. Noordzij, J. Bremerich, B. Stieltjes, V. Parmar, J. Cyriac, G. Sommer, and A.W. Sauter, Assessment of a deep learning algorithm for the detection of rib fractures on whole-body trauma computed tomography. *Korean Journal of Radiology*, 2020. 21(7): p. 891-899.
41. Thian, Y.L., Y. Li, P. Jagmohan, D. Sia, V.E. Yao Chan, and R.T. Tan, Convolutional neural networks for automated fracture detection and localization on wrist radiographs. *Radiology: Artificial Intelligence*, 2019. 1(1).
42. Lee, C., J. Jang, S. Lee, Y.S. Kim, H.J. Jo, and Y. Kim, Classification of femur fracture in pelvic X-ray images using meta-learned deep neural network. *Scientific reports*, 2020. 10(1): p. 13694.
43. Al-Helo, S., R.S. Alomari, S. Ghosh, V. Chaudhary, G. Dhillon, M.B. Al-Zoubi, H. Hiary, and T.M. Hamtini, Compression fracture diagnosis in lumbar: A clinical CAD system. *International Journal of Computer Assisted Radiology and Surgery*, 2013. 8(3): p. 461-469.
44. Badgeley, M.A., J.R. Zech, L. Oakden-Rayner, B.S. Glicksberg, M. Liu, W. Gale, M.V. McConnell, B. Percha, T.M. Snyder, and J.T. Dudley, Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2019. 2(1).
45. Derkatch, S., C. Kirby, D. Kimelman, M.J. Jozani, J.M. Davidson, and W.D. Leslie, Identification of Vertebral Fractures by Convolutional Neural Networks to Predict Nonvertebral and Hip Fractures: A Registry-based Cohort Study of Dual X-ray Absorptiometry. *Radiology*, 2019. 293(2): p. 405-411.
46. Olczak, J., J. Pavlopoulos, J. Prijs, F.F.A. IJpma, J.N. Doornberg, C. Lundström, J. Hedlund, and M. Gordon, Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders - an introductory reference with a guideline and checklist proposal. [Accepted for publication in *Acta Orthopaedica*].
47. Cruz Rivera, S., X. Liu, A.W. Chan, A.K. Denniston, M.J. Calvert, A.I. Spirit, C.-A.W. Group, A.I. Spirit, C.-A.S. Group, A.I. Spirit, and C.-A.C. Group, Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*, 2020. 26(9): p. 1351-1363.
48. Liu, X., S. Cruz Rivera, D. Moher, M.J. Calvert, A.K. Denniston, A.I. Spirit, and C.-A.W. Group, Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit Health*, 2020. 2(10): p. e537-e548.





Ningaloo Reef, Exmouth, Western Australia, Australia

**Artificial Intelligence Fracture Recognition
on Computed Tomography: Review of
Literature and Recommendations**

L.H.M. Dankelman

S.R.A. Schilstra

F.F.A. IJpma

J.N. Doornberg

J.W. Colaris

M.H.J. Verhofstad

M.M.E. Wijffels

J. Prijs

on behalf of Machine Learning Consortium

ABSTRACT

Purpose

The use of computed tomography (CT) in fractures is time consuming, challenging and suffers from poor inter-surgeon reliability. Convolutional neural networks (CNNs), a subset of artificial intelligence (AI), may overcome shortcomings and reduce clinical burdens to detect and classify fractures. The aim of this review was to summarize literature on CNNs for the detection and classification of fractures on CT scans, focusing on its accuracy and to evaluate the beneficial role in daily practice.

Methods

Literature search was performed according to the PRISMA statement, and Embase, Medline ALL, Web of Science Core Collection, Cochrane Central Register of Controlled Trials and Google Scholar databases were searched. Studies were eligible when the use of AI for the detection of fractures on CT scans was described. Quality assessment was done with a modified version of the methodologic index for nonrandomized studies (MINORS), with a seven-item checklist. Performance of AI was defined as accuracy, F1-score and area under the curve (AUC).

Results

Of the 1140 identified studies, 17 were included. Accuracy ranged from 69 to 99%, the F1-score ranged from 0.35 to 0.94 and the AUC, ranging from 0.77 to 0.95. Based on ten studies, CNN showed a similar or improved diagnostic accuracy in addition to clinical evaluation only.

Conclusions

CNNs are applicable for the detection and classification of fractures on CT scans. This can improve automated and clinician-aided diagnostics. Further research should focus on the additional value of CNN used for CT scans in daily clinics.

INTRODUCTION

The use of computed tomography (CT) in trauma care is substantially increasing. In the Netherlands, over 2 million CT scans were made in 2019 and this number increases each year.¹ Total-body CTs are increasingly used in acute trauma settings and can be more cost-effective than standard radiological imaging.² Increased use of imaging strains radiologists, to the point of creating a shortage of radiologist in hospitals.³ Examining CT scans and radiographs to detect and classify fractures can be time consuming, challenging, and poor inter-observer variability among radiologists and (experienced) clinicians can be substantial.³ Artificial intelligence (AI) could play a big role optimizing workflows in the acute setting and allow clinicians to spend their time more effectively.

AI can execute different tasks, ranging from searching the web to self-driving cars—tasks that until a few years ago could only be performed by humans. Deep learning (DL) is a subset of machine learning (ML) that uses mainly convolutional neural networks (CNNs).⁴ CNNs are combinations of artificial neuron layers with different units. These units operate like neurons of our brain.³ CNNs can learn to recognize discriminative features from data and assign importance to various aspects in the image and to differentiate one from another. An example of data used to train an ankle fracture CT CNN is presented in Supplemental Video 1. While most earlier AI methods have led to applications with subhuman performance, recent CNNs are able to match and even surpass the capacity of humans detecting certain fractures on radiographs, focusing on isolated fracture types per model.^{5–9} The strength of computers is their ability to evaluate a vast number of examinations rapidly, consistently and without exhaustion.

When clinicians are aided by DL-based automatic fracture detection algorithms, the accuracy of clinical diagnosis might improve and time to diagnosis reduced, which can be useful in, among others, an emergency setting. Various studies have successfully applied CNNs to detect fractures of various body parts on radiographs.^{5–9} The results in detecting and classifying fractures on radiographs by CNNs are promising. However, only a few studies have developed CNNs for the detection of fractures on CT scans. Therefore, we conducted this systematic review to give an overview of studies using AI with CT scans to detect or classify fractures. The aim of this study was to answer the following questions: 1) What is the accuracy of a CNN in detecting fractures on CT scans? 2) Does the use of CNNs with CT scans improve the diagnostic performance of clinicians?



MATERIALS AND METHODS

Article selection, quality assessment and data extraction

A systematic literature search was performed according to the PRISMA statement¹⁰ (Fig. 1) and conducted in the following libraries: Embase, Medline ALL, Web of Science Core Collection, Cochrane Central Register of Controlled Trials and Google Scholar. The search strategy was formulated together with a librarian (see appendix 1).

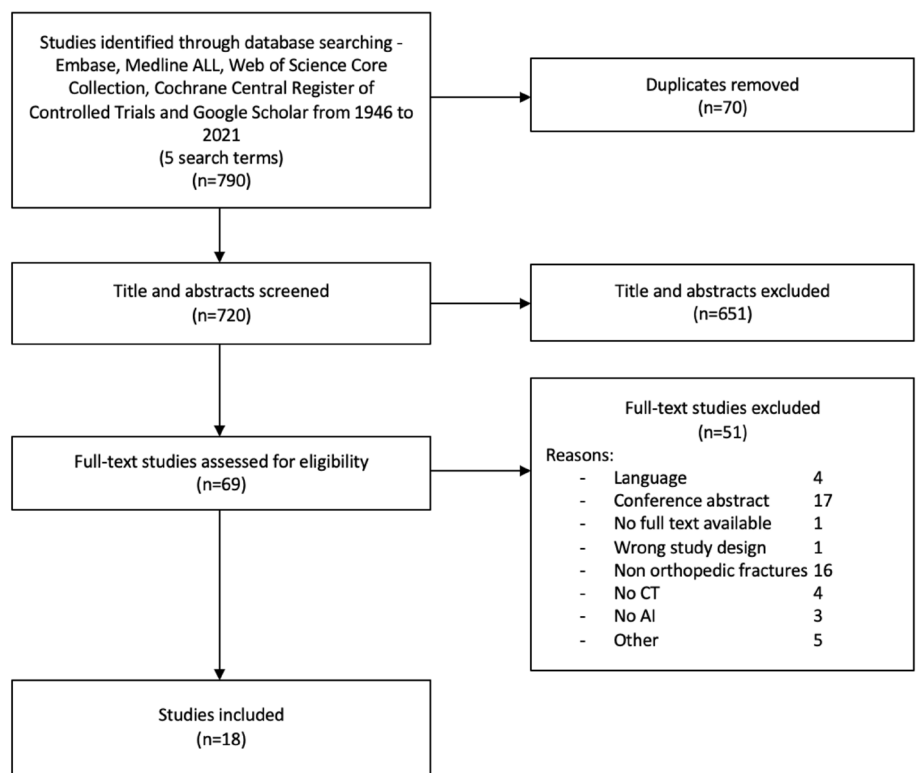


Figure 1. PRISMA flowchart

All published articles were searched. After removing duplicates, titles and abstracts of the potentially eligible articles were independently screened by two reviewers (LD, SS). Subsequently, full-text screening was performed using the predefined criteria to check eligibility. If the conclusion was inconsistent, a third reviewer was consulted (JP). Articles met the inclusion criteria if AI was used to detect fractures on CT scans in an orthopaedic trauma setting. The defined exclusion criteria were: review articles or letters, conference abstracts, technique papers, studies using robots, animal and cadaveric studies, non-orthopaedic

fractures and studies not published in English or Dutch. Covidence (Veritas Health Innovation, Melbourne, Australia) was used for the screening process and full-text review.

The quality of all included articles was assessed by two independent reviewers (LD, SS). In case of a disagreement, a third reviewer was consulted (JP). For the quality assessment, a modified version of the methodologic index for nonrandomized studies (MINORS) instrument was used, including the following items: disclosure, study aim, input features, ground truth, dataset distribution, performance metric and AI model (Table 1). Studies with low scores on three or more items were excluded. Standardized forms were used to extract and record data (Microsoft Excel Version 16.21; Microsoft Inc, Redmond, WA, USA).

Outcome Measures

In this study, the primary outcome was performance of the CNNs used, measured by their accuracy, F1-scores and area under the curve (AUC). Seventeen studies met the inclusion criteria and were used to answer this research question. To answer the secondary question in this study, ten studies comparing performance of the CNN to performance clinicians were used. The data points collected from each study were: author, year of publication, anatomical location of the fracture, AI models used (type), imaging direction of CT slices, output classes, ground truth label assignment, number of patients and performance metric (e.g., accuracy, AUC curve) (Table 2).

Output classes included fracture detection (i.e., fracture yes/no) and/or classification (i.e., OA/OTA classification). All studies described the detection of fractures by the CNN, and seven studies also performed fracture classification.

Studies used accuracy, F1-score and AUC to measure the performance of the model. The F1-score ($2 * ((\text{precision} * \text{recall}) / (\text{precision} + \text{recall}))$) is the harmonic mean of the precision (positive predictive value) and recall (sensitivity) of the test, where it requires both to be high for a favorable F1-score. The highest possible value is 1.0, indicating a perfect precision and recall, and the lowest possible value is 0. If not assessed, the F1-score was calculated when precision and recall were stated. The area under the curve (AUC) is a score to measure the ability of a classifier to distinguish between classes. The score lies between 0.5 (a classifier equal to that of chance) and 1 (an excellent classifier). Where possible, accuracy and/or F1-scores were calculated to facilitate comparison between studies.



Table 1. Quality assessment according to MINORS criteria

Author, year	Study type	Disclosure	Study aim	Input feature	Ground truth	External validation method	Performance metric	AI model
Castro-Zunty et al [17]	Classification	1	1	1	1	1	1	1
Dreizin et al [18]	Classification	1	1	1	1	1	1	1
Hu et al [13]	Detection/classification	1	1	0	1	1	1	1
Jin et al [27]	Detection	1	1	1	1	1	1	1
Kaiume et al [19]	Detection	1	1	1	1	1	1	1
Meng et al [20]	Detection/classification	1	1	1	1	1	1	1
Pranata et al [15]	Detection/classification	1	1	1	0	1	1	1
Raghavendra et al [16]	Detection	1	1	1	0	1	1	1
Roth et al [11]	Detection	0	1	0	0	1	1	1
Small et al [14]	Detection	1	1	0	1	1	1	1
Ukai wet al [21]	Detection	1	1	1	1	1	1	1
Voter et al [28]	Detection	1	1	1	1	1	1	1
Weikert et al [22]	Detection	1	1	1	1	1	1	1
Yacoub et al [23]	Detection	1	1	1	1	1	1	1
Yamamoto et al [12]	Detection	0	1	1	1	1	1	1
Yoon et al [24]	Detection/classification	1	1	1	1	1	1	1
Zhou et al [25]	Detection/classification	1	1	1	1	1	1	1
Zhou et al [26]	Detection/classification	1	1	1	1	1	1	1

Table 2. Description of studies

Author, year	Anatomical location	AI models used (Type)	Imaging direction of CT slices	Output classes	Ground truth label assignment	Number of fractures	Performance metric	Performance	Comparison CNN vs. radiologist
Castro-Zunti et al., 2021 [17]	Ribs	CNN (InceptionV3)	Axial	3 subclasses	2 radiologists	N = 612 ^a	Accuracy AUC	3 classes: 96.00% Binary: 97.76% Binary: 94.7% (95% CI 94.1–95.3) ^c	Yes
Dreizin et al., 2021 [18]	Pelvic	CNN (ResNetXt-50 + LSTM)	Axial, coronal, sagittal	2	3 radiologists	N = 373 ^a	Accuracy F1 score	85%; 74% (discriminating translational and rotational instability respectively) 0.63; 0.77	Yes
Hu et al., 2021 [13]	Ribs	CNN (SGANet)	Axial	2	2 doctors	N = 398	Accuracy, F1 score	82.54% 0.7843	Yes
Jin et al., 2020 [27]	Ribs	DL (FracNet)	Axial	2	2 radiologists and 2 junior radiologists	N = 7473	(Sensitivity)	92.9%	Yes
Kaiume et al., 2021 [19]	Ribs	CNN (DenseNet + SSD)	Axial	2	2 radiologists	N = 256	F1 score;	0.711	No
Meng et al., 2021 [20]	Ribs	CNN (VRB-Net)	Axial	4 subclasses	2 senior radiologists	N = 861	Accuracy F1 score	86.3% 0.940	Yes
Pranata et al., 2019 [15]	Calcaneus	CNN (ResNet; VGG)	Coronal, sagittal, axial	2	N.A	N = 1931	Accuracy	ResNet: 80–98% VGG: 92–98%	No
Raghavendra et al., 2018 [16]	Thoracolumbar	CNN	Sagittal	2	N.A	N = 700	Accuracy	99.1% max; 96.51% average	No
Small et al., 2021 [16]	Spine	CNN (ResNet)	Axial, coronal, sagittal	2	2 neuroradiologists	N = 143	Accuracy	(95% CI, 90–94%)	Yes



Table 2. Description of studies (continued)

Author, year	Anatomical location	AI models used (Type)	Imaging direction of CT slices	Output classes	Ground truth label assignment	Number of fractures	Performance metric	Performance	Comparison CNN vs. radiologist
Ukai et al., 2021 [21]	Pelvic	CNN (YOLOv3)	Axial, coronal, sagittal (3D)	3	Orthopedic surgeons	N = 389	F1-score AUC	0.853 0.824	No
Voter et al., 2021 [28]	CSFx	AI DSS (Aidoc)	Axial	2	Neuroradiologist; radiology	N = 173	F1-score ^b	0.453 [*]	No
Weikert et al., 2020 [22]	Ribs	CNN (ResNet)	Axial	3 subclasses	Written CT reports approved by a board certified radiologist	N = 159	Accuracy F1 score	90.2% (95% CI 87.3–92.6) 0.85	Yes
Yacoub et al., 2021 [23]	VCF	CNN (AI-Rad Companion)	Sagittal	2	2 radiologists	N = 100 ^a	AUC F1-score ^b	0.82 (95% CI 0.73–0.89) 0.352 ^b	Yes
Yamamoto et al., 2020 [12]	Pelvic	CNN (VGG-16)	Sagittal, axial, coronal (3D)	2	4 residents and 5 orthopaedic specialists	N = 103 ^a	Accuracy F1-score ^b	69.4% 0.578 ^b	No
Yoon et al., 2020 [24]	Femur	CNN (Faster R-CNN)	Sagittal, axial, coronal (3D)	2 groups, 10 subclasses	Orthopedic surgeons	N = 3343	Accuracy	Per class: 97% 95% 94% 92% 90%	No

Table 2. Description of studies (continued)

Author, year	Anatomical location	AI models used (Type)	Imaging direction of CT slices	Output classes	Ground truth label assignment	Number of fractures	Performance metric	Performance	Comparison CNN vs. radiologist
Zhou et al., 2021 [25]	Ribs	CNN (Faster R-CNN, ResNet-101)	Axial	3	2 musculoskeletal radiologists, 2 senior radiologists, thoracic surgeon	N = 4215	F1 score Accuracy AUC	Model I vs. Model I/T 0.814 vs. 0.875 0.816 vs. 0.847 0.378 vs. 0.839	Yes
								78.8% vs. 85.2% 81.3% vs. 90.4% 73.9% vs. 88.5% 83.6% vs. 90.7% 88.7% vs. 94.2% 77.0% vs. 90.5%	
Zhou et al., 2020 [26]	Ribs	CNN (Faster R-CNN, YOLOv3)	Axial	3	2 musculoskeletal radiologists, 2 senior radiologists, thoracic surgeon	N = 1079 ^a	F1 score	0.849 0.856 0.770 Mean: 0.825	Yes

AI/artificial intelligence, DSS decision support systems, CNN Convolutional Neural Networks, DL deel learning, AUC Area Under the Curve, LSTM a long short-term memory network, DenseNet Densely connected convolutional Network, SSD single shot multibox detector, ResNetResidual network, SGANet slice grouping and aggregation network, VGG Visual geometry group, VRB-net V-net, ResNey and Bottleneck ResNet Network, YOLOv3 You Only Look Once, version 3

^aNumber of fractures not given, number of patients stated

^bF1-score calculated with

^cAUC score given in percentages



Quality appraisal

The modified MINORS tool included the following items: disclosure, study aim, input feature, ground truth, dataset distribution and performance metric (Table 1). Disclosure was reported in all but two studies.^{11, 12} All studies clearly stated their study aim, model used and how performance was measured. The input feature was not clearly specified in three studies.^{11, 13, 14} These studies did not mention what the inclusion and exclusion criteria were. Three studies did not specify the ground truth (the reference standard used in AI).^{11, 15, 16} One study was excluded after the quality assessment, because it scored too low on three items: disclosure, input feature, and ground truth.¹¹

RESULTS

Included studies

The search yielded a total of 1140 articles. After duplicate removal, 720 abstracts were screened. Sixty-nine studies were selected for full-text screening, of which eighteen remained. No new eligible studies were identified through screening the reference lists. One study was excluded after quality assessment, because the risk of bias was deemed too high due to unclear reporting of disclosure, input feature and ground truth.¹¹ Seventeen studies were used for analysis.

Description of studies

All seventeen studies used a CNN to detect and /or classify fractures on CT scans.¹²⁻²⁸ Eight studies addressed detection of rib fractures^{13, 17, 19, 20, 22, 25-27}, three studies the performance for detection^{12, 21} and classification¹⁸ of pelvic fractures, four for detection of spine fractures^{14, 16, 23, 28}, one for detection and classification of femur fractures²⁴ and one of calcaneal fractures¹⁵. Fourteen studies used two output classes (fracture yes/no).

One study on spine fractures used three output classes: completely displaced, incompletely displaced and compression fracture.¹⁴ In addition, two studies used fresh, healing and old fracture as output classes.^{25, 26} In 12 studies, the ground truth for diagnosis and classification of the fractures was the conclusion of two or more experts, who interpreted the CT scans independently.^{12-14, 17-20, 23, 25-28} One study used radiology reports from routine care as ground truth.²² Two studies did not specify how many experts provided the ground truth.^{21, 24} Thereby, two studies did not report the ground truth.^{15, 16} The number of patients included in the studies ranged from 39¹⁹ to 8529²⁰ fractures.

Primary outcome: the performance of CNN

The performance was defined in various ways among studies. Accuracy on detection and/or classification was measured in eleven studies^{12–18, 20, 22, 24, 25}, ranging from 69.4%¹² to 99.1%¹⁶. Eight studies used the F1-score to assess performance instead: in two the F1-score was assessed for the classification of healing status^{25, 26}, in one for displacement²¹, and in five^{13, 18–20, 22} for the detection of fractures. Additionally, we calculated the F1-scores in three studies^{12, 23, 28} to facilitate comparison. F1-scores ranged from 0.35 in Yacoub et al.²³ to 0.94 in Meng et al.²⁰ Four studies reported the AUC as a performance metric^{17, 21, 23, 25}, ranging from 0.770²⁵ to 0.947¹⁷. Zhou et al.²⁵ reported the AUC on classification of challenging fractures compared to the other three studies with more simple fracture detection. One study just reported a sensitivity of 92.9%.²⁷

In Castro-Zunti et al.¹⁷, the accuracy and AUC scores of four different AI models were compared for 612 patients. They found that the CNN model InceptionV3 achieved the highest average accuracy of 96%, when the CT slices were divided into three classes (acute, old (healed) and normal (non-fractured)). In Yoon et al.²⁴ the data were divided into ten classes (based on the AO/OTA classification²⁹) and the accuracy of the different numbers of output classes was reported for 85 patients. Binary classification (no fracture vs fracture) achieved the highest accuracy of 97%. When the data were divided into more classes (AO/OTA classification²⁹), the accuracy decreased to the lowest value of 90% for ten classes, as compared to the ground truth by orthopaedic surgeons. Dreizin et al.¹⁸ reported the superiority of translational instabilities (85%) over rotational ones (74%) on the accuracy and F1-score of their model¹⁸ for 373 patients. Zhou et al.²⁵ reported improved performance on 1020 patients using CTs combined with patient information compared (accuracy for three different models: 85.2%, 90.4% and 88.5%) to just CTs alone (accuracy for three different models: 78.8%, 81.3% and 73.9%).²⁵ In another—earlier—study, Zhou et al.²⁶ reported that the mean F1-score of healing rib fractures was the highest and of old fractures the lowest (0.856 vs. 0.770).

In Fig. 2, the amount CTs for training, validation and testing are plotted against the accuracy, with increasing accuracy from left to right. The study with the most CTs reported an average accuracy of 92%.¹⁴ The highest accuracy of 97% was reported in a study¹⁷ with only 612 CTs.

In summary, the reported outcomes on accuracy (ranging from 69.4 to 99.1%), the F1-score (from 0.35 to 0.94), the AUC (from 0.770 to 0.947) and the sensitivity (92.9%) were assessed on different classifications, CNN models and training, validation and testing sets across the included studies.



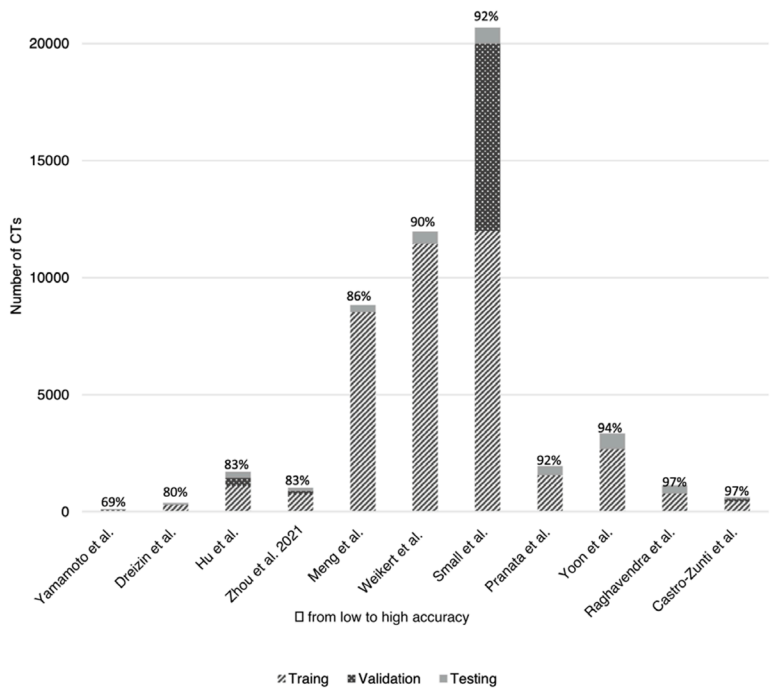


Figure 2. Correlation between accuracy and total number of CTs

Secondary outcome: CNN and clinicians

Ten out of seventeen studies compared a CNN model to the diagnostic performance of radiologists^{14, 17, 22, 23, 25-27} or radiology reports^{13, 22, 23}. Seven^{14, 17, 22, 23, 25-27} out of these ten studies compared the sensitivity of a CNN model to radiologists. In three studies^{17, 26, 27}, the CNN model solely or as an additional CNN model resulted in a higher sensitivity compared to the radiologist alone. Three studies showed a similar sensitivity for CNN and radiologist,^{22, 23, 25} and one¹⁴ showed a decrease in sensitivity with CNN. Four studies reported a significant reduction in time to diagnosis when a radiologist was aided by a CNN.^{20, 25-27}

Two out of ten studies compared the accuracy of CNN vs clinicians.^{18, 20} In Meng et al.²⁰, junior radiologists significantly improved their accuracy when assisted by a CNN for detection and classification of fractures. Experienced radiologists showed similar improvement.²⁰ In Dreizin et al.¹⁸, the model was equivalent in accuracy compared to radiologists. One study showed that when CNN is combined with clinical reports, the number of missed diagnoses is reduced by 88%.¹³

In summary, the four studies^{13, 20, 26, 27} that reported the performance of a CNN as an aid for the radiologist showed that CNN increases the performance of detection and classification of fractures. Twelve^{13, 15-20, 22, 24-27} out of seventeen studies concluded that the use of a CNN

improved or could improve clinical care. In the remaining five studies, three studies^{14, 21, 28} recommend CNN as a second-stage interpretation to assist radiologists, in one performance was inferior to clinical radiology reports²³ and lastly, one did not report on improvement¹².

DISCUSSION

In this systematic review, the results of several studies using AI for fracture detection and classification—in particular convolutional neural networks (CNNs)—were analyzed. The included studies reveal that CNNs show good performance in detecting and classifying various fractures on CT scans. The use of CNNs may add value and efficiency to several components of the skeletal imaging workflow in trauma care. The overall conclusion in most of studies was that CNNs are applicable in aiding clinicians, by reducing both time to diagnosis and number of missed diagnoses while improving the diagnostic performance. In addition, CNNs have proven to be very consistent, in contrast to the high inter-observer variability among radiologists and surgeons, when interpreting CT scans.³ Due to the scarcity of studies reviewing the place of CNNs in trauma CT imaging, the search strategy was very broad, and various libraries were queried. In addition, this study looks at the comparison of CNN versus clinicians or CNN as an assistant for clinicians.

This study should be interpreted in light of strengths and weaknesses. First, comparability of the studies is limited, because some fractures may be easier to detect, have different characteristics, and are in different surrounding anatomical structures than others. However, the results of the studies show comparable performances across the board and this heterogeneity did not affect answering our research questions. Secondly, different definitions for the ground truth were used among the various studies. For example, ground truth labels might be determined by various numbers of radiologists with different levels of expertise. An important note is that all these reference standards are subject to human biases. Lastly, to date, only a small number of studies have investigated the use of AI for fracture detection on CT scans, in limited patient group sizes. This may overestimate the potential benefit of AI, and therefore, future research should overcome this shortcoming. In addition, for the use of CNN models in daily practice, these models need to be further developed, with greater training and testing sets, external validation and prospective validation. However, if the beneficial effect of AI in fracture diagnosing and treatment results in improvement, this might impede extensive changes for the daily clinic. Strengths include the search of multiple databases, the use of a modified MINORS that included CNN-specific factors such as the input feature, ground truth, dataset distribution and performance metric. Future studies investigating AI on CTs for fracture detection and classification should include a wide data base of training, validation and testing sets, report demographic and diagnostic performance metrics, external validation of the CNN model [30] and the investigation of more common fractures (for example, wrist and ankle).



In general, for CNNs, it is assumed that the larger the dataset, the higher the performance. Training with a small dataset is a major cause of overfitting and does not lead to suitable generalization of performance. Due to the heterogeneity of the studies, straightforward conclusions for the recommended size of datasets cannot be drawn. However, a clear correlation for all fractures sites between accuracy and data size, with some studies reaching perfect accuracy with small datasets of less than 1200 CT scans, seems to be lacking. Taking this in consideration, in combination with the limited time of experts to provide high-quality labels, we recommend a stepwise approach of small dataset that increases in increments until adequate performance, or plateau is reached.

Most studies used the same base CNN architectures. Five studies used ResNet.^{14, 15, 18, 22, 25} They showed a similar accuracy, while investigating different anatomical locations. Two studies used YOLOv3 and both showed similar F1-scores.^{21, 26} Two studies used the CNN model VVG-16.^{12, 15} The accuracy measured in these studies was divergent. Pranata et al.¹⁵ presented a very accurate CNN model for detection of calcaneal fractures, while the accuracy found for detection of pelvic fractures¹² was significantly lower. A reason for this difference could be the group size of both studies; 1931 calcaneal fractures vs. 103 pelvic fractures. Furthermore, the stability of the pelvis is based both on bony and/or ligamentous injury, a much more challenging task compared to finding cortical fractures.

RestNet (or a modified version) was the most used CNN network, with reported accuracies between 73 and 98%. The best-performing model was reported by Raghavendra et al.¹⁶ that showed an average accuracy of 96.51%. This model was developed by the authors, however, without external validation which warrants some caution in interpretation of the results.³⁰ Less than half (6/17) of all studies reported the use of an external validation. To implement in clinical practice, external validation of CNN models is crucial to explore transportability and bias³⁰ and will be the topic of future studies.

Other fields are ahead of orthopaedics with regard to the use of CNNs as computer-aided detection. CNNs have been reported in oncology for: the classification of biopsy-proven masses and normal tissue on mammograms³¹, classification of skin cancer³² and the automated detection of pathological mediastinal lymph nodes in lung cancer³³. CNNs have been shown to improve diagnostic performance in detection of lung nodes and coronary artery calcium on CTs in lung cancer screening.³⁴ The use of CNNs in fracture detection and classification is only following in the footsteps of much further developments in other specialties.

In conclusion, CNNs can detect fractures and important fracture characteristics on CT scans, which may be used to guide treatment and optimize diagnosis of fractures. In addition, computers can evaluate a vast number of examinations rapidly, consistently and without exhaustion. If CNNs are trained well, using at least multiple experts to provide the ground truth, this could reduce the inter-observer variability plaguing daily practice, and be a valuable application in a trauma setting by reducing time to diagnosis. Further research is needed to explore strengths and weaknesses of CNNs in an acute trauma setting.

REFERENCES

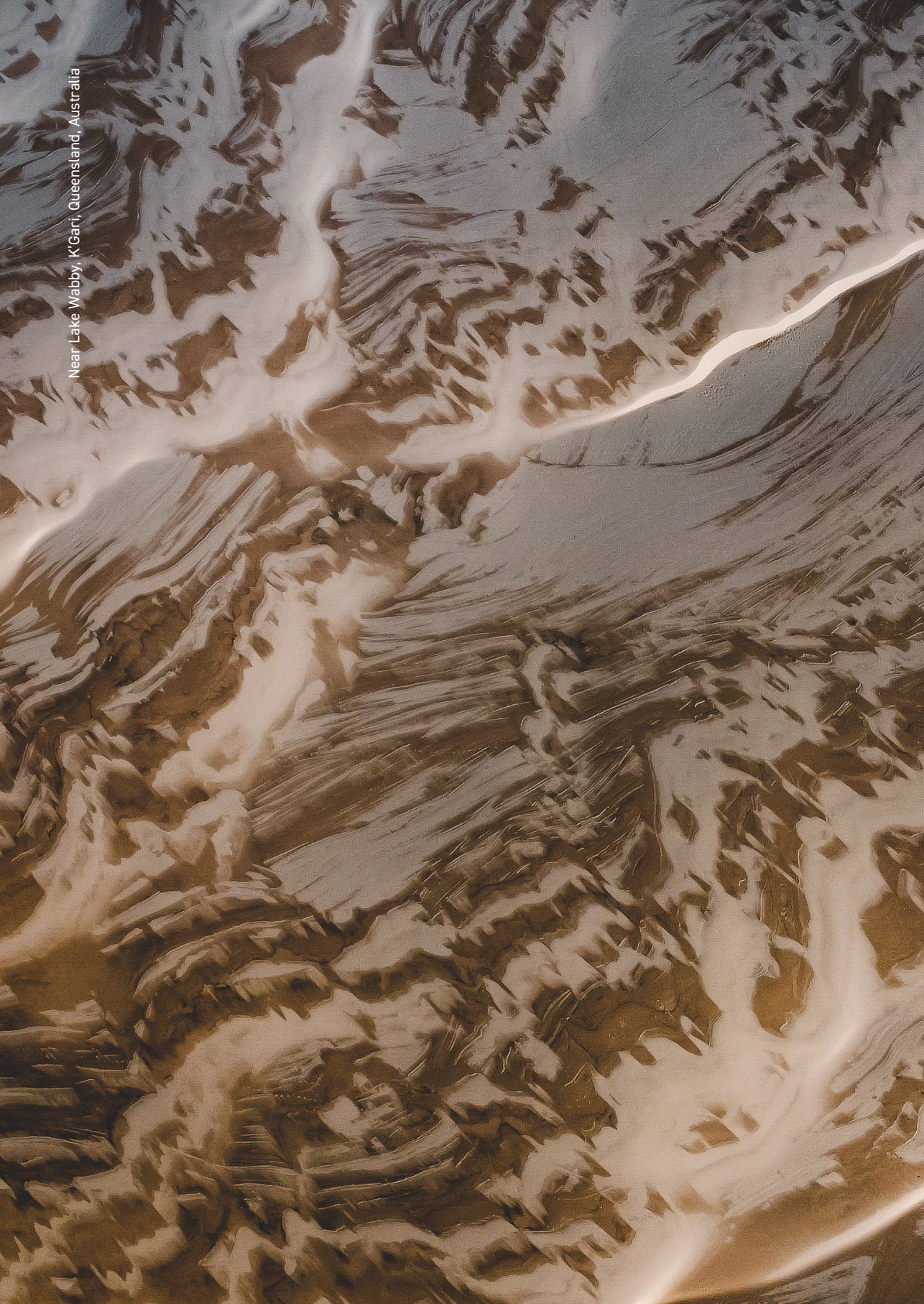
1. RIVM. Trends in het aantal CT-onderzoeken 2021 [Available from: <https://www.rivm.nl/medische-stralingstoepassingen/trends-en-stand-van-zaken/diagnostiek/computer-tomografie/trends-in-aantal-ct-onderzoeken>].
2. Treskes K, Sierink JC, Edwards MJR, Beuker BJA, Van Lieshout EMM, Hohmann J, et al. Cost-effectiveness of immediate total-body CT in patients with severe trauma (REACT-2 trial). *Br J Surg*. 2021;108(3):277-85.
3. Kalmet PHS, Sanduleanu S, Primakov S, Wu G, Jochems A, Refaee T, et al. Deep learning in fracture detection: a narrative review. *Acta Orthop*. 2020;91(2):215-20.
4. Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. *Annu Rev Biomed Eng*. 2017;19:221-48.
5. Adams M, Chen W, Holcdorf D, McCusker MW, Howe PD, Gaillard F. Computer vs human: Deep learning versus perceptual training for the detection of neck of femur fractures. *J Med Imaging Radiat Oncol*. 2019;63(1):27-32.
6. Chung SW, Han SS, Lee JW, Oh KS, Kim NR, Yoon JP, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop*. 2018;89(4):468-73.
7. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A*. 2018;115(45):11591-6.
8. Blüthgen C, Becker AS, Vittoria de Martini I, Meier A, Martini K, Frauenfelder T. Detection and localization of distal radius fractures: Deep learning system versus radiologists. *Eur J Radiol*. 2020;126:108925.
9. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthop*. 2017;88(6):581-6.
10. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Int J Surg*. 2010;8(5):336-41.
11. Roth HR, Wang YN, Yao JH, Lu L, Burns JE, Summers RM. Deep convolutional networks for automated detection of posterior-element fractures on spine CT. 2015;9785.
12. Yamamoto N, Rahman R, Yagi N, Hayashi K, Maruo A, Muratsu H, et al. An automated fracture detection from pelvic CT images with 3-D convolutional neural networks. 2020.
13. Hu Y, He X, Zhang R, Guo L, Gao L, Wang J. Slice grouping and aggregation network for auxiliary diagnosis of rib fractures. *Biomed Signal Process Control*. 2021;67.
14. Small JE, Osler P, Paul AB, Kunst M. Ct cervical spine fracture detection using a convolutional neural network. *Am J Neuroradiol*. 2021;42(7):1341-7.
15. Pranata YD, Wang KC, Wang JC, Idram I, Lai JY, Liu JW, et al. Deep learning and SURF for automated classification and detection of calcaneus fractures in CT images. *Comput Methods Programs Biomed*. 2019;171:27-37.
16. Raghavendra U, Bhat NS, Gudigar A, Acharya UR. Automated system for the detection of thoracolumbar fractures using a CNN architecture. *Future Generation Computer Systems-the International Journal of Esience*. 2018;85:184-9.
17. Castro-Zunti R, Chae KJ, Choi Y, Jin GY, Ko SB. Assessing the speed-accuracy trade-offs of popular convolutional neural networks for single-crop rib fracture classification. *Comput Med Imaging Graph*. 2021;91.
18. Dreizin D, Goldmann F, LeBedis C, Boscak A, Dattwyler M, Bodanapally U, et al. An Automated Deep Learning Method for Tile AO/OTA Pelvic Fracture Severity Grading from Trauma whole-Body CT. *J Digit Imaging*. 2021;34(1):53-65.



19. Kaiume M, Suzuki S, Yasaka K, Sugawara H, Shen Y, Katada Y, et al. Rib fracture detection in computed tomography images using deep convolutional neural networks. *Medicine (Baltimore)*. 2021;100(20):e26024.
20. Meng XH, Wu DJ, Wang Z, Ma XL, Dong XM, Liu AE, et al. A fully automated rib fracture detection system on chest CT images and its impact on radiologist performance. *Skelet Radiol*. 2021.
21. Ukai K, Rahman R, Yagi N, Hayashi K, Maruo A, Muratsu H, et al. Detecting pelvic fracture on 3D-CT using deep convolutional neural networks with multi-orientated slab images. *Sci Rep*. 2021;11(1):11716.
22. Weikert T, Noordtzi LA, Bremerich J, Stieltjes B, Parmar V, Cyriac J, et al. Assessment of a deep learning algorithm for the detection of rib fractures on whole-body trauma computed tomography. *Korean J Radiol*. 2020;21(7):891-9.
23. Yacoub B, Kabakus IM, Schoepf UJ, Giovagnoli VM, Fischer AM, Wichmann JL, et al. Performance of an Artificial Intelligence-Based Platform Against Clinical Radiology Reports for the Evaluation of Noncontrast Chest CT. *Acad Radiol*. 2021.
24. Yoon SJ, Kim TH, Joo SB, Oh SE. Automatic multi-class intertrochanteric femur fracture detection from ct images based on ao/ota classification using faster r-cnn-bo method. *J App Biomed*. 2020;18(4):97-105.
25. Zhou QQ, Tang W, Wang J, Hu ZC, Xia ZY, Zhang R, et al. Automatic detection and classification of rib fractures based on patients' CT images and clinical information via convolutional neural network. *Eur Radiol*. 2021;31(6):3815-25.
26. Zhou QQ, Wang J, Tang W, Hu ZC, Xia ZY, Li XS, et al. Automatic detection and classification of rib fractures on thoracic ct using convolutional neural network: Accuracy and feasibility. *Korean J Radiol*. 2020;21(7):869-79.
27. Jin L, Yang J, Kuang K, Ni B, Gao Y, Sun Y, et al. Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet. *EBioMedicine*. 2020;62.
28. Voter AF, Larson ME, Garrett JW, Yu JPJ. Diagnostic Accuracy and Failure Mode Analysis of a Deep Learning Algorithm for the Detection of Cervical Spine Fractures. *AJNR Am J Neuroradiol*. 2021;42(8):1550-6.
29. Association OT. Fracture and dislocation classification compendium *Journal of Orthopaedic Trauma* 2018 32.
30. Oliveira ECL, van den Merkhof A, Olczak J, Gordon M, Jutte PC, Jaarsma RL, et al. An increasing number of convolutional neural networks for fracture recognition and classification in orthopaedics : are these externally validated and ready for clinical application? *Bone Jt Open*. 2021;2(10):879-85.
31. Giger ML. Machine Learning in Medical Imaging. *J Am Coll Radiol*. 2018;15(3 Pt B):512-20.
32. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-8.
33. Wallis D, Soussan M, Lacroix M, Akl P, Dubouchier C, Buvat I. An [18F]FDG-PET/CT deep learning method for fully automated detection of pathological mediastinal lymph nodes in lung cancer patients. *Eur J Nucl Med Mol Imaging*. 2022;49(3):881-8.
34. Chamberlin J, Kocher MR, Waltz J, Snoddy M, Stringer NFC, Stephenson J, et al. Automated detection of lung nodules and coronary artery calcium using artificial intelligence on low-dose CT scans for lung cancer screening: accuracy and prognostic value. *BMC Med*. 2021;19(1):55.



Near Lake Wabby, K'Gari, Queensland, Australia



**Development and External Validation of
Automated Detection, Classification, and
Localization of Ankle Fractures Inside the Black
Box of a Convolutional Neural Network (CNN)**

J. Prijs

Z. Liao

M-S To

J.W. Verjans

P.C. Jutte

V.M.A. Stirler

J. Olczak

M. Gordon

D. Guss

C.W. DiGiovanni

R.L. Jaarsma

F.F.A. IJpma

J.N. Doornberg

on behalf of the Machine Learning Consortium

ABSTRACT

Purpose

Convolutional neural networks (CNNs) are increasingly being developed for automated fracture detection in orthopaedic trauma surgery. Studies to date, however, are limited to providing classification based on the entire image—and only produce heatmaps for approximate fracture localization instead of delineating exact fracture morphology. Therefore, we aimed to answer (1) what is the performance of a CNN that detects, classifies, localizes, and segments an ankle fracture, and (2) would this be externally valid?

Methods

The training set included 326 isolated fibula fractures and 423 non-fracture radiographs. The Detectron2 implementation of the Mask R-CNN was trained with labelled and annotated radiographs. The internal validation (or 'test set') and external validation sets consisted of 300 and 334 radiographs, respectively. Consensus agreement between three experienced fellowship-trained trauma surgeons was defined as the ground truth label. Diagnostic accuracy and area under the receiver operator characteristic curve (AUC) were used to assess classification performance. The Intersection over Union (IoU) was used to quantify accuracy of the segmentation predictions by the CNN, where a value of 0.5 is generally considered an adequate segmentation.

Results

The final CNN was able to classify fibula fractures according to four classes (Danis-Weber A, B, C and No Fracture) with AUC values ranging from 0.93 to 0.99. Diagnostic accuracy was 89% on the test set with average sensitivity of 89% and specificity of 96%. External validity was 89–90% accurate on a set of radiographs from a different hospital. Accuracies/ AUCs observed were 100/0.99 for the 'No Fracture' class, 92/0.99 for 'Weber B', 88/0.93 for 'Weber C', and 76/0.97 for 'Weber A'. For the fracture bounding box prediction by the CNN, a mean IoU of 0.65 (SD $\pm$ 0.16) was observed. The fracture segmentation predictions by the CNN resulted in a mean IoU of 0.47 (SD $\pm$ 0.17).

Conclusions

This study presents a look into the 'black box' of CNNs and represents the first automated delineation (segmentation) of fracture lines on (ankle) radiographs. The AUC values presented in this paper indicate good discriminatory capability of the CNN and substantiate further study of CNNs in detecting and classifying ankle fractures.

INTRODUCTION

Convolutional neural networks (CNNs) are increasingly being developed in orthopaedic trauma surgery for automated detection and classification of fractures.^{1–11} General benefits include the fact that they (a) do not suffer from mental or physical fatigue compared to clinicians, (b) are consistent in their assessment because they are not limited by surgeon bias or poor inter-surgeon reliability^{12–15}, and (c) can perform at or above the level of consensus agreement from a panel of experienced surgeons and radiologists.^{1, 5, 10, 11, 16} To date, most studies that have developed CNNs for fracture detection and classification primarily apply models that classify based on the entire or cropped input image.^{1, 5, 7, 10, 11, 16–18} In contrast, newer computer vision techniques can detect, segment (i.e. exact delineate the suggested location of the fracture (Fig. 5)), and classify fracture patterns.

Automated delineation of fracture lines gives us insight into what the algorithm ‘sees’, and may help foster clarity for the as yet ill-defined role of artificial intelligence (AI) in the field of computer vision for fracture recognition.^{19, 20} The next level of CNN studies in our field report detailed segmentation by a CNN of the second intact metacarpal²¹ on plain radiographs, vertebrae on computed tomography (CT)²², and femora on magnetic resonance imaging (MRI)^{23, 24}. To the best of our knowledge, however, detailed segmentation of fracture lines on radiographs has yet to be reported.

In 2020, Olczak and colleagues successfully applied a CNN for ankle fracture classification⁸ using the imagelevel classification model ResNet²⁵, but without automated delineation of the fracture. It remains the only fracture recognition paper for patients sustaining ankle trauma to date. Ideally, CNNs should combine object detection with segmentation, and thus offer localization and classification simultaneously—for example to better guide junior doctors during their early learning curves by presenting an exact visual outline of the fracture line itself. In addition, CNNs are often trained with large datasets without selecting cases that facilitate the most efficient training rate for the CNN (i.e. learning rate). This results in a large portion of unnecessarily labelled and/or annotated cases, because these contribute minimally to the performance of the model.

Therefore, we aimed to develop a CNN that detects (i.e. fracture yes/no), classifies (according to AO/OTA 44/Weber A, B and C²⁶), and localizes (with exact delineation/segmentation of an ankle fracture). The following questions will be answered in this paper: (1) What are the diagnostic performance characteristics (accuracy, sensitivity, specificity) and area under the receiver operator characteristic curve (AUC) of a CNN that classifies, localizes, and segments a lateral malleolus ankle fracture?, (2) Is this CNN externally valid?, and (3) Does application of a preliminary CNN that selects an appropriate training set result in an efficient training rate for the CNN?



MATERIALS AND METHODS

This study was approved by our Regional Review Board, according to the Declaration of Helsinki under number 13991.

Guidelines

This study was conducted according to the Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research²⁷ as well as the CONSORT-AI²⁸ the SPIRIT-AI²⁹, MI-CLAIM³⁰, and the CAIR checklist³¹.

Dataset

For this study, 12.000 radiographic ankle examinations with standard views (AP, Mortise and Lateral) were retrospectively collected from our Level 1 Trauma Centre, between January 2016 and December 2020. Studies were filtered using keywords in radiology reports to create an index database containing isolated fibular fractures and a non-fracture database (Fig. 1).

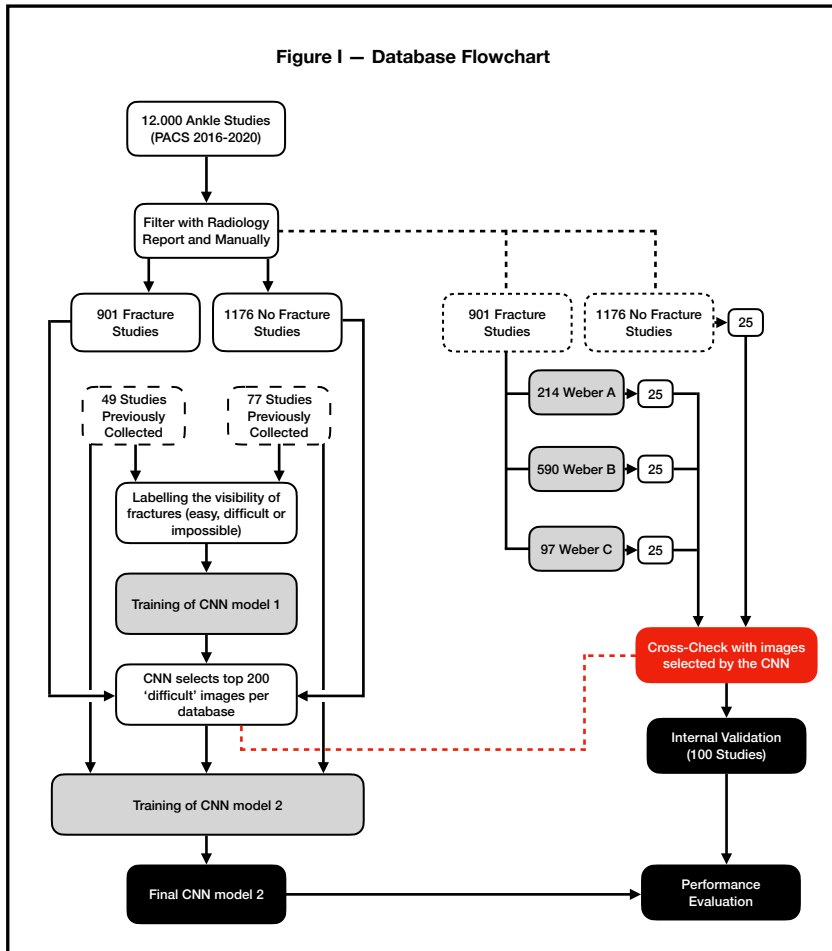


Fig. 1. Workflow used to create the final convoluted neural network (CNN) for the classification of ankle fractures. This involves a two-stage approach. An initial CNN was trained to select cases that were considered difficult—for example, fractures that were hard to appreciate—for classification. Subsequently, the final CNN was trained using these radiographs selected by the former CNN

Three independent observers manually reviewed and classified the radiographs according to the AO/OTA 44/ Weber A, B, and C²⁶, thereby excluding malleolar fractures where the tibia was involved. Any disagreements were resolved by discussion with a fourth independent senior observer. Data curation further excluded radiographs with fractures and pathology other than a fibular fracture, old fractures, presence of callous or cast, radiographs of poor quality (i.e. radiographs of patients that would be sent back to radiology in clinic), open physes, radiological views of insufficient quality, and presence of plates or screws.

Preliminary CNN model

To improve efficiency in labelling and segmentation, a preliminary CNN was trained to provide model-assisted labelling and annotations. For the annotation task, the DeepLab V3+³² architecture with MobileNet V2³³, pre-trained on ImageNet³⁴ data, was used. For the classification task, a separate CNN with a MobileNet V2³³ backbone and a softmax classifier were used. Training data for the preliminary CNN consisted of 147 radiographs with a fracture and 228 without a fracture. Using Labelbox³⁵, the images were manually labelled for visibility of the fracture (easy, difficult, or impossible) and annotated by two independent observers for the following: shape of tibia/fibula and fracture. Bounding boxes were created around the borders of these respective annotations.

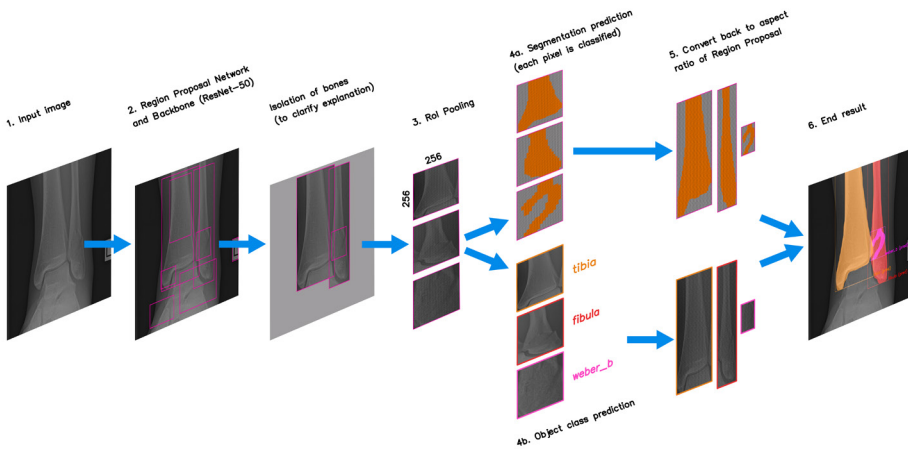


Fig. 2 This figure presents how the final convoluted neural network (CNN) goes from the input image (1) to the final prediction (6). The region proposal network and backbone create countless bounding boxes (2), where each box has a high likelihood of the presence of an object. Then, the region of interest (RoI) crops the bounding boxes to fit fixed dimensions, in this case 256x256 pixels (3). These cropped images are then used to simultaneously segment (4a) and classify (4b). Finally, the cropped images are then resized to their original dimensions (5) and presented on top of the input image as predictions (6)

Final CNN model (Fig. 2)

For final model development, the Detectron2³⁶ implementation of the Mask R-CNN was used. The backbone of the Mask R-CNN model was set to the Microsoft Research Asia version ResNet-50³⁷, pre-trained on ImageNet³⁴. The ResNet-101 variation of the backbone was tested but did not result in significant improvement. An instance segmentation model can segment individual objects (i.e. bones) by combining object detection (bounding box) and semantic segmentation (Fig. 3). The simplified explanation order in which Mask R-CNN does this is as follows (Fig. 2): (1) The radiograph is fed into the CNN; (2) the backbone (ResNet-50) together

with the Region Proposal Network (RPN) creates many bounding boxes with each proposal being an object; (3) each region proposal is resized by Region of Interest (RoI) pooling to fit fixed height and width dimensions of 256×256 ; (4a) Mask R-CNN classifies each pixel in a region proposal to create a segmentation; (4b) simultaneously, Mask R-CNN uses object class prediction on each region proposal; (5) predictions are reverted back to original height and width dimensions and projected onto the output image.

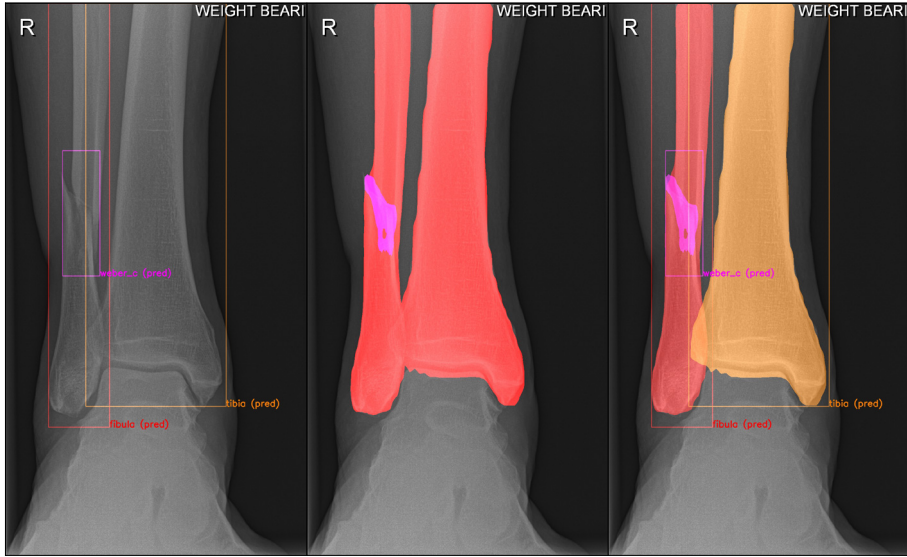


Fig. 3 From left to right: Object detection, semantic segmentation, and instance segmentation

Training of final CNN model

The training set included 326 fracture and 423 non-fracture radiographs, which were labelled and annotated in Labelbox.³⁵ Standard data augmentation (random cropping and horizontal flip operations) was used to improve the generalization of the model. To reduce bias, cases were re-weighted according to their prevalence. The annotated bounding boxes are used as the Ground Truth for the RPN. Training was completed at 64 epochs (64 iterations of the complete dataset) after 90 min. The training starts from an initial learning rate of 0.05 down to 1/10 every 1000 steps. Each step is commonly known as a mini-batch iteration; in this study, we loaded 12 images per mini-batch.

Evaluation of final CNN model

Twenty-five patients of each class (AO/OTA 44/Weber A, B, C or No Fracture), were randomly selected by the computer—and cross-checked with the 400 (by the preliminary CNN) selected 'difficult' images—for the internal validation set (also known as 'test set'), to assess the



patient-level accuracy. The final prediction was the class with the highest combined prediction value among all radiographic views. The ground truth was the consensus between three experienced fellowship-trained trauma surgeons. Consensus was achieved on all cases; however, ambiguous cases (low inter-observer agreement) were put in a clinically challenging set and swapped with randomly selected patients to ensure objective measurement (high inter-observer agreement) of model performance. After assessing performance using nonambiguous cases (clinically easy internal validation), these were then put back into the internal validation set to assess the effect of clinically challenging cases (clinically challenging internal validation) on performance of the model.

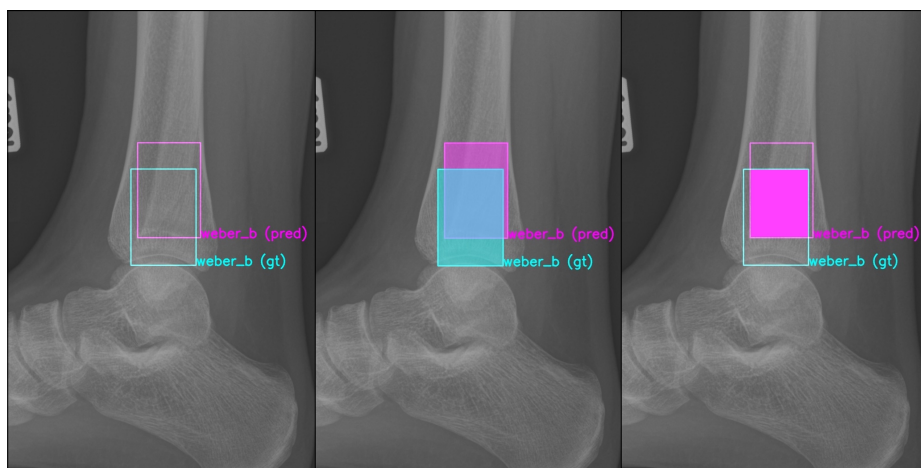


Fig. 4 From left to right: Ground truth (gt) versus prediction (pred), area of union (gt+pred), and area of overlap

To assess transportability and generalizability of the model, external validation was performed using 167 cases from our second Level-1 Trauma hospital in the Netherlands with the same methodology as for the internal validation.

Due to a difference in protocol, these did not contain mortise views.

For assessment of the image-level accuracy, Intersection over Union (IoU; also known as the Jaccard index, Fig. 4) was used to quantify accuracy of the segmentation predictions by the CNN, where 0 indicates no overlap at all and 1 a perfect overlap. Due to the complex nature of CNNs, it cannot be assumed that predictions will perfectly match the ground truth; therefore, IoU is used as it is an indicator of overlap. Generally, an IoU > 0.5 is considered a good prediction.^{38, 39}

Statistical analysis

Accuracy (defined as the percentage of cases correctly predicted by the CNN), sensitivity, specificity, and AUC were calculated for each (non) fracture class. The AUC reflects the discriminative ability of the CNN to separate classes, an AUC of 1.0 corresponds to a prediction with perfect discriminatory performance, whereas 0.5 indicates a prediction equal to chance.

Statistical analysis was performed using Python 3.9.0 [Python Software Foundation, Beaverton, United States] with the modules: pandas, cv2, NumPy, sklearn, and plotly.

RESULTS

Test set—performance of CNN model 2 on clinically ‘Easy’ cases (Fig. 5)

The final CNN was able to classify fibula fractures according to four classes (Danis-Weber A, B, C and No Fracture) with AUC values ranging from 0.93–0.99 and 89% accuracy (Tables 1, 2 and Fig. 5). Best accuracy was observed for the ‘No Fracture’ class with 100% and ‘Weber B’ with 92%. Accuracies of 88% and 76% were observed for classes ‘Weber C’ and ‘Weber A’, respectively. Specificity, however, was 100% for both of those two fracture classes. In the ‘Weber C’ group, three cases were misclassified of which two were subtle fractures that were picked up on the lateral radiograph but missed or misclassified as ‘Weber B’ on the anteroposterior and mortise views, and one was a steep oblique fracture line misclassified as ‘Weber B’. From the ‘Weber A’ group, the six patients that were misclassified, five had transverse fractures at the level of the ankle joint (the line between ‘A’ or ‘B’ classification), and one was a subtle fracture. Two examples of misclassifications are shown in Fig. 6.

Test set—performance of CNN model 2 on clinically ‘Difficult’ cases

With reintroduction of ambiguous cases, AUC values ranged from 0.90 to 0.98 and accuracy decreased by 4% to an average of 85%. Performance metrics per class are given in Tables 1 and 2. Besides minor changes in other classes, the ‘Weber C’ class was most affected, where accuracy decreased from 88% (22/25) to 72% (18/25). Compared with the initial internal validation set, the ‘Weber C’ class had two extra ‘Weber B’ misclassifications. These occurred with a steep oblique fracture line, and two extra misclassifications as ‘No Fracture’ occurred when there was a presence of high ‘Weber C’ fracture.



Table I — Combined Radiograph Confusion Matrix and Accuracy

Performance on Clinically 'Easy' Cases					
	Predicted				Accuracy
	Weber A	Weber B	Weber C	No-Fracture	
Weber A	19	3	0	3	76%
Weber B	0	23	0	2	92%
Weber C	0	2	22	1	88%
No-Fracture	0	0	0	25	100%

Performance on Clinically 'Difficult' Cases					
	Predicted				Accuracy
	Weber A	Weber B	Weber C	No-Fracture	
Weber A	19	2	0	4	76%
Weber B	0	23	0	2	92%
Weber C	0	4	18	3	72%
No-Fracture	0	0	0	25	100%

Table II — Sensitivity, Specificity and AUC per class

Performance on Clinically Easy Cases			
	Sensitivity	Specificity	AUC
Weber A	76%	100%	0,93
Weber B	92%	93%	0,97
Weber C	88%	100%	0,99
No-Fracture	100%	92%	0,99

Performance on Clinically Difficult Cases			
	Sensitivity	Specificity	AUC
Weber A	76%	100%	0,93
Weber B	92%	93%	0,97
Weber C	72%	100%	0,9
No-Fracture	100%	88%	0,98

Test set—accuracy of segmentation (i.e. delineation of the fracture line) (Fig. 7)

Quality of the predicted fracture segmentations by the CNN was quantified by the IoU (Figs. 2 and 7). For the fracture bounding box prediction by the CNN, a mean IoU of 0.65 (SD±0.16) was observed. The much more challenging fracture polygon segmentation predictions by the CNN resulted in a mean IoU of 0.47 (SD ± 0.17).

External validation of CNN model 2

On the clinically 'easy' external validation set without ambiguous cases, the model achieved AUC values ranging from 0.83 to 0.95 and an overall accuracy of 90% (Table 3).

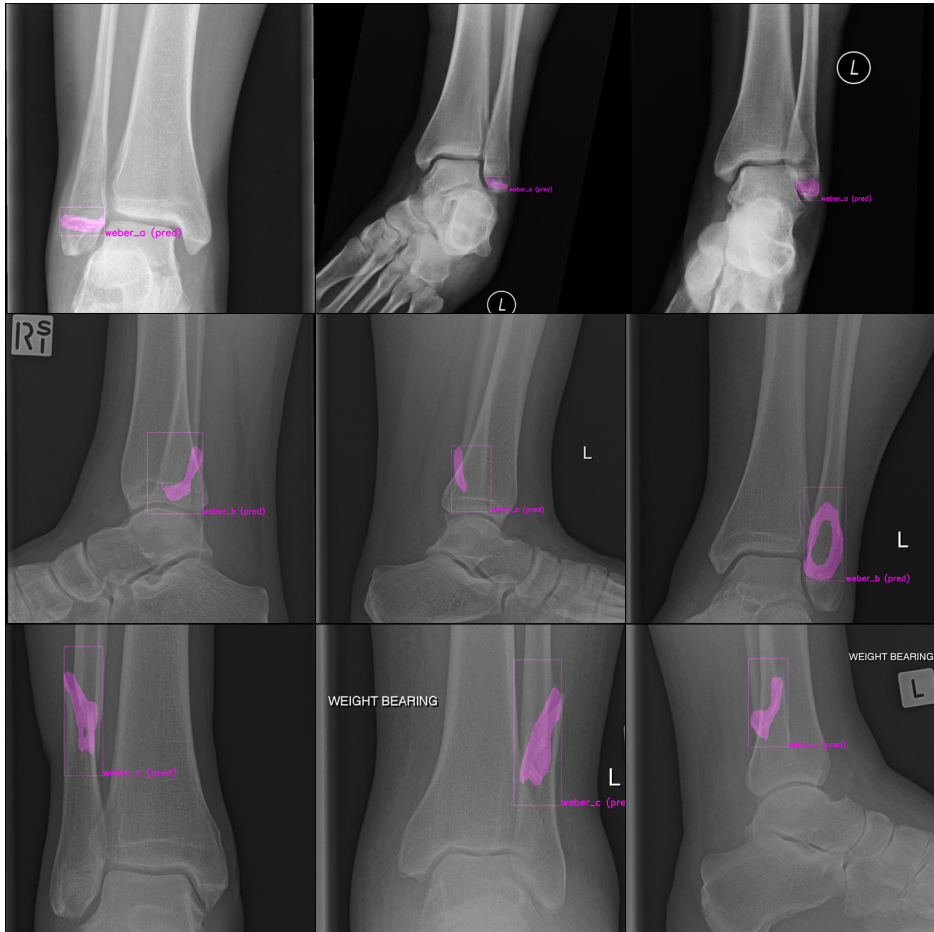


Fig. 5 Selection of correct classifications by the final convoluted neural network

Best accuracies of 99% and 92% were achieved for the 'No Fracture' and 'Weber B' classes, respectively, while the 'Weber C' and 'Weber A' classes resulted in the least accurate predictions with 71% and 64%, respectively.

When ambiguous cases were introduced, the model achieved AUC values ranging from 0.84 to 0.92 and accuracy to 89%. Compared to the former external validation, all fracture classes' accuracies were affected by 1–2% (Table 3). A similar pattern to the internal validation was observed; the model struggled with ambiguous cases; however, in contrast to the internal validation, no specific class was more affected than others.

Table III — External Validation Performance

Clinically 'Easy' External Validation				
	Accuracy	Sensitivity	Specificity	AUC
Weber A	64%	64%	100%	0,88
Weber B	92%	92%	98%	0,98
Weber C	71%	71%	100%	0,83
No-Fracture	99%	99%	86%	0,9

Clinically 'Difficult' External Validation				
	Accuracy	Sensitivity	Specificity	AUC
Weber A	62%	62%	100%	0,86
Weber B	90,0%	90,0%	96%	0,92
Weber C	70,0%	70,0%	100%	0,84
No-Fracture	99%	99%	89%	0,89

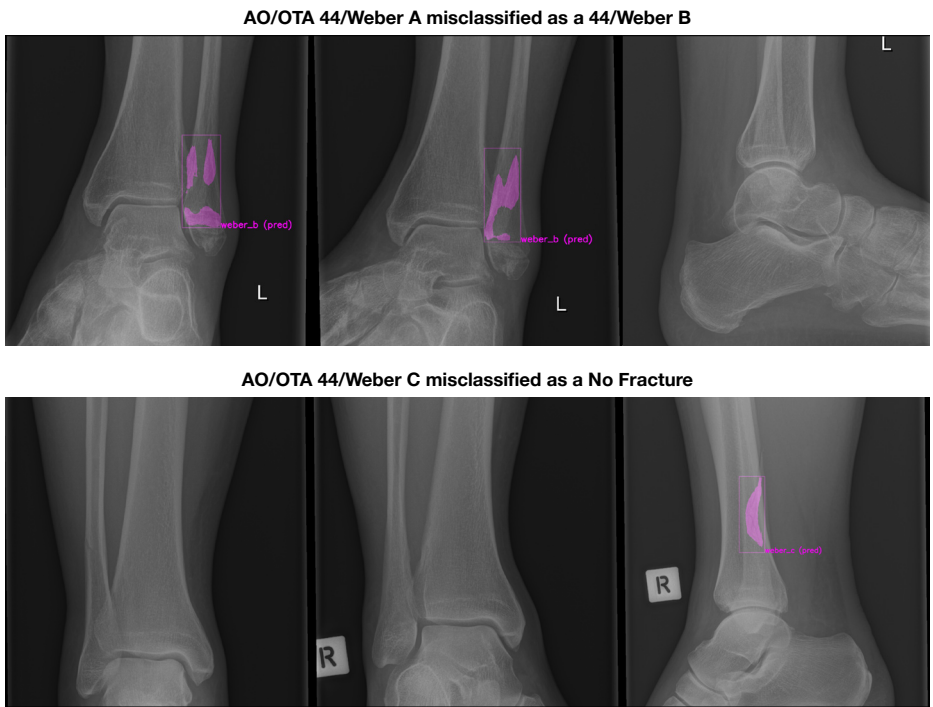


Fig. 6 AO/OTA 44/Weber A misclassified as a 44/Weber B, AO/OTA 44/Weber C misclassified as a No Fracture

DISCUSSION

To date, studies on the clinical application of AI in the field of computer vision have not deployed CNNs to automatically delineate fractures, which can reduce the black box effect as well as aide less experienced doctors who are still in their early learning curve. Moreover, external validity of current CNNs for fracture recognition in orthopaedic trauma is scarce.³ In this study, we developed a CNN that can detect, classify, and create detailed segmentations of fracture lines in ankle fractures (AO/OTA 44/Weber A, B and C) with an overall accuracy of 89%. In addition, it was found to be externally valid on radiographs from Level I Trauma Centre on a different continent, with an average accuracy of 89–90%. We used a preliminary CNN to select fractures that were difficult to appreciate, aiming for the most efficient training rate per image for the final CNN.

As with any study, this information must be interpreted with respect to its scientific strengths and weaknesses. One limitation is that existing classification systems suffer from varying inter-observer reliability, affecting performance of the model.¹⁴ Another stems from evaluating the CNN using retrospective instead of prospective data, although the internal and external validation were collected from multiple years and thus simulate clinical practice. Also, as training data did not include fractures with concomitant joint dislocations, the CNN is unlikely to recognize this significant fracture subset. These study shortcomings, however, are counterbalanced by several notable merits, including that this is the first paper in the field of orthopaedic trauma to describe a pixel perfect segmentation of fracture lines on plain radiographs—compared to rough predictions using heat/activation maps described in literature—and use a preliminary CNN to select cases to train the final CNN model. Another strength is external validation of the CNN in assessing generalizability and possible bias of the model on data different than that used for development. Moreover, labelling of the internal validation set was done by three independent experienced surgeons, and a consensus was used as the ground truth. Finally, the Mask R-CNN [40] used in this investigation represents a state-of-the-art CNN that accepts entire radiographs as input image for the detection, classification, and segmentation tasks, whereas commonly used CNN models often warrant cropping and are more difficult to implement in clinical practice.

In 2020, Olczak et al. [8] were the first and the only ones thus far to report the use of a CNN in classifying ankle fractures. The current study adds to our knowledge by presenting a CNN that was developed to create a detailed localization and segmentation of fracture lines on radiographs. This may improve clinical reasoning and diagnostics by giving junior clinicians a visual guide and simultaneously reduces the ominous ‘black box’ effect, which facilitates a feedback loop for an ongoing learning curve. Furthermore, this study reports an improvement in the discriminatory performance (AUC values) compared to the study from Olczak and colleagues [8], using less than a fourth (approximately 250 cases) of the 1064 cases for the corresponding classes without tibial involvement (AO/OTA 44A1, B1, B2.1 and C1.1).



Accuracy cannot be compared as the latter study did not report an accuracy of their CNN in classifying ankle fractures. As labelling and annotating is very labour-intensive and qualified experts' time is often limited, training an initial CNN to select optimal cases for training the CNN increases efficiency.

Although accuracy and the AUCs were high, the CNN misclassified 11 out of 100 patients in our test set (Figs. 5, 6, and 7). It should be noted, however, that the AUC is close to 1 (indicating almost perfect discriminatory performance), ranging from 0.93 to 0.99. Accuracy depends greatly on individual cases in the internal validation set. Interestingly, the same cases considered ambiguous by surgeons (i.e. poor inter-observer reliability) were also the ones that CNN had difficulties with. Since CNNs can only be as good as their training, it should be noted that without an absolute truth (e.g. a CT scan), current computers can only be trained to approach the performance of surgeons—but they cannot surpass it. If one defines the ground truth as a consensus agreement, however, at least some inherent surgeon bias can be eliminated. The AO/OTA 44/Weber A and C classifications were most susceptible for misclassification, together accounting for nine of the 11 errors. The recognition of higher Weber C-type injuries might be limited by the fact that the CNN appears to have no positional awareness and seems to classify based purely on fracture configuration. Another explanation might be the alternative CNN shortcoming in trying to detect features that exist at the margin of an image; similarly, CNNs are likely dependent on the variability in what gets captured in a given radiograph, since for multiple reasons this clearly varies image to image. Since Weber C injuries can have dynamic instability or exist at a level not identified on non-stressed or more limited exposure radiographs, it makes sense that this is where these algorithms seems to fall short and demonstrate room for improvement. For the segmentation task, the average IoU value for the bounding boxes was good, even though one study suggests that IoU is optimal for round shapes, but not for elongated ones⁴¹ such as those used in this study. As expected, the highly variable fracture line segmentations resulted in a lower IoU compared to the bounding box. However, the average IoU was still close to 0.5, suggesting an overall accurate fracture line segmentation despite the great variation in fracture configurations.

It is worth noting that accuracy was the highest when detecting a lack of fracture, doing so with 100% accuracy. Thus, while much of the efforts of this study were to distinguish between various fracture patterns, the ability of the same process to exclude fracture is inherently useful to those working in the emergency or urgent care setting who simply need guidance as to when to seek additional orthopaedic consultation.

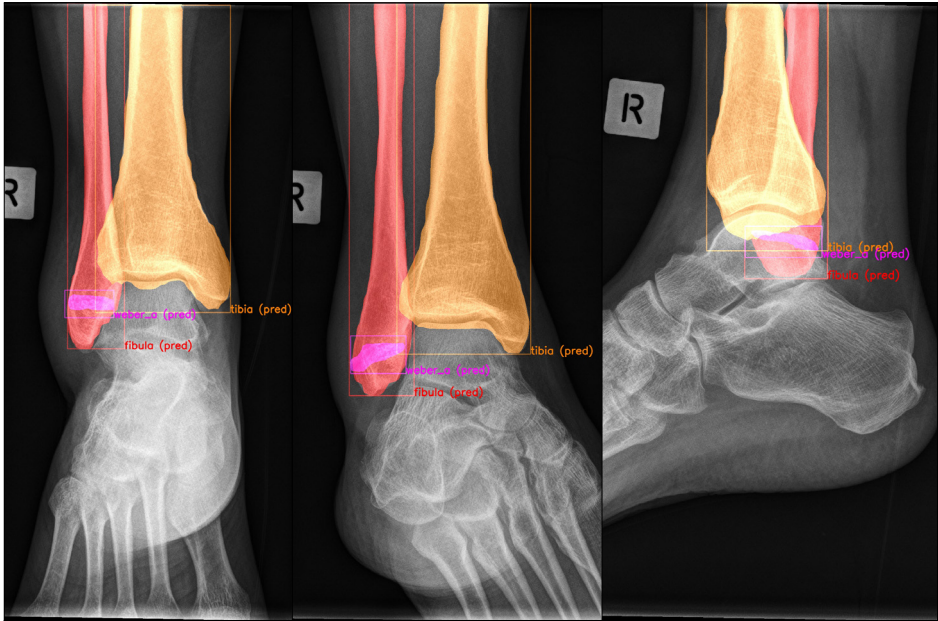


Fig. 7 Segmentations and classifications of the final convoluted neural network for AO/OTA 44/Weber A (top), B (middle), and C (bottom)

Performance is often reduced when assessed with an external validation set^{4, 16, 42}, as there are many observer and machine-dependent variances between hospitals. Therefore, geographical external validation is a stringent and crucial test towards clinical implementation of these models. Enabling the external validation set to usually have two views (anteroposterior and lateral) instead of three (mortise) improved classification of fractures that were only detected on one of the views, with a caveat that the CNN has to be more confident of its 'Fracture' classification than its 'No Fracture' classification. A notable distinction here is that when there were three views and the fracture was seen on the lateral view but not on the anteroposterior and mortise, it was always classified as 'No Fracture'.

In summary, this early work on automated detection in orthopaedic imaging demonstrates remarkable future potential despite several shortcomings noted in its current level of development. In conclusion, even though object detection has been employed for certain other types of fractures and imaging modalities, this study presents the first automated segmentation of fracture lines on ankle radiographs. The accuracy and AUC values presented in this paper certainly fortify a role for CNNs in detecting and classifying ankle fractures. Moreover, using a preliminary CNN to identify cases resulted in a network that was accurate enough to be externally valid in another hospital, surely important for reducing the workload of creating high-quality data for training of CNNs.



ACKNOWLEDGEMENTS

No funding has been received for this study.

On behalf of the Machine Learning Consortium: Kaan Aksakal, Britt Barvelink, Michel van der Bekerom, Benn Beuker, Anne Eva Bultra, Luisa Oliviera e Carmo, Joost Colaris, Huub de Klerk, Andrew Duckworth, Kaj ten Duis, Eelco Fennema, Jorrit Harbers, Ran Hendrickx, Marilyn Heng, Sanne Hoeksema, Mike Hogervorst, Bhavin Jadav, Julie Jiang, Aditya Karhade, Gino Kerkhoffs, Joost Kuipers, Charlotte Laane, David Langerhuizen, Bart Lubberts, Wouter Mallee, Haras Mhmud, Mostafa El Moumni, Patrick Nieboer, Koen Oude Nijhuis, Peter van Ooijen, Jacobien Oosterhoff, Jai Rawat, David Ring, Sanne Schilstra, Joseph Schwab, Sheila Sprague, Sjoerd Stufkens, Elvira Tijdens, Puck van der Vet, Jean-Paul de Vries, Klaus Wendt, Matthieu Wijffels, David Worsley.

REFERENCES

1. Adams M, Chen W, Holcldorf D, McCusker MW, Howe PD, Gaillard F. Computer vs human: Deep learning versus perceptual training for the detection of neck of femur fractures. *J Med Imaging Radiat Oncol*. 2019;63(1):27–32.
2. Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, McConnell MV, Percha B, Snyder TM, Dudley JT. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med*. 2019;2:31.
3. Oliveira ECL, van den Merkhof A, Olczak J, Gordon M, Jutte PC, Jaarsma RL, Ijpma FFA, Doornberg JN, Prijs J. An increasing number of convolutional neural networks for fracture recognition and classification in orthopaedics: are these externally validated and ready for clinical application? *Bone Jt Open*. 2021;2(10):879–85.
4. Choi JW, Cho YJ, Lee S, Lee J, Lee S, Choi YH, Cheon J-E, Ha JY. Using a dual-input convolutional neural network for automated detection of pediatric supracondylar fracture on conventional radiography. *Invest Radiol*. 2020;55(2):101–10.
5. Chung SW, Han SS, Lee JW, Oh KS, Kim NR, Yoon JP, Kim JY, Moon SH, Kwon J, Lee HJ, Noh YM, Kim Y. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop*. 2018;89(4):468–73.
6. Kitamura G, Chung CY, Moore BE 2nd. Ankle fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, de novo training, and multiview incorporation. *J Digit Imaging*. 2019;32(4):672–7.
7. Langerhuizen DWG, Bulstra AEJ, Janssen SJ, Ring D, Kerkhoffs G, Jaarsma RL, Doornberg JN. Is deep learning on par with human observers for detection of radiographically visible and occult fractures of the scaphoid? *Clin Orthop Relat Res*. 2020;478(11):2653–9.
8. Olczak J, Emilson F, Razavian A, Antonsson T, Stark A, Gordon M. Ankle fracture classification using deep learning: automating detailed AO Foundation/Orthopedic Trauma Association (AO/OTA) 2018 malleolar fracture identification reaches a high degree of correct classification. *Acta Orthop*. 2021;92(1):102–108. <https://doi.org/10.1080/17453674.2020.1837420>
9. Oosterhoff JHF, Doornberg JN, Machine Learning C. Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of Gartner's hype cycle. *EFORT Open Rev*. 2020;5(10):593–603.
10. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedistlevel accuracy using a deep convolutional neural network. *Skeletal Radiol*. 2019;48(2):239–44.
11. Yamada Y, Maki S, Kishida S, et al. Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: ensemble decision-making with antero-posterior and lateral radiographs. *Acta Orthop*. 2020;91(6):699–704. <https://doi.org/10.1080/17453674.2020.1803664>.
12. Mellema JJ, Doornberg JN, Molenaars RJ, Ring D, Kloen P, Traumaplatform Study C, and Science of Variation G. Tibial plateau fracture characteristics: reliability and diagnostic accuracy. *J Orthop Trauma*. 2016;30(5):e144–51.
13. Mellema JJ, Doornberg JN, Molenaars RJ, Ring D, Kloen P, C. Traumaplatform Study C, and G. Interobserver reliability of the Schatzker and Luo classification systems for tibial plateau fractures. *Injury*. 2016;47(4):944–9.
14. Malek IA, Machani B, Mevcha AM, Hyder NH. Interobserver reliability and intra-observer reproducibility of the Weber classification of ankle fractures. *J Bone Jt Surg Br*. 2006;88(9):1204–6.



15. Yin MC, Yuan XF, Ma JM, Xia Y, Wang T, Xu XL, Yan YJ, Xu JH, Ye J, Tong ZY, Feng YQ, Wang HB, Wu XQ, Mo W. Evaluating the reliability and reproducibility of the AO and LaugeHansen classification systems for ankle injuries. *Orthopedics*. 2015;38(7):e626–30.
16. Blüthgen C, Becker AS, Vittoria de Martini I, Meier A, Martini K, Frauenfelder T. Detection and localization of distal radius fractures: Deep learning system versus radiologists. *Eur J Radiol*. 2020;126:108925.
17. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, Hanel D, Gardner M, Gupta A, Hotchkiss R, Potter H. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci*. 2018;115(45):11591–6.
18. Lee C, Jang J, Lee S, Kim YS, Jo HJ, Kim Y. Classification of femur fracture in pelvic X-ray images using meta-learned deep neural network. *Sci Rep*. 2020;10(1):13694.
19. Liu KL, Wu T, Chen PT, Tsai YM, Roth H, Wu MS, Liao WC, Wang W. Deep learning to distinguish pancreatic cancer tissue from non-cancerous pancreatic tissue: a retrospective study with cross-racial external validation. *Lancet Digit Health*. 2020;2(6):e303–13.
20. Myers TG, Ramkumar PN, Ricciardi BF, Urish KL, Kipper J, Ketonis C. Artificial intelligence and orthopaedics: an introduction for clinicians. *J Bone Jt Surg Am*. 2020;102(9):830–40.
21. Tecle N, Teitel J, Morris MR, Sani N, Mitten D, Hammert WC. Convolutional neural network for second metacarpal radiographic osteoporosis screening. *J Hand Surg Am*. 2020;45(3):175–81.
22. Löffler MT, Jacob A, Scharr A, et al. Automatic opportunistic osteoporosis screening in routine CT: improved prediction of patients with prevalent vertebral fractures compared to DXA. *Eur Radiol*. 2021;31(8):6069–6077. <https://doi.org/10.1007/s00330-020-07655-2>.
23. Deniz CM, Xiang S, Hallyburton RS, Welbeck A, Babb JS, Honig S, Cho K, Chang G. Segmentation of the proximal femur from MR images using deep convolutional neural networks. *Sci Rep*. 2018;8(1):16485.
24. Memis A, Varli S, Bilgili F. Semantic segmentation of the multiform proximal femur and femoral head bones with the deep convolutional neural networks in low quality MRI sections acquired in different MRI protocols. *Comput Med Imaging Graph*. 2020;81: 101715.
25. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.
26. Meinberg EG, Agel J, Roberts CS, Karam MD, Kellam JF. Fracture and dislocation classification compendium-2018. *J Orthop Trauma*. 2018;32(Suppl 1):S1–170.
27. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S, Berk M. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. 2016;18(12): e323.
28. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Spirit AI, Group C-AW. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364–74.
29. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, Spirit AI, Group C-AW, Group C-AS, Group C-AC. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351–63.
30. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, Arnaout R, Kohane IS, Saria S, Topol E, Obermeyer Z, Yu B, Butte AJ. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320–4.
31. Olczak J, Pavlopoulos J, Priejs J, et al. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. *Acta Orthop*. 2021;92(5):513–525. <https://doi.org/10.1080/17453674.2021.1918389>.
32. Chen L, Papandreou G, Schroff F, Adam H (2017) Rethinking Atrous Convolution for Semantic Image Segmentation. *arxiv*: 1706.05587.

33. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition, 2018; p. 4510–520.
34. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, p. 248–255.
35. Labelbox, "Labelbox," Online, 2022. [Online]. Available: <https://labelbox.com>. Accessed 2021.
36. Wu Y, Kirillov A, Massa F, Lo W-Y, Girshick R. Detectron2. 2019.
37. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016; p. 770–78.
38. Zhu W, Liu C, Fan W, Xie X. DeepLung: 3D Deep Convolutional Nets for Automated Pulmonary Nodule Detection and Classification. bioRxiv. 2017.
39. Qi Y, Zhao J, Shi Y, Zuo G, Zhang H, Long Y, Wang F, Wang W. Ground truth annotated femoral X-ray image dataset and object detection based method for fracture types classification. IEEE Access. 2020;8:189436–44.
40. He K, Gkioxari G, Doll-PR, Girshick RB. Mask R-CNN. In: 2017 IEEE International Conference on computer vision (ICCV), 2017;p. 2980–988.
41. JinL,YangJ,KuangK,NiB,GaoY,SunY,GaoP,MaW,Tan
42. M, Kang H, Chen J, Li M. Deep-learning-assisted detection and segmentation of rib fractures from CT scans: development and validation of FracNet. EBioMedicine. 2020;62:103106.
43. Zhou Q-Q, Wang J, Tang W, Hu Z-C, Xia Z-Y, Li X-S, Zhang R, Yin X, Zhang B, Zhang H. Automatic detection and classification of rib fractures on thoracic CT using convolutional neural network: accuracy and feasibility. Korean J Radiol. 2020;21(7):869–79.



ABSTRACT

Background

Advances in medical imaging have made it possible to classify ankle fractures using Artificial Intelligence (AI). Recent studies have demonstrated good internal validity for machine learning algorithms using the AO/OTA 2018 classification. This study aimed to externally validate one such model for ankle fracture classification and ways to improve external validity.

Methods

In this retrospective observation study, we trained a deep-learning neural network (7,500 ankle studies) to classify traumatic malleolar fractures according to the AO/OTA classification. Our internal validation dataset (IVD) contained 409 studies collected from Danderyd Hospital in Stockholm, Sweden, between 2002 and 2016. The external validation dataset (EVD) contained 399 studies collected from Flinders Medical Centre, Adelaide, Australia, between 2016 and 2020. Our primary outcome measures were the area under the receiver operating characteristic (AUC) and the area under the precision-recall curve (AUPR) for fracture classification of AO/OTA malleolar (44) fractures. Secondary outcomes were performance on other fractures visible on ankle radiographs and inter-observer reliability of reviewers.

Results

Compared to the weighted mean AUC (wAUC) 0.86 (95%CI 0.82–0.89) for fracture detection in the EVD, the network attained wAUC 0.95 (95%CI 0.94–0.97) for the IVD. The area under the precision-recall curve (AUPR) was 0.93 vs. 0.96. The wAUC for individual outcomes (type 44A-C, group 44A1-C3, and subgroup 44A1.1-C3.3) was 0.82 for the EVD and 0.93 for the IVD. The weighted mean AUPR (wAUPR) was 0.59 vs 0.63. Throughout, the performance was superior to that of a random classifier for the EVD.

Conclusion

Although the two datasets had considerable differences, the model transferred well to the EVD and the alternative clinical scenario it represents. The direct clinical implications of this study are that algorithms developed elsewhere need local validation and that discrepancies can be rectified using targeted training. In a wider sense, we believe this opens up possibilities for building advanced treatment recommendations based on exact fracture types that are more objective than current clinical decisions, often influenced by who is present during rounds.

INTRODUCTION

With artificial intelligence's (AI) growing success in image analysis, AI interventions are rapidly being developed and applied in medical diagnostics.¹ Many studies have reported promising results, reaching close to perfect accuracy on basic pathology detection tasks, illustrating that accuracy in elementary pathology detection should be relatively easy to attain. The promise of AI interventions lies in their ability to solve complex tasks in different scenarios. For example, classifying fractures into meaningful features could give clinical guidance, drive treatment decision-making, or predict clinical outcomes. However, as researchers develop models under controlled conditions, few have reproduced their results.

A meta-analysis by Liu et al. reported the lack of external validation in deep learning [2]. For example, a systematic meta-analysis by Oliveira e Carmo et al. found 36 papers using deep learning for orthopedics. Only three were externally validated, i.e., tested on independent data from a different site^{3,4,5,6} (See Supplement 1, Table S1.) Similarly, a systematic review of orthopedic machine learning models predicting surgery outcomes by Groot et al. found that only 10/59 studies had externally validated their models.⁷ There are currently many initiatives to improve the quality of reporting AI studies in medicine, for example, via checklists for consistent and relevant reporting and external validation.^{1, 8} For a predictive model to be helpful, it must work and be tested in clinical environments other than what the model has been trained on – also called external validation – and thus be generally applicable.

There are three major classification systems for ankle fractures. Previously, we showed that deep-learning models can classify ankle fractures according to the AO Foundation/Orthopedic Trauma Association (AO/OTA) classification. The AO/OTA standard classifies fractures based on their visual appearance in radiographic examinations, making it well-suited to AI image classification. This classification is influenced by the very popular Lauge-Hansen (LH) system, which is widely used in clinical practice and categorizes fractures based on the injury mechanism. The LH system's reliance on such non-visual factors presents challenges for this study, where the injury mechanisms were missing. At the same time, the AO/OTA classification can be seen as an extension of the Danis-Weber classification.

We have previously reached a model performance of weighted average area under the receiver operating characteristic curve (AUC and wAUC) 0.90 (95%CI 0.82–0.94)⁹ using internal validation test data – data from the same site as the test data. Given that such a model aims to facilitate classification and decision-making in an emergency setting, we needed to validate its performance in the clinic, not in the training setting. This paper examines the external validation of an AI model for classifying ankle fractures according to the AO/OTA standard. External validation consists of applying a model to independent data from a site different from the one used for training. It aims to see how relevant and generalizable a model is in a clinical context. Our primary aim was to study the effect of transferring a model to a different setting,



i.e., the model's external validity, and to study ways to improve the external validity of a machine learning model. Our secondary aim was to explore the AO/OTA classification more broadly.

MATERIAL AND METHODS

The study was a retrospective external validation cohort study.

Ethics approval and consent to participate

Ethical approval for the collection of Flinders/external validation dataset was obtained from the Central Adelaide Local Health Network Human Research Ethics Committee (CALHN HREC) reference number: 13991, Authorization date: 21 December 2020. In accordance with the ethical permit, no individual or informed consent from participants was required. In accordance with the specific consent for inclusion into this study, the data was not considered patient data.

Ethical approval for the Danderyd/internal validation dataset was obtained by the Regional Ethics Committee for Stockholm, Sweden (Dnr. 2014/453-31/3, April 9, 2014). According to the ethical approval, no individual or informed consent from participants was required, as the data did not constitute human data after anonymous collection.

Ethical approval to use the external dataset for this study was also obtained from the Swedish Ethical Review Authority, Sweden (Dnr. 2023-07151-01).

The need for informed consent for the use of the data for the study waived. The data was anonymous radiographs without personal identifiable information, it would not be possible to identify individuals and informed consent was waived.

Data collection and pre-processing

Training and internal validation dataset (IVD)

The training data came from a retrospective cohort of trauma radiographs (initial imaging performed at the emergency department at the time of injury) collected from Danderyd University Hospital (Stockholm, Sweden) between 2002 and 2015. This level 2 trauma center had a referral area of approximately 350,000 people during that period. The data was collected anonymously and only coded with a unique patient identifier, but the radiologist report was included. No injury or population data (such as age, gender, trauma mechanism, etc.) was collected. We used the radiologist reports to generate initial fracture/no fracture labels. These labels have been improved over successive studies through manual review by radiologists and orthopedic consultants.^{9, 10, 11} All examinations visualizing the ankle were included, and only pediatric studies (open physis) were excluded because they are classified differently.

Four hundred patients (409 exams, including all available views visualizing the ankle) were previously selected for the internal validation dataset (IVD). Our previous study had a 2/3 bias towards fractures in the IVD to ensure sufficient fractures to classify and compare rarer

fractures. We did not specify the fracture type, so tibia, malleolus, fibula, or foot fractures were included.^{9,10} As part of active training, we added 2664 fractures to the training dataset to align it more with the EVD distribution. We used model-based selection, i.e., the model screened ankle studies from the Danderyd dataset and selected cases where the model flagged categories of interest or where the probability for the predicted class was low, i.e., had high uncertainty. These were then manually reclassified. No patient was present in both the training and IVD set. See Fig. 1. For training details, see Supplement 2, available online.

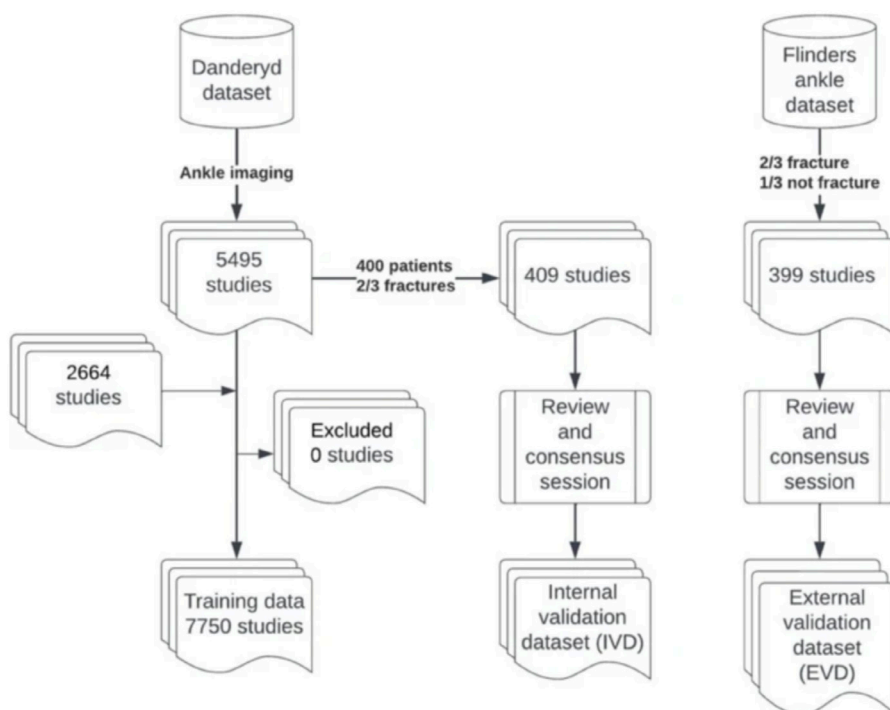


Fig. 1 Study flowchart

External validation dataset (EVD)

The external validation dataset (EVD) was a subset of 12,000 radiographic ankle examinations collected from Flinders University Medical Centre (Adelaide, Australia), a level 1 Trauma Centre, between 2016 and 2020.

Studies were filtered using keywords in radiology reports to create an index database containing isolated fibular and lateral malleolus fractures and a non-fracture database. While only trauma radiographs were included, this included one-week follow-ups and weight-bearing images. Projections were three standard views (AP, mortise, and lateral). Exclusion criteria were any pathology other than a fibula or lateral malleolus fracture, old fractures, callous

or cast presence, radiographs of poor quality, open physes, radiological views of insufficient quality, and the occurrence of plates or screws were also excluded.

Three hundred ninety-nine examinations were randomly selected from the Flinders dataset and provided as an external validation dataset (EVD) for this study, with a 2/3 selection bias towards studies containing a fracture. The Flinders set was anonymized and provided without reports, injury, or population parameters¹² (Table 1).

Table 1 Properties of the internal validation dataset (IVD) and external validation dataset (EVD)

Dataset properties	IVD	EVD
Cases	409	399
Projections	≥ 4	3
Focus	Ankle study	Lateral malleolar fracture
Timing	Initial imaging	Initial imaging, one-week follow-up, weight-bearing
Implants & casts	Yes	No
Open physes	No	No
Excluded on imaging quality	None	Insufficient quality views Poor quality images Severely displaced fractures

Fracture	Cases	Percent (%)	Cases	Percent (%)
Base	253	61,9%	277	69,4%
Malleolar ^a	216	52,8%	274	68,7%
Fibula ^b	37	9,0%	3	0,8%
Previous fracture/other ^b	134	32,8%	15	3,8%
Foot ^b	57	13,9%	2	0,5%

Numbers are based on ground truth labelling by reviewers after the consensus session

^aDistinguishing between isolated fibula and lateral malleolar fracture can be subjective. In the absence of talar dislocation, we reviewed the radiologist's report for indications of direct trauma. Additionally, we assessed for specific characteristics, such as the presence of more transverse fractures, which are more common in isolated fibula fractures, as opposed to C-category fractures that often extend beyond 3 cm. While these criteria are not entirely objective, we aimed to apply them consistently across cases to minimize variability

^bDenotes fractures and outcomes that were flagged as fractures during study selection but are secondary outcomes

Annotation protocol for the EVD

The four reviewers (FIJ, EA, JD, and MG) were consultant or senior consultant orthopedic trauma surgeons. All underwent a training session to ensure familiarity with the labeling platform (the Raiddex platform developed by DeepMed AB) and agreement on the AO/OTA 2018 ankle fracture classification. Each reviewer labeled the EVD independently at the original

image resolution. Labeling was distributed so that three reviewers examined each study independently. After the independent labeling, we held a consensus session to review the cases where there were discrepancies between observers, and a majority vote decided the final classification. The result was the ground truth EVD. We have previously established the annotations for the training data and the IVD ground truth.⁹

Model and evaluation

Image pre-processing, network architecture (modified ResNet-based¹³ neural network model developed in PyTorch), parameters, training, and output evaluation were consistent with Olczak et al. 2021¹ and identical for the IVD and the EVD. The network scaled down exams to reduced-size images for training and assessment. The network was always trained for 300 epochs, and we did not stop early.

The software used in our previous study was unsuitable for this study's experiments. Instead, we used an identically trained network on the same IVD. Due to the random nature of model training, the exact performance for the initial IVD varied slightly from our previous study. After initial evaluation, we were dissatisfied with the model performance and noted a notable difference in the distribution of Type A fractures. As part of active learning, we: 1) expanded the training data with previously unlabeled ankle imaging from the training site, focusing on type A fracture. These were labeled by FW (medical student) and JO (medical doctor). We could not preselect fracture type among those previously unclassified images. 2) we actively focused on Type A fracture prediction edge cases. Fractures in the training set classified as Type A with the lowest probability or where type A fracture was the second most likely type (but another type won out) were rigorously reexamined after each training epoch by MG (senior orthopedic consultant). By focusing on the lowest probability type A fractures and almost type A, we hoped to reduce the uncertainty in the type A classes. In addition to adding more training data, we trained the model on different image resolutions. We report the results for the 400×400-sized images as the primary outcome. At higher resolutions, there was no performance increase.

The model classified studies by examining all available images individually and independently for each possible class. There were 39 outcomes for ankle (AO/OTA 44) fractures and, as many classes for fibula (4), tibial (43), foot fractures, and one additional for fracture-yes/no. The model selected the most probable AO/OTA class (top-1 classification) for the series outcome.⁹ Class outcomes, i.e., fracture yes/no, type (A–C – i.e., Danis-Weber), group (A1–C3), and subgroup (A1.1–C3.3) are determined independently of each other. We trained a network without pre-training, then used the resulting trained AI model to classify the IVD and EVD and compared the results to the ground truths. See Supplement 2 (available online) for details on the network, modeling parameters, and all possible outcomes for the network. We report our findings per the CAIR checklist¹ and follow the TRIPOD statement¹⁴.



Statistical analysis

Primary outcomes

The area under the receiver operating-characteristic curve (AUC) and the area under the precision-recall curve (AUPR) for malleolar fractures (AO/OTA 2018 bone-location 44). Top-1 classification is used for determining outcomes for each level, i.e., fracture/ vs. no fracture, type 44A-C, group 44A1-44C3, and subgroup 44A1.1-44A3.3, i.e., 40 possible outcomes for ankle fractures. While some outcomes overlap, each was decided individually. We used frequency-weighted means as summary statistics [1] and calculated 95% confidence intervals (CI) with bootstrapping. We do not report outcomes with single cases, as it is impossible to calculate CI for these outcomes.

Although AUC and accuracy are often used to report performance in CNN models, a multi-label classifier—such as that used in this study—benefits from a metric that can more accurately capture its inherent class imbalance between the many groups.

For the AUPR, a random classifier will perform proportionally to the number of positive outcomes for that class, i.e., $AUPR_{random} = (\text{number of cases for the class} / \text{total number of cases})$. If a dataset consists of 10% of class *X*, a random classifier should deliver an AUPR of 0.1 for class *X*, and anything above that is better than chance.¹⁵ Therefore, we also report when the AUPR outperforms a random classifier – i.e. when the lower 95%CI bound is better than the random classification. We only measured the top-1 performance (i.e., no points for being close to the correct answer).

After enhancing training (with active learning, additional training data, and increased image resolution), we only tested the model (on the EVD and IVD) once for each resolution. This was done to eliminate the risk of overfitting the EVD.

We compare the classifier's performance on both datasets and report according to the Clinical AI Research (CAIR) and TRIPOD checklists.

Secondary outcomes

Compare the classification between observers (before the consensus session) and performance for non-malleolar fracture outcomes. We use Cohen's kappa to compare two reviewers and intra-class correlation (ICC) to compare all reviewers. We use ICC and kappa as rough indicators of the difficulty of the classification task.

RESULTS

Compared to the IVD, the EVD had fewer displaced fractures and no casts or implants. The EVD included studies labeled “weight-bearing,” indicating that these were not fresh injuries at the time of examination (i.e., from the emergency department at the time of injury). The EVD had three views per study, while the IVD had four or more views. The IVD had 216 ankle

fractures out of 409 cases (53%), compared to 274 out of 399 ankle fractures (69%) in the EVD set (Table 1). The fracture incidence was similar, and type B fractures dominated both settings. Type A fractures were three times more prevalent in the EVD. There were more non-malleolar fractures in the IVD than in the EVD (Table 2 and Table 3). The EVD also had less severe fractures, e.g., more B1 fractures, less B3, and very few fibula fractures.

Table 2 Prediction outcomes for the internal validation dataset (IVD)

DANDERYD – Internal validation set (IVD) (409 cases)					
Malleolar fractures					
	Cases	AUC (95% CI)	ΔAUC	AUPR (95% CI)	ΔAUPR
Fracture	216	0.95 (0.94–0.97)	0.03	0.96 (0.94–0.97) ^a	0.03
44A					
Base	32	0.84 (0.76–0.92)	0.04	0.46 (0.11–0.35) ^a	0.23
44A1	22	0.84 (0.76–0.92)	-0.03	0.37 (0.08–0.29) ^a	0.19
44A1.1	6	0.88 (0.79–0.97)	-0.01	0.04 (0.01–0.10)	0.00
44A1.2	7	0.84 (0.69–1.00)	-0.02	0.30 (0.01–0.21)	0.22
44A1.3	9	0.82 (0.69–0.96)	0.03	0.18 (0.01–0.22)	0.11
44A2	7	0.99 (0.97–1.00)	0.15	0.52 (0.01–0.47)	0.28
44A2.1	5	0.99 (0.97–1.00)	0.09	0.41 (0.00–0.56)	0.15
44A2.3	2	0.99 (0.99–1.00)	0.14	0.25 (0.00–0.04)	0.23
44A3	2	0.95 (0.86–1.04)	-0.02	0.08 (0.03–0.17) ^a	0.01
44B					
Base	137	0.96 (0.93–0.92)	0.04	0.92 (0.88–0.95) ^a	0.05
44B1	67	0.95 (0.93–0.98)	0.05	0.77 (0.67–0.86) ^a	0.14
44B1.1	39	0.90 (0.87–0.94)	0.07	0.37 (0.25–0.51) ^a	0.06
44B1.2	26	0.94 (0.91–0.97)	0.07	0.40 (0.22–0.60) ^a	0.15
44B1.3	2	0.96 (0.90–1.02)	0.04	0.06 (0.01–0.23) ^a	0.03
44B2	38	0.86 (0.80–0.92)	0.01	0.40 (0.25–0.56) ^a	0.04
44B2.1	20	0.91 (0.85–0.97)	0.05	0.37 (0.20–0.55) ^a	0.14
44B2.2	16	0.88 (0.77–1.00)	-0.01	0.35 (0.15–0.53) ^a	0.13
44B2.3	2	0.87 (0.68–1.07)	-0.05	0.03 (0.00–0.11) ^a	0.00
44B3	32	0.92 (0.89–0.96)	0.06	0.50 (0.27–0.59) ^a	0.03
44B3.1	12	0.90 (0.83–0.97)	0.04	0.18 (0.06–0.34) ^a	0.02
44B3.2	13	0.92 (0.88–0.96)	0.08	0.20 (0.08–0.35) ^a	-0.04
44B3.3	6	0.96 (0.93–0.99)	0.02	0.16 (0.03–0.30) ^a	0.06



Table 2 Prediction outcomes for the internal validation dataset (IVD) (continued)

DANDERYD – Internal validation set (IVD) (409 cases)					
Malleolar fractures					
	Cases	AUC (95% CI)	ΔAUC	AUPR (95% CI)	ΔAUPR
44C					
Base	47	0.93 (0.89–0.97)	0.05	0.73 (0.61–0.82) ^a	0.20
44C1	24	0.90 (0.84–0.97)	0.05	0.42 (0.27–0.63) ^a	0.18
44C1.1	17	0.93 (0.87–0.99)	0.03	0.39 (0.21–0.60) ^a	0.16
44C1.2	5	0.86 (0.75–0.97)	-0.01	0.05 (0.01–0.12)	0.01
44C1.3	2	0.93 (0.83–1.02)	0.02	0.04 (0.01–0.14) ^a	0.02
44C2	18	0.93 (0.90–0.97)	-0.02	0.40 (0.16–0.58) ^a	-0.05
44C2.1	6	0.86 (0.74–0.99)	-0.08	0.22 (0.01–0.51)	0.07
44C2.2	3	0.99 (0.99–1.00)	0.08	0.32 (0.00–0.62)	0.28
44C2.3	9	0.92 (0.88–0.96)	0.03	0.11 (0.04–0.21) ^a	0.00
44C3	5	0.98 (0.97–1.00)	0.07	0.29 (0.02–0.67) ^a	0.21
44C3.1	3	0.96 (0.90–1.03)	0.29	0.16 (0.00–0.50)	0.15
		Weighted mean AUC	Δ	Weighted mean AUPR	Δ
		0.93	+0.04	0.65	+0.08

Reported with the area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (AUPR). The outcome measures for the most important groups. 95% confidence intervals (CI) are computed using bootstrapping. The "base case" corresponds to the Danis-Weber classes (AO/OTA 44A, 44B, and 44C). Outcomes with ≤ 1 instance are not reported. Radiographs at 400 × 400px resolution. ΔAUC and ΔAUPR was the difference in AUC and AUPR comparing the actively trained network to the pre-active training network at 256 × 256px resolution. Increasing resolution prior to active learning had no effect on performance

^aIndicates that the AUPR with 95% CI exceeds random AUPR

Table 3 Prediction outcomes for the external validation dataset (EVD)

FLINDERS – External validation dataset (EVD) (399 cases)					
Malleolar fractures (44)					
	Cases	AUC (95% CI)	ΔAUC	AUPR (95% CI)	ΔAUPR
Fracture	274	0.86 (0.82–0.89)	0.03	0.93 (0.91–0.95) ^a	0.00
44A					
Base	94	0.74 (0.68–0.80)	0.12	0.52 (0.40–0.61) ^a	0.20
44A1	93	0.75 (0.69–0.81)	0.14	0.57 (0.47–0.66) ^a	0.25
44A1.1	5	0.63 (0.33–0.94)	-0.07	0.04 (0.00–0.16)	0.02
44A1.2	28	0.78 (0.69–0.87)	0.15	0.26 (0.11–0.43) ^a	0.14
44A1.3	60	0.68 (0.61–0.76)	0.08	0.30 (0.20–0.41) ^a	0.10
44B					
Base	142	0.90 (0.87–0.93)	0.03	0.84 (0.78–0.89) ^a	0.03
44B1	116	0.84 (0.80–0.88)	0.03	0.68 (0.58–0.76) ^a	0.07
44B1.1	87	0.80 (0.75–0.85)	0.05	0.47 (0.37–0.56) ^a	0.06
44B1.2	27	0.80 (0.72–0.88)	0.02	0.19 (0.11–0.31) ^a	0.03
44B1.3	2	0.60 (0.17–1.02)	-0.30	0.01 (0.00–0.02)	-0.01
2	21	0.85 (0.75–0.94)	0.10	0.32 (0.17–0.50) ^a	0.19
44B2.1	18	0.85 (0.75–0.95)	0.12	0.33 (0.12–0.55) ^a	0.24
44B2.2	3	0.93 (0.88–0.99)	0.00	0.05 (0.00–0.17)	-0.03
44B3	5	0.82 (0.61–1.04)	-0.06	0.19 (0.01–0.58)	0.11
44B3.1	5	0.82 (0.63–1.02)	-0.05	0.12 (0.00–0.25)	0.07
44C					
Base	38	0.89 (0.82–0.96)	0.04	0.63 (0.46–0.78) ^a	-0.06
44C1	28	0.90 (0.84–0.96)	0.08	0.42 (0.26–0.65) ^a	0.07
44C1.1	27	0.90 (0.84–0.97)	0.07	0.44 (0.25–0.62) ^a	0.10
44C2	9	0.92 (0.82–1.01)	-0.04	0.19 (0.05–0.36) ^a	-0.40
44C2.1	9	0.90 (0.79–1.02)	-0.04	0.16 (0.05–0.31) ^a	-0.38
		Weighted mean AUC	Δ	Weighted mean AUPR	Δ
		0.83	+0.06	0.64	+0.07

Reported with the area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (AUPR). The outcome measures for the most important groups. 95% confidence intervals (CI) are computed using bootstrapping. The “base case” corresponds to the Danis-Weber classes (AO/OTA 44A, 44B, and 44C). Outcomes with ≤ 1 instance are not reported. Radiographs at 400 × 400px resolution. ΔAUC and ΔAUPR was the difference in AUC and AUPR comparing the actively trained network to the pre-active training network at 256 × 256px resolution. Increasing resolution prior to active learning had no effect on performance

^aIndicates that the AUPR with 95% CI exceeds random AUPR



Danderyd (IVD)

While the AUC was good for type A fractures in the IVD, AUPR was only better than chance for three outcomes (base case/ type “A”, subgroup A1.1 and group A3).

Type B fractures were the most numerous in the IVD. All had good to excellent AUC, and all had AUPR better than random, even rare outcomes such as B1.3, B2.3, and B3.3.

The network had excellent AUC and AUPR for type C fractures. The model performed better than random for the same type C outcomes in the Danderyd set (base, C1, C1.1, C2) as it did for the Flinders set, except C2.1.

The wAUC increased by 0.04 (from 0.89 to 0.93), and the wAUPR increased from 0.57 to 0.65 for the IVD. A random classifier would give a wAUPR of 0.23. See Table 2.

The model was less accurate for malleolar fracture detection (“base” AO/OTA 44) on the EVD than the IVD. The EVD dataset is less diverse, with fewer outcomes (23 vs. 36 AO/OTA outcomes). Notable was that fracture detection (fracture “yes”/“no”) had AUC 0.86 (0.82–0.89) for the EVD vs. AUC 0.95 (0.94–0.97) for the IVD.

Flinders (EVD)

Type A fractures were the second most numerous in the EVD, as was in group A1. Type A fractures had the lowest AUC, but only A1.1 performed worse than a random classifier measured in AUPR. However, there were few outcomes against which to measure performance. Figure 2 shows type A fractures and how the network classified them incorrectly.

For type B fractures, the base case performed well. While AUC was decent for all type B outcomes, four out of ten cases did not reach better AUPR than a random classifier (i.e., B1.3, B2.2, B3, and B3.1).

Type C fractures performed well, as did all four subclasses of type C outcomes. See Table 3.

Figure 3 shows an example of a type A1.3 fracture in the EVD, incorrectly classified as a type B fracture. Figure 4 shows examples of type B1.2 fracture incorrectly classified as type C.

Table 3 displays the change in performance for every class from active learning for the EVD. Most notable is a drop in performance for group C2 fractures, where group C2 and subgroup C2.1 decreased considerably in AUPR (-0.40 and -0.38, respectively). The wAUC increased by 0.06 (0.77 to 0.83), and the wAUPR increased from 0.57 to 0.63 for the EVD. A random classifier would give a wAUPR of 0.32.



Fig. 2 Incorrectly classified cases where the network failed to detect Type A, sorted from lowest probability to highest

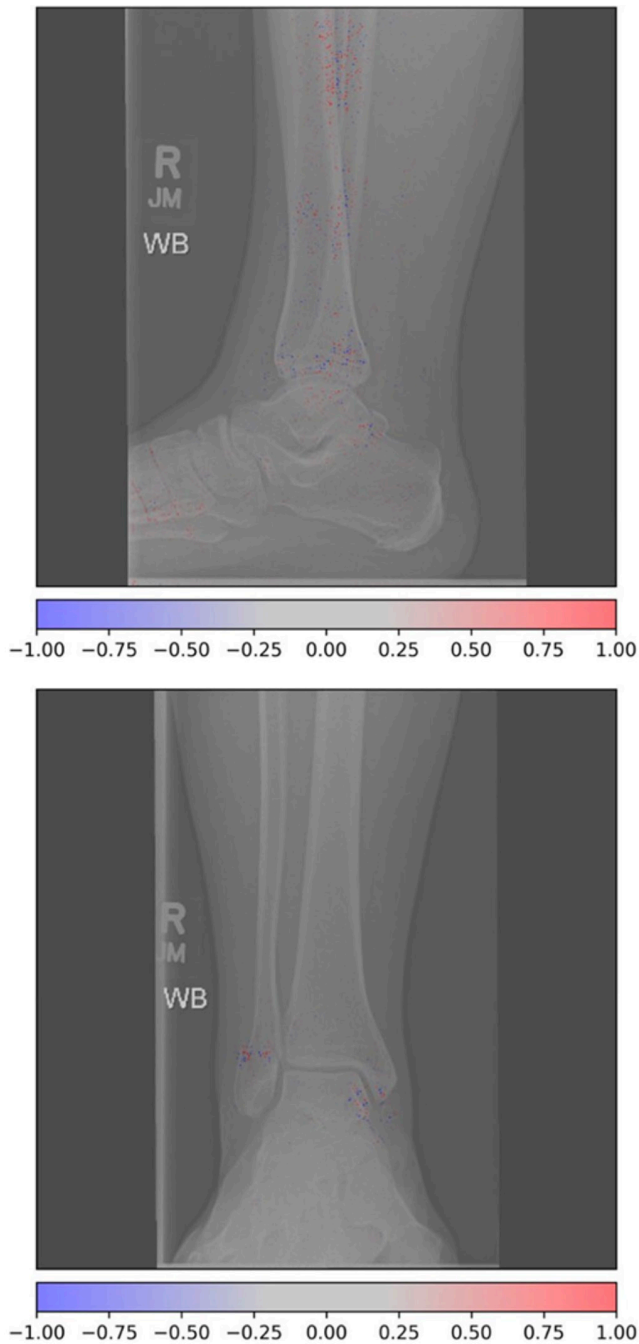


Fig. 3 Activation heatmap where a type 44A1.3 fracture is incorrectly classified as a type B fracture. The activations show what the model reacts to classify fractures. Study from the external validation data.

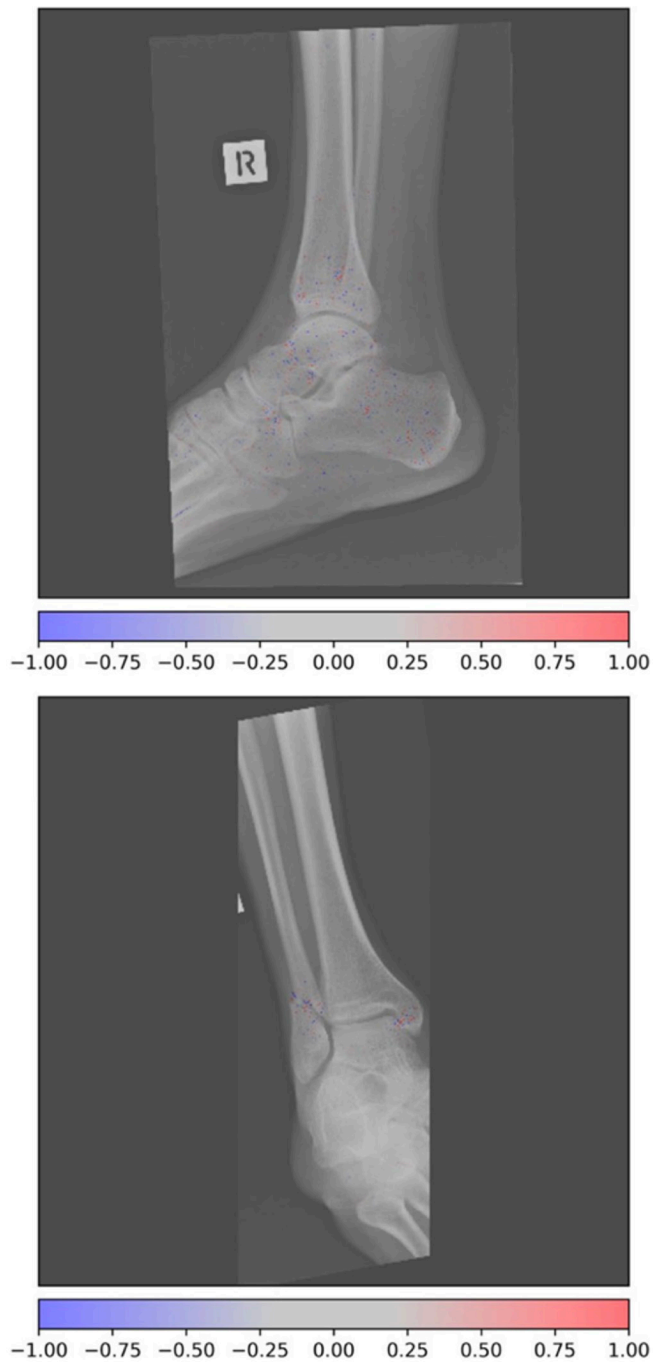


Fig. 4 Activation heatmap of a type 44B1.2 fracture, incorrectly classified as a type C fracture. The activations show what the model reacts to classify fractures. Study from the external validation data.

Secondary outcomes: Intra-observer measurements (ICC and Cohen's Kappa)

Intra-observer measurements for malleolar fracture have ICC 0.86 and kappa 0.85, and for the type of fracture, 44A-C, ICC 0.76–0.84 and kappa 0.85–0.78. Less severe fractures, i.e., A1, B1, and C1, had higher kappa values than more severe cases, perhaps because they are more prevalent. ICC and kappa are poor for most other outcomes, and we consider the task challenging for humans. Several classes had no interobserver agreement (kappa 0), mainly because they were not represented in the EVD or were so few that any disagreement or agreement had a disproportionate influence. See Supplement 1, Table S2 (available online).

We report performance for secondary outcomes, like non-malleolar fractures, along with more in-depth data and complete experiment readouts in Supplement 3 (400 × 400 pixels, available online) and Supplement 4 (256 × 256 pixels, available online). Supplement 4 (available online) reports the initial EVD performance before retraining and active learning.

DISCUSSION

This study aimed to externally validate a complex multi-label AO/OTA 2018 ankle fracture detection model. Few models are externally validated, and we found none as multifaceted as the AO/OTA in our study. In this study, we wanted to establish a baseline against which to compare future models. We found a gap in our model performance under external validation and reported a way of actively improving performance. We found that the model performed very well on an external validation set. Our model classified fractures much better than chance for all outcomes and indicates authentic learning utility for classifying ankle fractures in an external setting.

This study used the AO/OTA 2018 ankle classification system. A widely used alternative, alone or in conjunction, is the Lauge-Hansen (LH) system. The LH classification system aims to predict fracture patterns and ligamentous injuries based on injury mechanisms. Many studies have shown that LH is only partially valid or reproducible. Lindsjö, as far back as 1985, raised the question of poor reproducibility of LH between different populations based on previous studies.¹⁶ Later studies repeated these findings of poor reproducibility.^{17, 18, 19, 20, 21} An MRI study by Gardner et al. 2006 found that LH had limitations in predicting ligamentous injuries and soft-tissue damage.¹⁹ These findings were replicated by Kwon et al. in 2010 and 2012 using actual injury footage.^{22, 23, 24} Boszczyk et al. 2018, came to the same conclusion based on radiographs and patient-reported injury mechanisms.²¹ Patton et al. 2022 came to similar conclusions based on CT and complete patient workups.²⁵ Both Michelson et al. and Haraguchi and Armingier failed to reproduce Lauge-Hansens's results in physical experiments. They concluded that the LH system could not be used to predict injury mechanisms or injury patterns.^{26, 27} The AO/OTA standard launched the Danis-Weber system. Danis-Weber is based on the location of the lateral malleolus fracture about the syndesmosis. AO/OTA then extends the Danis-Weber

classification to consider the medial and posterior malleolus injuries and grades fractures based on physical appearance.^{9, 28} The main critique of the AO/OTA ankle system is that it is complex and that isolated medial malleolus fractures are treated as distal tibial fractures.^{20, 29}

Our goal was to develop AI models for rapid, easy, and accurate fracture classification and clinical decision-making. LH is not well suited to predicting injury mechanisms from radiographs in its current form, whereas AO/OTA is imaging-based. In the clinical context, AO/OTA (complete or simplified Danis-Weber) and LH are often used in conjunction to guide treatment decisions. The classifications are similar, and conversions between the two systems have been suggested, but no fully agreed-upon complete conversion exists.^{24, 30-34}

Model training

The training of AI models often comes down to hidden factors and confounders that are only sometimes related to actual pathology detection. For example, in a multicenter study, Badgeley et al. found that logistic and healthcare system parameters were often responsible for prediction. Without them, performance fell to that of a random classifier.³⁵ Subjecting the model to another dataset exposes it to a different data distribution—called a dataset shift³⁶—and is crucial for evaluating models. It should be integral to the model training and development stage. If the model only performs well on the data it was trained on or from one hospital, we can quantify this. It reduces the risk of presenting overfitted models as research progresses. In this study, the Flinders data has a different distribution and priori probabilities than the training data. For example, there were three times as many type A fractures. The Flinders data (e.g., EVD) had three images per study compared to four or more for Danderyd. The presence of follow-up images, e.g., weight-bearing one week after the trauma – was not a part of the network training. A non-displaced “weight-bearing” exam would signal a less severe injury to a human reviewer, whereas the network did not recognize this signal. We expected the IVD performance to be somewhat better. For both datasets, AUC and AUPR are better than random for all outcomes. Few AI models are validated, making it difficult to assess how general and transferable these models are to other settings and what performance we can expect in our study. For the three studies, Oliveira e Carmo et al. found performance was not affected dramatically for the validation set³ (see Supplement 1, Table S1a, available online). However, those studies evaluated models with just two or three outcomes. The AO/OTA classification, as used in our research, had 40 outcomes for ankle fractures – not all mutually exclusive.

As we were dissatisfied with the performance of the EVD, we tried multiple strategies to improve performance. We increased image resolution, which did not affect EVD performance. We tried to drop views to make the training data resemble the EVD data (three standard views in EVD vs four or more in the training data). Neither had any performance effect, and we speculated that type A fracture signs had a too-discrete training signal for the network. Only after active training (i.e., additional training data focusing on the problematic type A classes) could we improve performance by increasing resolution. Yet, we did not see any rise



in performance past 400×400 px. However, it can be desirable to reverse this generalization process on the externally validated model in a clinical application, i.e., honing it in the local setting. This would be done by actively retraining the externally valid model on data from the clinic where it is being used.

To our knowledge, this is the first study that externally validates such a complex fracture classification model and raises the question of what we can expect. Our model performed well compared to other multinomial classifiers, even on EVD data^{11, 37-43} (Supplement 1, Table S1b, available online). While we must take care when applying the algorithm to a new environment, it appears to work satisfactorily. Lim et al. (2014) found that many of the most common orthopedic procedures had poor evidence-based medicine support and were unnecessary.⁴⁴ We believe that tools such as this algorithm and evolutions could be part of the solution towards a more stringent and evidence-based treatment, for example, by reducing ambiguity in treatment decisions, identifying failure patterns, or automating data reporting to registries.

Limitations and strengths

In alignment with our previous studies^{9, 10}, we initially tested our trained model on 256×256 radiographs but had difficulties capturing type A fractures. We attributed this to them being rarer in the training data. Type C injuries were also uncommon, but the network performed better. Our experience was that the radiological footprint of type A fractures was less clear as these injuries tended to be less severe. The model captured Type A fractures after increasing the radiograph resolution and actively training for them. We found no benefit in going beyond 400×400 radiographs for our data.

As not all outcomes were sufficiently prevalent in the test data, we could not quantify all outcomes with reasonable confidence intervals. This was evident in cases where upper AUC confidence interval bounds exceeded 1.00 (i.e., 100% accuracy). Similarly, outcomes with few test cases (5 or less) AUPR could not be shown to outperform random guessing.

We did not have the population distribution for either dataset. The original training data was anonymized upon collection and did not come with population parameters. It consisted of all available trauma radiographs at Danderyd at that time. We have only excluded pediatric ankles. This makes it representative of the area from where it was collected. The Flinders data concentrated on lateral malleolus injuries and excluded casts and displaced fractures but included weight-bearing images and exams that were not concurrent with the injury. Therefore, it was impossible to determine how representative the EVD was of training data regarding population.

CT and MRI scans and operative findings are essential to the AO/OTA classification. CT scans are considered the gold standard in visualization. Neither dataset had access to CT scans, MRI scans, or patient journals. The lack of additional modalities or patient records made ligamentous injuries more challenging to classify and can affect the ground truth. However,

this accurately simulated the daily clinical practice in many clinical situations where the initial assessment is performed on a radiograph.

The validation sets were limited in size for a model with so many possible outcomes. Several outcomes were scarce in the EVD and IVD, making the variability extremely large. We have only validated our model on this site. If we were to look at a different hospital, we would get different results. Of course, this is true with all external validation.

The software used in our previous study was unsuitable for this study's experiments. Due to the random nature of model training, the exact performance for the IVD and EVD will vary slightly. However, comparing the previous model with the updated software where active learning was performed would be erroneous. Instead, we replicate the initial experiment.

Conclusions, interpretation & generalizability

To our knowledge, this is the first paper that externally validates a multi-label radiographic ankle fracture classifier of this complexity. Despite considerable differences in the data makeup, we illustrate active learning strategies to improve external validity. Our model could successfully be used for complex ankle fracture classification at a different hospital, which is not to say that it will work equally well at all hospitals. We concur with the literature that the clinical relevance of published AI models must be proved through external validation. As clinical AI goes beyond simply stating the obvious "fracture or no fracture," this becomes even more true.



REFERENCES

1. Olczak J, Pavlopoulos J, Prijs J, IJpma FFA, Doornberg JN, Lundström C, et al. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklistproposal. *Acta Orthop*. 2021;14:1–13. Page 12 of 13
2. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and metaanalysis. *Lancet Dig Health*. 2019;1(6):e271–97.
3. Oliveira e Carmo L, van den Merkho A, Olczak J, Gordon M, Jutte PC, Jaarsma RL, et al. An increasing number of convolutional neural networks for fracture recognition and classification in orthopaedics. *Bone Jt Open*. 2021;2(10):879–85.
4. Blüthgen C, Becker AS, de Vittoria Martini I, Meier A, Martini K, Frauenfelder T. Detection and localization of distal radius fractures: Deep learning system versus radiologists. *Eur J Radiol*. 2020;126:108925.
5. Choi JW, Cho YJ, Lee S, Lee J, Lee S, Choi YH, et al. Using a dual-input convolutional neural network for automated detection of pediatric supracondylar fracture on conventional radiography. *Invest Radiol*. 2020;55(2):101–10.
6. Zhou QQ, Wang J, Tang W, Hu ZC, Xia ZY, Li XS, et al. Automatic detection and classification of rib fractures on thoracic CT using convolutional neural network: accuracy and feasibility. *Kor J Radiol*. 2020;21(7):869–79.
7. Groot OQ, Bindels BJJ, Ogink PT, Kapoor ND, Twining PK, Collins AK, et al. Availability and reporting quality of external validations of machinelearning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthop*. 2021;92(4):385–93.
8. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020S;26(9):1364–74.
9. Olczak J, Emilson F, Razavian A, Antonsson T, Stark A, Gordon M. Ankle fracture classification using deep learning: automating detailed AO Foundation/Orthopedic Trauma Association (AO/OTA) 2018 malleolar fracture identification reaches a high degree of correct classification. *Acta Orthop*. 2021;92(1):102–8.
10. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthop*. 2017;88(6):581–6.
11. Lind A, Akbarian E, Olsson S, Näsell H, Sköldenberg O, Razavian AS, et al. Artificial intelligence for the classification of fractures around the knee in adults according to the 2018 AO/OTA classification system. *PLoS One*. 2021;16(4):e0248809.
12. Prijs J, Liao Z, To MS, Verjans J, Jutte PC, Stirler V, et al. Development and external validation of automated detection, classification, and localization of ankle fractures: inside the black box of a convolutional neural network (CNN). *Eur J Trauma Emerg Surg*. 2023;49(2):1057–69.
13. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv:151203385 [cs]. 2015 Dec 10; Available from: <http://arxiv.org/abs/1512.03385>. Cited 2021 Dec 7
14. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55–63.
15. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349800/>. Cited 2020 Aug 6

16. Lindsjö U. Classification of ankle fractures: The Lauge-Hansen or AO system? *Clin Orthop Relat Res.* 1985;199:12–5.
17. Thomsen NO, Overgaard S, Olsen LH, Hansen H, Nielsen ST. Observer variation in the radiographic classification of ankle fractures. *J Bone Joint Surg Br.* 1991;73(4):676–8.
18. Nielsen JØ, Dons-Jensen H, Sørensen HT. Lauge-Hansen classification of malleolar fractures: An assessment of the reproducibility in 118 cases. *Acta Orthop Scand.* 1990;61(5):385–7.
19. Gardner MJ, Demetrakopoulos D, Briggs SM, Helfet DL, Lorich DG. The ability of the Lauge-Hansen classification to predict ligament injury and mechanism in ankle fractures: an MRI study. *J Orthop Trauma.* 2006;20(4):267–72.
20. Fonseca L, Nunes I, Nogueira R, Martins G, Mesencio A, Kobata S. Reproducibility of the Lauge-Hansen, Danis-Weber, and AO classifications for ankle fractures. *Revista Brasileira de Ortopedia (English Edition).* 2017;1:53.
21. Boszczyk A, Fudalej M, Kwapisz S, Błoński M, Kiciński M, Kordasiewicz B, et al. X-ray features to predict ankle fracture mechanism. *Forensic Sci Int.* 2018;1(291):185–92.
22. Kwon JY, Chacko AT, Kadzielski JJ, Appleton PT, Rodriguez EK. A novel methodology for the study of injury mechanism ankle fracture analysis using injury videos posted on YouTube.com. *J Orthop Trauma.* 2010;24(8):477.
23. Rodriguez EK, Kwon JY, Chacko AT, Kadzielski JJ, Lindsay H, Appleton PT. An update on assessing the validity of the Lauge Hansen classification system for In-vivo ankle fractures using youtube videos of accidentally sustained ankle fractures as a tool for the dynamic assessment of injury. *Harvard Orthop J.* 2012;14:40–3.
24. Rodriguez EK, Kwon JY, Herder LM, Appleton PT. Correlation of AO and Lauge-Hansen classification systems for ankle fractures to the mechanism of injury. *Foot Ankle Int.* 2013;34(11):1516–20.
25. Patton BK, Orfield NJ, Clements JR. Does the Lauge-Hansen injury mechanism predict posterior Malleolar fracture morphology? *J Foot Ankle Surg.* 2022;61(6):1251–4.
26. Michelson J, Solocoff D, Waldman B, Kendell K, Ahn U. Ankle fractures. The Lauge-Hansen classification revisited. *Clin Orthop Relat Res.* 1997;345:198–205.
27. Haraguchi N, Armingier RS. A new interpretation of the mechanism of ankle fracture: JBJS. *J Bone Joint Surg Am.* 2009;1(91):821–9.
28. Meinberg EG, Agel J, Roberts CS, Karam MD, Kellam JF. Fracture and dislocation classification compendium-2018. *J Orthop Trauma.* 2018;32(Suppl1):S1-170.
29. Glen LZQ, Wong JYS, Tay WX, Li TP, Phua SKA, Manohara R, et al. Weber ankle fracture classification system yields greatest interobserver and intraobserver reliability over AO/OTA and Lauge-Hansen classification systems under time constraints in an Asian population. *J Foot Ankle Surg.* 2023;62(3):505–10.
30. Harper MC. Ankle fracture classification systems: a case for integration of the Lauge-Hansen and AO-Danis-Weber schemes. *Foot Ankle.* 1992;13(7):404–7.
31. Budny AM, Young BA. Analysis of radiographic classifications for rotational ankle fractures. *Clin Podiatr Med Surg.* 2008;25(2):139–52.
32. Chen DW, Li B, Yang YF, Yu GR. AO and Lauge-Hansen classification systems for ankle fractures. *Foot Ankle Int.* 2013;34(12):1750–1750.
33. Tartaglione JP, Rosenbaum AJ, Abousayed M, DiPrea JA. Classifications in brief: Lauge-Hansen classification of ankle fractures. *Clin Orthop Relat Res.* 2015;473(10):3323–8.
34. Rydberg EM, Zorko T, Sundfeldt M, Möller M, Wennergren D. Classification and treatment of lateral malleolar fractures - a single-center analysis of 439 ankle fractures using the Swedish Fracture Register. *BMC Musculoskelet Disord.* 2020;21(1):521.



35. Badgeley MA, Zech JR, Oakden-Rayner L, Glucksberg BS, Liu M, Gale W, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *Npj Digital Med.* 2019;2(1):1–10.
36. MIT Press. Dataset shift in machine learning. Quiñero-Candela J, editor. Cambridge, Mass: MIT Press; 2009. 229 p. (Neural information processing series). <https://mitpress.mit.edu/9780262545877/dataset-shift-in-machine-learning/>.
37. Dreizin D, Goldmann F, LeBedis C, Boscak A, Dattwyler M, Bodanapally U, et al. An automated deep learning method for tile AO/OTA pelvic fracture severity grading from trauma whole-body CT. *J Digit Imaging.* 2021;34(1):53–65.
38. Qi Y, Zhao J, Shi Y, Zuo G, Zhang H, Long Y, et al. Ground truth annotated femoral X-ray image dataset and object detection based method for fracture types classification. *IEEE Access.* 2020;8:189436–44.
39. Tanzi L, Vezzetti E, Moreno R, Aprato A, Audisio A, Massè A. Hierarchical fracture classification of proximal femur X-ray images using a multistage deep learning approach. *Eur J Radiol.* 2020;1(133):109373.
40. Yoon SJ, Hyong Kim T, Joo SB, Eel OhS. Automatic multi-class intertrochanteric femur fracture detection from CT images based on AO/OTA classification using faster R-CNN-BO method. *J Appl Biomed.* 2020;18(4):97–105.
41. Lee KM, Lee SY, Han CS, Choi SM. Long bone fracture type classification for limited number of CT data with deep learning. In: *Proceedings of the 35th Annual ACM Symposium on Applied Computing.* New York, NY, USA: Association for Computing Machinery; 2020. p. 1090–5. Available from: <https://doi.org/10.1145/3341105.3373900>. Cited 2022 Jan 6
42. Olsson S, Akbarian E, Lind A, Razavian AS, Gordon M. Automating classification of osteoarthritis according to Kellgren-Lawrence in the knee using deep learning in an unfiltered adult population. *BMC Musculoskelet Disord.* 2021;22(1):844.
43. Chung SW, Han SS, Lee JW, Oh KS, Kim NR, Yoon JP, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop.* 2018;89(4):468–73.
44. Lim HC, Adie S, Naylor JM, Harris IA. Randomised trial support for orthopaedic surgical procedures. *PLoS One.* 2014;9(6):e96745.





CHAPTER

6

Artificial Intelligence and 3D-Guided Surgery in Orthopaedic Trauma: Why, How, and What

J.N. Doornberg
F.F.A. IJpma
P. Jayakumar
V.M.A. Stirler
N. Assink
J. Prijs

Book chapter in Rockwood and Green, 2024

ARTIFICIAL INTELLIGENCE IN ORTHOPAEDIC TRAUMA

Hype or Disruptive Innovation?

On August 3, 2023, the authors of this chapter entered the following prompt into the Chat Generative Pretrained Transformer (ChatGPT; OpenAI, San Francisco, CA): “Please write a scientific introduction for Rockwood and Green; covering artificial intelligence (AI) and its basics, the applications in orthopaedic trauma, and future perspective.”

ChatGPT⁸⁹ returned the following:

“In the rapidly evolving landscape of orthopaedic trauma management, the integration of Artificial Intelligence (AI) has emerged as a transformative paradigm, promising to revolutionize traditional approaches and enhance clinical outcomes.

AI, a multidisciplinary field at the intersection of computer science and healthcare, encompasses the creation of intelligent systems capable of mimicking human cognitive functions such as learning, reasoning, and problem-solving. Its fundamental principles, rooted in machine learning and data analysis, empower AI systems to extract meaningful insights from vast and complex datasets, thereby augmenting diagnostic accuracy, treatment planning, and prognostic capabilities. Within the purview of orthopaedic trauma, AI's prowess has found a compelling canvas for applications ranging from automated fracture detection and classification to personalized surgical navigation and predictive modeling. As we stand on the precipice of a new era in medical practice, this chapter endeavors to elucidate the fundamental underpinnings of AI, elucidate its current applications in orthopaedic trauma, and extrapolate the tantalizing future prospects that hold the promise of reshaping the landscape of fracture care.”

AI—the science and engineering of intelligent machines, especially computer programs—now permeates health care. The application of AI in the field of orthopaedics is a new and transformative frontier of data science, bearing the hope and promise of revolutionizing ways to promote health and deliver health care across specialties, including orthopaedic trauma.^{23,}

^{63, 86, 87, 93}

After promising studies^{78, 79, 82, 105}, many reasoned that “with Watson's help” AI would change our medical world. In contrast, critics argued hope and hype would crash in the trough of disillusionment, also coined the “valley of death” for many AI-driven applications, following Gartner's hype cycle⁸⁷ (Fig. 9-1). Indeed, until 2022^{90, 101}, there had not been a single AI-driven clinical application in orthopaedic trauma that changed our day-to-day practice.^{19, 59, 67} For example, machine learning did not outperform traditional logistic regression algorithms for predictive modeling of clinical outcomes in trauma data sets.⁸⁸ Nevertheless, many continued

to believe—and write—that AI had the potential to greatly impact our field of orthopaedics.^{19, 23, 59, 63, 67, 86-88, 93} Moreover, the rapid expansion of interest in this technology has gained an additional boost since the development of generative AI—a class of machine learning trained using large volumes of text, image, and audio data to generate new, natural human-like content.

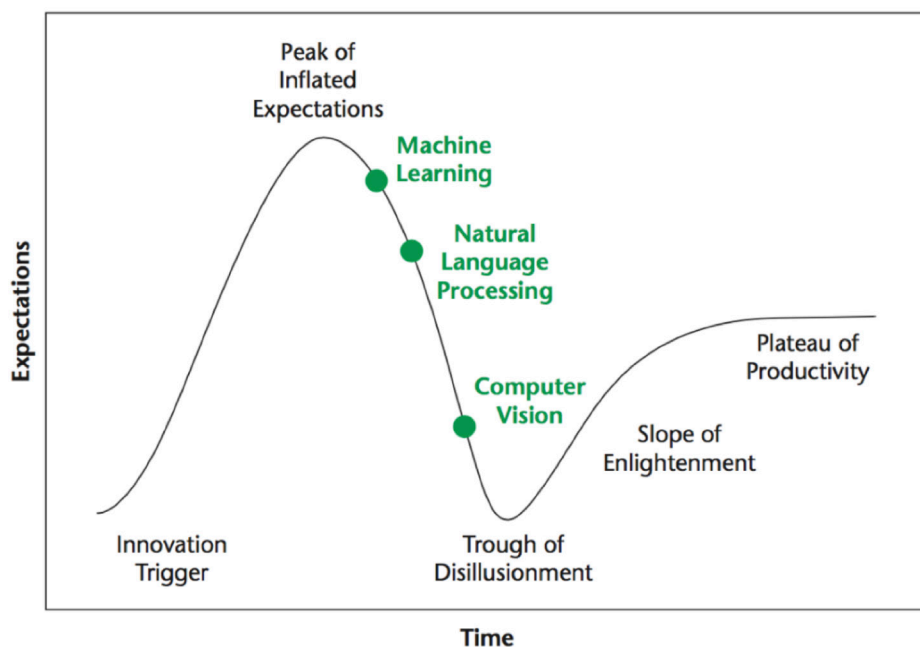


Figure 9-1. Gartner hype cycle, or Scott parabola, moving from innovation trigger, to peak of inflated expectations, via the trough of disillusionment (also referred to as “the valley of death” for many AI-driven applications), to ultimately reach a plateau of productivity. (Reproduced from Oosterhoff JHF, et al. Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of Gartner’s hype cycle. *EFORT Open Rev.* 2020;5(10):593–603. Adapted from Gartner Hype Cycle for Artificial Intelligence. 2019. gartner.com/smarterwithgartner.)

Large language models (LLMs) or chatbots, such as ChatGPT, are a form of generative AI, trained using high volumes of text data, to perform a range of language-specific tasks. True disruption came early 2023, when ChatGPT became mainstream. To illustrate the level of disruption, the authors of this chapter prompted ChatGPT to write several paragraphs throughout this chapter, making it a virtual coauthor. Chatbot writing is a hot topic at schools and universities and among editors of scientific literature.^{58, 59, 61}

Generative AI, LLM in particular, convinced the public of AI’s unparalleled potential, its undisputable strengths, and its inherent flaws, dangers even [100]. While current versions of these chatbots generate content with a remarkable level of sophistication, there will undoubtedly be barriers to overcome and continual improvement to be made to enable real

world impact (Fig. 9-2). The authors argue that 2024 may be the year to leave the trough of disillusionment, ascend the slope of enlightenment, and finally discover the first useful applications of AI in orthopaedic trauma in day-to-day clinical practice.^{23, 54, 55, 60, 62, 85-88, 91, 92} We are ever curious what the plateau of productivity will bring to our patients' care.⁵⁵

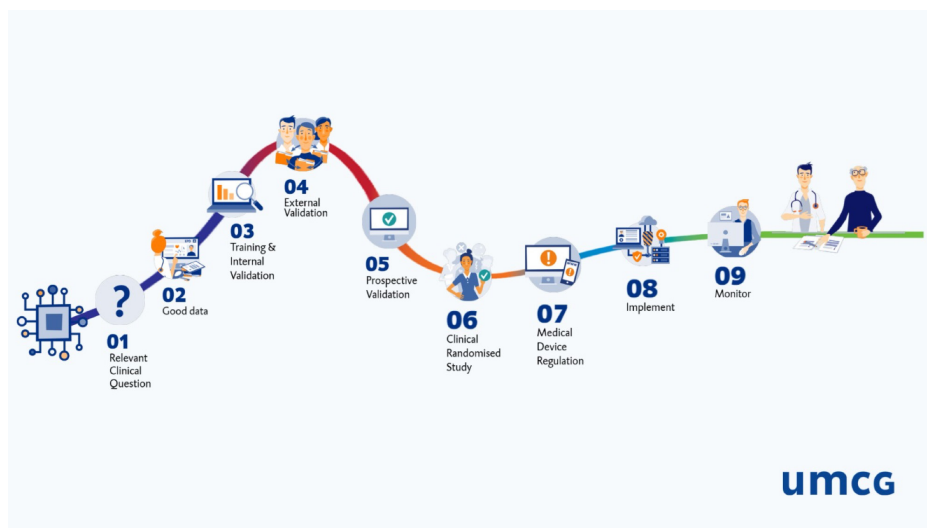


Figure 9-2. AI development: algorithm to patient. Barriers to overcome and continual improvement to enable real-world impact pathway from a clinically relevant question through all the required steps to reach models ready for clinical application.

A Brief History of AI

AI dates back to the mid-20th century. In 1943, Warren McCulloch and Walter Pitts proposed the first theoretical computational model, which served as a foundation for the development of machine learning (a branch of AI focused on the use of data, algorithms, and statistical models to imitate human learning with increasing accuracy). Machine intelligence was also integral to cracking Germany's enigma coding encryption during the Second World War by a team led by British mathematician and cryptanalyst Alan Turing at Bletchley.¹⁰⁰ In 1950, Turing coined the imitation game (or Turing test) as the test of a machine's ability to express intelligent behavior equivalent to or indistinguishable from that of a human.

In 1956, John McCarthy, a computer and cognitive scientist, coined the term artificial intelligence and, with a group of researchers including Marvin Minsky and Claude Shannon, organized the Dartmouth Conference, widely considered the birthplace of AI.⁸⁰ The 1950s also saw Frank Rosenblatt's design of the first neural network for computers, a method in AI and type of machine learning process called deep learning, in which computers process data via a series of interconnected nodes or neurons in a layered structure akin to the human

brain. Arthur Samuel developed the first self-learning computer and coined the term machine learning during this period.

The 1960s and 1970s saw substantial progress in the field of AI, with researchers developing symbolic reasoning systems, natural language processing algorithms, realistic pattern matching and substitution, and early versions of AI robots and chatbots. However, progress slowed in the mid-1970s onward as processing power struggled to keep up with the strong theoretical basis of the technology developed by scientists. Researchers and computer engineers failed to develop algorithms capable of handling the complexity and nuance of real-world problems. The development of expert systems—programs able to learn by asking experts in a given field how to respond in specific situations—in the 1980s and increasing expenditure on more sophisticated computing systems for machine learning and deep learning during the 1990s reignited the interest in AI.

With advancements in computing power and the access and availability of “big data”, machine learning algorithms became increasingly effective (moving from knowledge-driven approaches to data-driven approaches) at handling complex tasks, such as image and speech recognition. Defeat of world chess champion Garry Kasparov by IBM’s Deep Blue supercomputer in 1997 ushered in a new era of AI and a transition toward widespread, practical applications within software services and mobile devices, including the da Vinci robotics-assisted surgical system (Intuitive Surgical)—the first device to gain U.S. Food and Drug Administration approval for general laparoscopic surgery.

In 2006, Geoffrey Hinton coined the term deep learning to describe algorithms that could be trained to recognize objects and text in images and videos. The 21st century has also observed the development of AI-enabled autonomous vehicles, voice-activated virtual assistants, human gesture recognition solutions, increasingly humanoid robots, programs to detect previously undetected particles, computers that can teach themselves to walk, biometric recognition, and, in health care, increasingly effective detection and diagnosis of conditions such as cancer.

While the expectations for powerful AI solutions remain high, critics have also argued that many of these technologies have and will continue to plunge into the trough of disillusionment following Gartner’s hype cycle⁸⁷ (see Fig. 9-1). Despite the array of powerful functions provided by Watson—IBM’s AI and Internet of Things (IoT) platform—announcement of its full withdrawal on December 1, 2023, provided a stark awareness of the potential lifespan of such technology and important considerations around integration into real-world contexts, including the range of settings in health care.

Overall, the authors of this chapter consider AI as a technology capable of driving all forms of innovation in orthopaedic trauma. We believe AI will help sustain an existing position in the market (sustaining innovation, such as current AI/computer vision solutions for detecting fractures); will provide gradual, continuous improvement of existing services (incremental innovation, such as improvements in interpretation of text and audio for automating coding); will form technologies that disrupt the entire market (disruptive innovation, such as AI- and



augmented reality-enabled surgery; and will spark breakthrough technology that transforms the industry by creating new markets (radical innovation, such as providing patients unparalleled access to medical information and engagement in health systems).^{23, 54, 55, 60, 62,}

85-88, 91, 92

WHY?

To Alleviate the Healthcare Burden to Society: Current Healthcare Systems Are Not Sustainable

"If the computer does the simple stuff, we have more time for our patients".⁹⁵ AI has demonstrated the ability to liberate surgeons from repetitive operational tasks, enabling them to spend more time on rewarding engagement with their patients, fellow clinicians at work, and family and friends outside work.

Common frustrations during an orthopaedic surgeon's daily work often arise from mundane tasks, navigating multiple platforms that lack interoperability, to achieve results that should be simple but instead demand multiple steps, such as ordering radiographs in the emergency department (ED), generating clinical documentation of an outpatient visit, triggering transitions of care and appointment scheduling, and operation notes and postoperative orders. Substantial cost savings may be achieved by automation of relatively basic, low-risk, routine processes and procedures.²²

AI can substantially augment a clinician's activities through pattern recognition at a pace and scale to improve diagnostic precision and efficiency. For more complex tasks, AI is more likely to complement physicians than to replace them. Speeding up diagnostics in simple cases creates more autonomy for junior doctors by supplying AI-driven expert advice in an instant, thereby closing the so-called expert gap.^{40, 52} AI models could make the ED more productive and safer with automated triage based on plain text or with early warning systems for adverse events.^{30, 35, 42} It is estimated that AI could reduce annual health care costs in the US, where expenditure is at an all-time high, by \$150 billion by 2026.⁴⁹

Overcome Human Biases: Decrease Undesired Practice Variation, Improve Pattern Recognition, and Reduce Biased Decision Making

AI could be an aid to critical thinking. For a patient with an ankle or distal radius fracture, the adage "What you get depends on where you live and who you see" is true.^{29, 51} One could argue that such practice variation is undesired. Medicine in general, and our field of orthopaedic trauma in particular, is rife with unsatisfactory levels of agreement between doctors and surgeons. One good example is the variability in fracture recognition and fracture classification, alongside relatively high variations in surgical decision making.^{15, 17, 18, 25-27, 36, 75, 76, 103, 106} In this

clinical scenario, AI algorithms should not just replicate human thinking processes but should aim to exceed them.^{1, 22}

AI is all about pattern recognition, at pace and at scale to improve diagnostic precision and efficiency. This capability is especially relevant to care of the orthopaedic trauma patient, which is often dependent on a surgeon's ability to recognize patterns such as (1) fracture patterns on radiographs (AI domain: computer vision)^{23, 54, 62, 86, 91}; (2) a combination of specific patient characteristics that could predict arthrosis, infection, or trigger adverse events and other clinically relevant outcomes (AI domain: machine learning prediction tools)⁶⁶; (3) an arrangement of words predictive of underlying anxiety or depression (AI domain: natural or LLM [3]; and (4) patient's facial expression¹⁰⁴ in order to integrate mental and social health in orthopaedic practice.⁴⁵

However, human behavior and performance, especially with regard to pattern recognition, has its limitations and is prone to several biases.⁴⁸ For instance, the recognition of fracture patterns is shown to differ greatly between surgeons such that different surgeons see different things. Moreover, humans are unaware of their blind spots (unknown unknowns), which has been coined WYSIATI, for "what you see is all there is", by Daniel Kahneman.⁴⁸ Biases and blind spots may systematically affect trauma patients by overlooking specific fracture patterns, by making treatment decisions based on familiarity and personal preference rather than evidence, and, more importantly, by failing to recognize aspects of individual mindset or circumstances (the "whole person lens") and focusing instead on pathophysiology and the familiar biomedical paradigm.

It is important to note that the complete elimination of human biases is not possible and the introduction of and perpetuation of biases within training algorithms remain a risk that should be controlled. For instance, the potential amplification of bias should be addressed when using clinical datasets from one population for predicting clinical outcomes in another. The importance of ensuring bias-mitigation strategies when building AI models is paramount.⁸³ Therefore, we should be mindful of the ground truth, not mistaking our biased opinions for a gold standard in supervised training of algorithms.¹³



HOW?

Guiding Principles for AI Model Development in Orthopaedic Trauma

ChatGPT⁸⁹ output:

"Basics of Artificial Intelligence

Artificial Intelligence encompasses a diverse range of techniques and methodologies that enable computers to simulate human-like intelligence. At its core, AI relies on machine learning algorithms that learn from data patterns and iteratively improve their performance

over time. These algorithms include supervised learning, where models are trained on labeled datasets, and unsupervised learning, which involves extracting insights from unlabeled data. Supervised learning can generally achieve good to excellent performance with limited data, but providing detailed information about the data is a costly effort, especially when these need to be provided by experts that have limited time—such as surgeons. This is one of the most challenging obstacles in creating accurate models in our field. Reinforcement learning further enables AI systems to make decisions by interacting with an environment and receiving feedback. Deep learning, a subset of machine learning and the most recent leap forward in the AI field, involves artificial neural networks inspired by the human brain's structure and function. These networks, often referred to as deep neural networks, are particularly adept at handling complex and unstructured data, such as images and text, making them invaluable tools in medical imaging analysis and natural language processing.”

Expanding Education: Training our Peers

Orthopaedic trauma surgeons should become familiar with the basic concepts, technical aspects, and implementation of AI technologies in orthopaedic trauma to better understand and critically appraise the validity of AI algorithms.⁴⁶ Avoidance will hinder the opportunities to optimize the use of these technologies as they become introduced in our daily practice and will magnify the risk of dependence as AI becomes part of the data infrastructure in health care, including electronic medical records (EMRs). Several educational initiatives specifically combining AI and health care exist, such as the European Union–funded AIPROHEALTH², which offers a free online course as well as certificate programs in AI for health care providers.

There is also a movement toward introducing education regarding AI in a more structured manner. The basics of AI should be included when training medical students, residents, and practice peers to understand the language of specialists in this field—our collaborating computer scientists.

Interdisciplinary Engagement: Bringing Computer Scientists and Engineers into Orthopaedic Trauma Care

Orthopaedic trauma teams should engage with computer scientists and engineers to help develop effective use cases and models designed to fit these functions in the real world. While computer scientists and engineering resources are often limited within health care, especially considering the pull of such talent into the medical technology and digital health industry, orthopaedic trauma offers several clinically meaningful (and commercially viable) opportunities for interdisciplinary engagement. A new generation of clinicians with a specialist interest in AI are emerging that understand the importance of bringing together good data infrastructures with appropriate AI techniques to match the specific clinical use case.

Data Integrity

Use of Unbiased Prospective Data in Orthopaedic Trauma

A model can perform only as well as the data used to train it. Collaboration between different institutions and surgeons is essential in generating sufficient quantities of high-quality data to train models. While recent endeavors using “big data” in orthopaedics are encouraging^{34, 53}, the availability of data of high volume (e.g., cohorts from large health systems, clinical registries, biobanks) and high quality is limited, especially in orthopaedic trauma. These data limitations, particularly in terms of retrospective data, may be due to the acute nature of trauma, scattered datasets of variable quality that are not networked between hospitals, and restrictions around the availability of data and experts to label the data itself. The ability to utilize prospective data in AI models is further challenged by the requirement of greater financial resources, often a lengthy inclusion period, and high variations in treatment standards, potentially leading to underpowered analyses due to insufficient numbers of a patients for a given injury type.

Overfitting and Ground Truth

Challenges of AI model development across these functions are multifold. Overfitting is a phenomenon that occurs when a model becomes exceedingly well adapted to the training data set and captures not just the main underlying patterns but also the noise and outliers specific to that data. Model overfit can occur when training deep learning models and as a result compromise its external validity and generalizability when applied to new data sets. This limitation renders the model less effective, as it tends to make predictions based on the noise it has learned rather than being anchored around the true and genuine patterns.

In a famous example, a model was trained to detect traumatic pneumothorax on chest radiographs. The data set used for training contained a relatively large number of radiographs with a chest drain inserted within the pleural cavity. Even though the model output was frequently correct, its prediction was based on the presence of a chest drain (noise) and not the pneumothorax itself (Fig. 9-3). Overfitting can have severe implications in relation to orthopaedic trauma imaging, leading to incorrect diagnoses or misguided treatment recommendations. Therefore, ensuring a model is robust and not overfit is paramount for maintaining accuracy and reliability in clinical applications. Regularization techniques, cross-validation, data augmentation, and increasing the diversity of training data are common methods employed to mitigate overfitting.



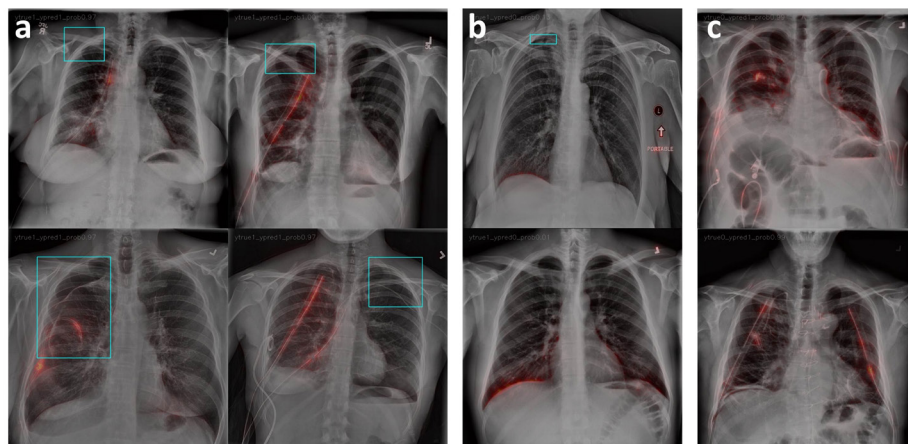


Figure 9-3. CNN was trained to detect traumatic pneumothorax on chest radiographs. The data set used for training contained a relatively large number of radiographs with a chest drain inserted within the pleural cavity. Even though the model output was frequently correct, its prediction was based on the presence of a chest drain (noise) and not the pneumothorax itself, shown in red.

In addition to variations in data, differences in labeling can also exert a notable impact on model performance. Label noise can significantly disrupt a convolutional neural network's (CNN's) effectiveness, and radiology reports, which often rely on a single observer, can vary in terms of expertise and accuracy. Relying solely on a single expert for data labeling carries the inherent risk of biasing the CNN toward that specific observer. Another method is to classify each fracture by multiple clinicians—to create the so-called ground truth by majority voted decision. However, requiring experts to label and annotate each individual case is extremely labor-intensive and repetitive. In addition, human bias is reduced but not eliminated when a majority voted decision is used. These limitations should all be considered when critically appraising deep learning algorithms. An example in fracture classification: What was considered the ground truth? Which imaging modality was used to determine this reference standard? Were there one or more observers classifying to determine “the truth”?

Moreover, it is challenging to find a sufficient number of qualified experts to provide accurate image labels when developing CNNs. Using reference standards such as CT scans as the ground truth is considered the best approach for training CNNs. However, these standards may not always be readily available, especially in cases involving “simple” fractures.

Appropriate Regulation: Enabling Responsible AI Through Robust Governance, Data Sharing, and Federated Learning

Appropriate levels of AI regulation are required globally, and effective regulatory standards by organizations within a region should be emulated and mirrored by others. For instance, the EU has developed strong general data protection and regulation guidelines that other countries should consider adopting to ensure adequate levels of data privacy, transparency, accountability, accreditation, and security. Specific safeguarding strategies should include careful documentation of data parameters, disclosure of methods for model development, data sources, and open acknowledgment of potential biases, risks, and limitations of models. Patients place substantial trust in health systems that use their personal health information, and these safeguards and protections should be adhered to. This aspect is especially important in regard to the trauma team, which is focused primarily on the acute management of patients and less concerned about data per se during emergent care. Leadership from professional societies and organizations (e.g., OTA, AO, national orthopaedic associations), patient advocacy groups, and academic experts within and outside orthopaedics should be convened for human oversight over development (reinforcement learning through human feedback) of AI models to ensure trust among all stakeholders in this arena.

Federated learning—a decentralized approach to AI and machine learning models in which raw data from edge devices are used to train models locally, thereby increasing data privacy, before being shared with global servers as needed—offers an enhanced approach that optimizes privacy while ensuring more streamlined flow of data between systems. This approach may allow safe and efficient model training across hospital networks and therefore more accurate development of AI solutions tailored to different patient populations experiencing orthopaedic trauma.

Open access to algorithms allows for independent validation and adherence to FAIR—findable, accessible, interoperable, and reusable—principles. Further, ethical and legal guardrails should be developed that are aligned with values of diversity, equity, and inclusion. Regulatory agencies should also be empowered with the authority to both enforce compliance and grant accreditations for the development and utilization of safe, accurate, and equitable AI models. Clinicians are ultimately responsible and liable for the outcomes of care delivered using AI and therefore should understand the basis of this tool like any other in their toolkit.

In the research domain, advances have been made around standards of reporting AI research, including CONSORT-AI⁶⁵, CAIR⁸⁵, and TRIPOD-AI²⁰.



WHAT?

This section will offer guidance for critical appraisal of AI tools in the upcoming era and will discuss examples of potential applications.

ChatGPT⁸⁹ output:

"Applications in Orthopaedic Trauma

In the realm of orthopaedic trauma, AI exhibits transformative potential across diverse domains. 1) AI-driven image analysis facilitates the rapid and accurate detection, classification, and segmentation of fractures from radiological images, expediting diagnosis and treatment planning. 2) Surgical interventions are bolstered by AI-powered preoperative simulations and intraoperative guidance, enhancing precision, and minimizing risks. Additionally, 3) AI-powered predictive models utilize patient data to forecast postoperative outcomes, aiding in shared decision-making and optimizing treatment strategies. Furthermore, AI-driven data mining and knowledge extraction from electronic health records contribute to evidence-based clinical guidelines and facilitate research endeavors.

There is a range of computer vision modalities, encompassing image classification, object detection and semantic segmentation. As we delve into each facet, we uncover their respective contributions to fracture assessment and therapeutic interventions, underscoring the transformative potential computer vision holds in orthopaedic trauma care."

AI-Driven Image Analysis: Computer Vision

Fracture Detection

Image recognition and classification models using AI (also termed computer vision) are increasingly being developed for automated fracture detection^{54, 62, 67, 84, 86, 91}, including some solutions that are commercially available⁵. Most models provide classification based on the entire image and produce heatmaps for approximate fracture localization (Fig. 9-4). New types of models developed by computer scientists allow for more advanced possibilities, such as accurate segmentation of bones and fractures. For example, the development of the Mask R-CNN makes it possible to combine conventional object detection with segmentation, an exact outline of the area of interest [91].

This neural network was used to train a model to detect, locate, and segment distal radius fractures (Fig. 9-5). The benefit of these models compared to conventional object detection lies in the user interface; that is, clinicians can more readily visualize and confirm whether the conclusion drawn by the computer is correct. This type of solution is an example of "explainable AI," in which the observer is able to understand what it is that the computer "sees".⁹¹

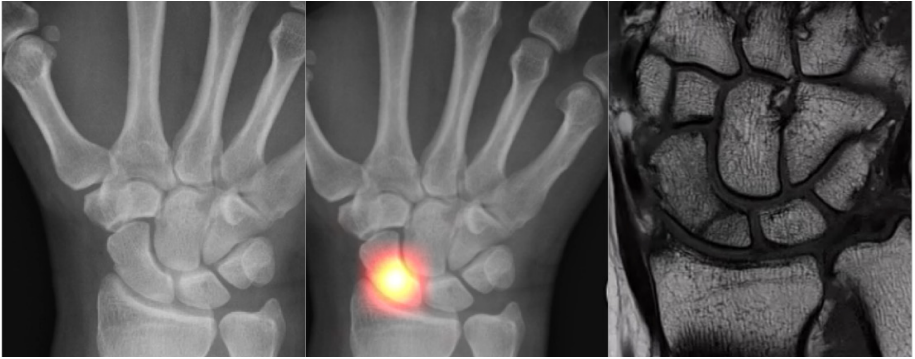


Figure 9-4. CNN for scaphoid fracture detection. Most models provide classification based on the entire image and produce heatmaps for approximate fracture localization.

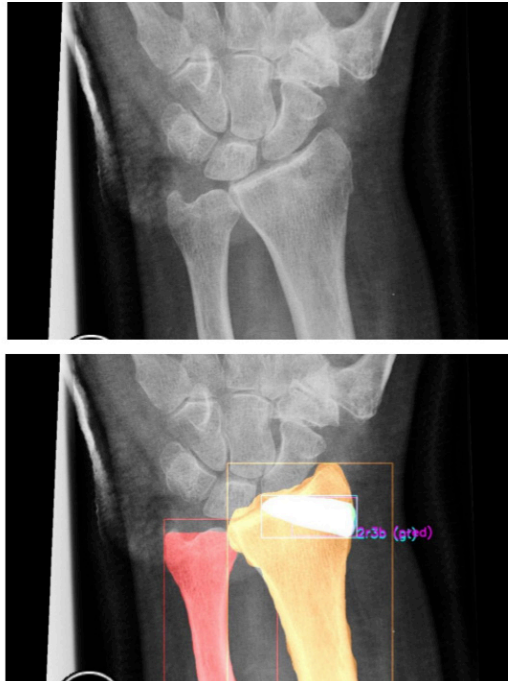


Figure 9-5. The development of the Mask R-CNN makes it possible to combine conventional object detection with segmentation—an exact outline of the area of interest. This neural network was used to train a model to detect, locate, and segment distal radius fractures.

While earlier computer vision models were developed for fractures that are easily recognized and classified clinically, growing experience with this technology is being used to solve increasingly complex tasks. For instance, source data from advanced imaging such as computed tomography (CT) and magnetic resonance imaging (MRI) are now being utilized as the ground truth and the visual input for AI models to detect injuries such as scaphoid fractures, which can be occult in 20% of patients, with increasing precision^{38, 39, 54} (see Fig. 9-4). AI models may recognize patterns that are not apparent to the clinician or radiologist, especially if they are challenging to diagnose on plain radiographs. The combination of improved reliability as compared to humans and a visual output with this prediction, fractures may be picked up on initial presentation and receive optimal treatment, such as, in our example of scaphoid fractures, avoiding overtreatment by either unnecessary cast immobilization or redundant advanced imaging with MRI or CT.⁹⁹

Fracture Classification

Training algorithms to classify fractures is a more challenging task given the marked variation in intra- and interobserver agreement on fracture classification, especially based on plain radiographs alone.^{15, 17, 18, 21, 26, 36, 75, 76, 81, 103, 106} While CT scans are commonly used to improve accuracy and definition of the fracture configuration, they are also subject to a level of inter-surgeon disagreement. As expressed by Doornberg et al., “Surgeons agree mostly with themselves, and not so much with each other”.²⁷ However, classification of each fracture by multiple clinicians to create the so-called ground truth by majority voted decision is the best we have.

Studies generally report results similar to human performance; this extends to the fact that more complicated classification systems suffer from lower reliability and accuracy, mirroring clinical practice. AI does not exceed surgeon performance for fracture classification.^{67, 84} These classification tasks, like detection tasks, are ubiquitous among early studies using AI models.^{67,}

^{84, 85}

These are examples of supervised learning: images are labeled by humans and a CNN is trained based on these truths. In contrast, unsupervised learning strategies can also be applied: the computer finds patterns in pathoanatomy on images, while outcome variables are given for automated cluster analyses. These AI algorithms do not just replicate human thinking processes—classification of fractures—but should aim to exceed them by offering new classifications.^{1, 22}

Prediction of Clinical Outcomes Based on Images: Multimodal Approach

Beyond automated fracture detection, a step change in AI applications includes the prognostic and diagnostic utility of these models to advance clinical outcomes. The prediction of clinical outcomes, based on trauma radiographs potentially augmented with a patient's personal characteristics, provides a multimodal inputs and outputs that may influence decision support. For instance, for patients with distal radius fractures, we can not only detect, segment, and

classify the fractures (see Fig. 9-5), but also gauge the likelihood of loss of threshold stability [4] (Fig. 9-6). This may guide ongoing treatment with a cast or conversion to offer fracture fixation to our patients in a shared decision-making process.

This type of data-driven approach to decision making is otherwise challenging, if not impossible, where surgeon-based estimations of fracture instability carry a 54% accuracy based on trauma and reduction radiographs, improving to 70% with advanced imaging based on CT interpretations as discovered in online experiments.⁴⁷ In contrast, a CNN can be trained with an area under the curve of 0.83 and 76% accuracy to predict fracture instability based on radiographs alone.⁴⁷

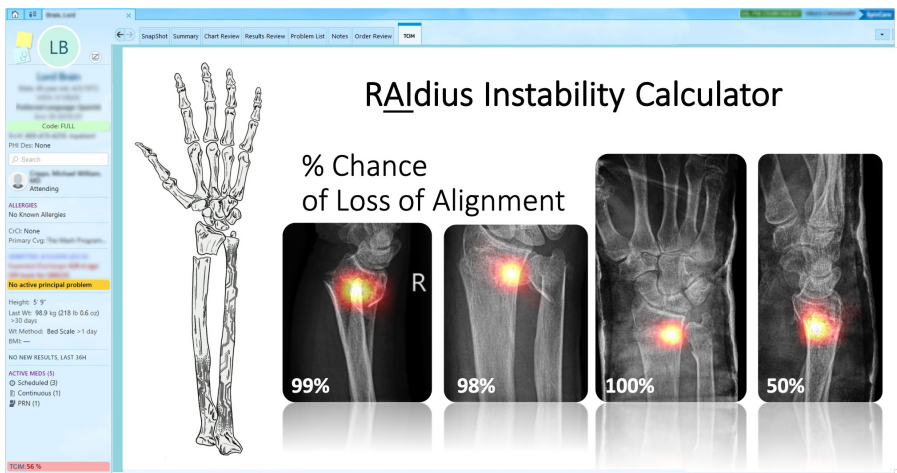


Figure 9-6. Prediction of distal radius fracture instability. For patients with distal radius fractures, we can not only detect, segment, and classify the fractures (see Fig. 9-5), but also gauge the likelihood of loss of threshold stability.

This line of inquiry is more challenging; for example, identifying patients at risk for failure of their hip fracture fixation based on FAITH data³³ applying machine learning techniques without imaging data failed¹⁰². However, we are confident that advancements will be made when trauma data sets become richer and AI methodology more sophisticated.

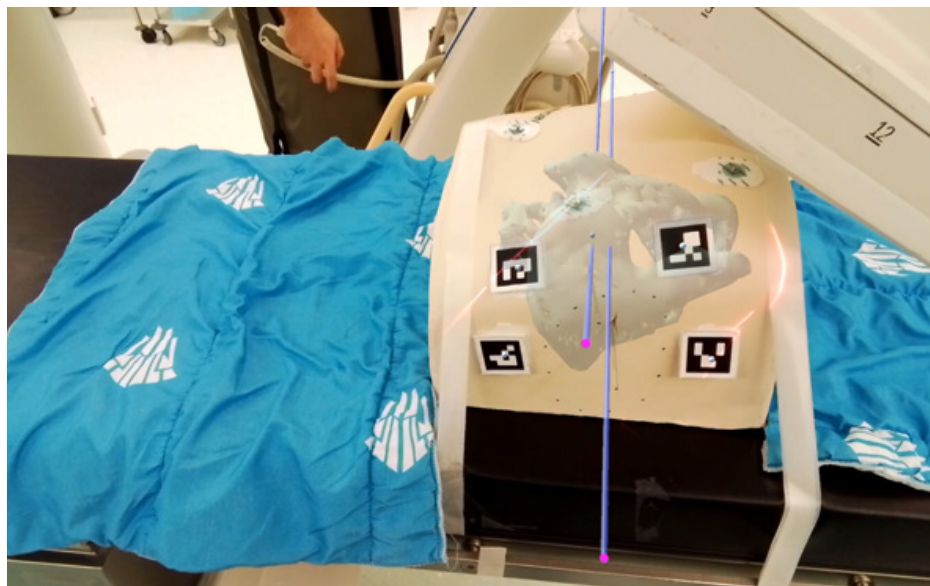
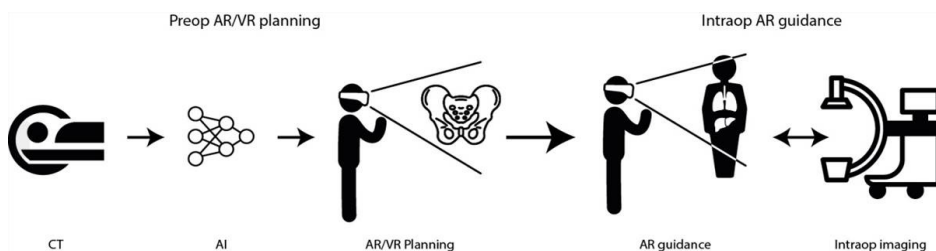


Figure 9-7. A: Using holograms in surgical preparation enhances the understanding of anatomy and pathology, facilitating spatial tasks like patient positioning and surgical trajectory planning. B: An example in orthopaedic trauma is augmented reality-guided sacroiliac screw placement. (Courtesy of Michiel Herteleer and Kobe Bamps, Leuven University, Belgium.)

Surgical Interventions: Augmented Reality

Accurate visualization of anatomic structures and fractures is pivotal for both diagnostic and therapeutic procedures. Deep learning-based AI models can process CT scans by segmenting anatomic structures and associated fractures with high precision.^{23, 64} Once segmented, these AI models can subsequently reconstruct a detailed 3D model of the fracture site. Such AI model outputs serve dual purposes: offering comprehensive anatomical insights and enabling individualized treatment planning.

Augmented reality technology further augments the utility of these AI model outputs by rendering them in the surgeon's field of view in real-time. This integration allows medical personnel, especially those in training, to interact with these 3D reconstructions in a more intuitive and immersive manner.³⁷ As a result, augmented reality could bridge the gap

between theoretical knowledge and practical application. For surgeons, the 3D models can be superimposed over the actual surgical site during both preoperative planning and the intraoperative procedure. This provides both the experienced surgeon and the trainee increased fidelity and information, including and improved understanding of the spatial relationship of the fracture fragments and the surrounding anatomy.

Such technology enables the integration and visualization of 3D models into clinical workflows, allowing surgeons to efficiently convert scans into 3D holograms in real life projection over the patient's affected anatomic site. In a cloud-based environment, medical scans can be automatically converted into 3D holograms, which can be experienced and manipulated in augmented reality by multiple users. Using holograms in surgical preparation enhances the understanding of anatomy and pathology, facilitating spatial tasks like patient positioning and surgical trajectory planning.^{32, 50} An example in orthopaedic trauma is augmented reality-guided sacroiliac (SI) screw placement (Fig. 9-7). However, clinical efficacy and accuracy have yet to be determined as compared to conventional methods.⁹⁴

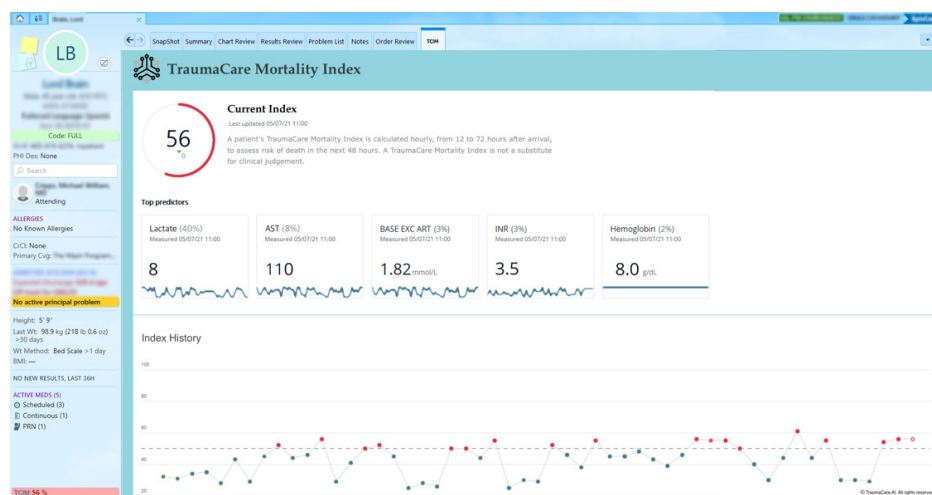


Figure 9-8. Example of a machine learning-driven probability calculator in electronic medical records, also known as predictive models, which use abstract data such as patient characteristics and lab values—in this clinical scenario, to predict mortality in trauma patients.

Steps Toward Clinical Practice (see Fig. 9-2)

Model Development

Model development consists of two phases: training and testing. The process of training an AI model involves a continuous and iterative approach. For computer vision, the model learns from a data set that contain annotated (highlighted region of interest) and/or labeled (image-level classification) cases. Predictive models use abstract data such as patient characteristics

and laboratory values (Fig. 9-8). Each complete cycle through the entire training data set is termed an epoch. Striking the right balance of epochs is crucial. Too few leads to suboptimal performance, and too many results in overfitting, where the model becomes exceptionally accurate on the training data but fails on new data.

The duration of an epoch can vary, spanning from hours to days, which is why computer scientists often work with batches (essentially small parts of the data set, consecutively) to monitor the learning rate of a model and to identify a plateau. This plateau represents a point at which there is a substantial diminishing return in the model's improvement with each pass of a batch during training. Throughout this training phase, the model tries to discern the essential data characteristics that align with the ground truth.

Subsequently, in the testing phase, the model evaluates instances it has not encountered previously and generates predictions based on its learned patterns. This is referred to as internal validation. Performance assessment is conducted by comparing these model predictions with the ground truth. It is crucial to acknowledge that during testing, the model's performance may not consistently mirror its real-world effectiveness, particularly when handling infrequent cases that are underrepresented in the data set. These rare instances often carry clinical significance, demanding a balance between effort invested and clinical relevance.

To mitigate this challenge, a common approach in computer vision is to introduce randomness into the data set, such as by applying image rotations, random cropping, stretching, and horizontal or vertical mirroring. This forces the neural network to identify features that remain unchanged despite manipulations and is commonly known as data augmentation.

External Validation

External validation is the next phase after internal validation (often referred to as the test set) when assessing a model. In external validation, the focus shifts from reproducibility to transportability. To ensure the adaptability of a model to various environments, several factors must be considered. These factors include demographic disparities, operator-dependent radiologic variables (e.g., angles, rotations, contrast enhancement, and radiation doses during radiograph or CT procedures), as well as variations in the make and quality of radiography equipment. It is imperative to thoroughly evaluate these factors before deploying a model to a different institution or setting.

The significance of this evaluation becomes clear in instances where models exhibit impressive performance on radiographs from a general population but struggle when confronted with diagnostically challenging cases. Typically, the actual model performance, assessed through external validation, tends to be lower compared to the performance measured using the data set employed during the model's development phase. This performance dip underscores the substantial differences in image characteristics between the

internal validation and geographical external validation data sets, underscoring the importance of geographical external validation.¹⁶

Prospective Validation and Clinical Efficacy

When a model has been through internal and external validation, confirming its predictive accuracy and robustness, the next step is to prospectively validate it—in real-time—and evaluate its performance with the struggles of daily practice. We often encounter radiographs shot from suboptimal angles, over- or underexposed images, and a great variance in demographics presenting to the emergency department, all of which the model needs to be able to handle. One of the great challenges is having a model incorporated in the imaging software used by the hospital. Ideally, the output of the model should be presented as an extra image or report to the clinician. However, when validating a model, this will inherently introduce a bias. One way to prevent a model from biasing clinicians in their decision making, before it is proven to be accurate in a clinical setting, is to blind the results from the clinician. In other words, we perform our assessment and treatments as we would normally do, and in the background the model presents their prediction or treatment to a researcher, which can then over many cases evaluate the real performance of the model in the hospital without compromising patient care.

Self-Learning

If the model has proved to be accurate and trustworthy in a clinical setting, it should be employed in daily practice to further improve its accuracy and thus its efficacy. In a perfect scenario, doctors would use the model's predictions in their decision making, and the model learns from the note if it was correct or not, so that it may learn from each encounter. Self-learning could be considered an area of controversy because it can be seen as doctors essentially training a model to replace their expertise. This poses another challenge: If models replace our expertise in certain diagnoses or predictions, how will we train the new generation of doctors to be able to correct or deviate from the model when they deem necessary, like the guidelines we may deviate from daily?



FUTURE PERSPECTIVE

ChatGPT⁸⁹ output:

"Future Perspectives

The burgeoning synergy between AI and orthopaedic trauma presents an exciting trajectory of advancement. The evolution of AI algorithms holds promise for real-time intraoperative decision support, where AI systems provide instantaneous insights to surgeons during

procedures. Collaborative robotic systems, guided by AI, may redefine surgical workflows and enable remote consultations, transcending geographical barriers. The harnessing of multi-modal data, encompassing patient genetics, biomechanics, and lifestyle factors, could pave the way for personalized treatment regimens tailored to individual needs. Moreover, AI's potential to decipher complex molecular interactions could engender breakthroughs in regenerative therapies and fracture healing mechanisms.

In summation, the convergence of AI and orthopaedic trauma signifies a watershed moment in modern medicine, heralding a future characterized by precision, efficiency, and patient-centered care. This chapter embarks on a comprehensive exploration of this dynamic alliance, delving into the foundational principles of AI, its current applications within the domain of orthopaedic trauma, and the tantalizing vistas that lie ahead. As we traverse this uncharted terrain, the symbiotic interplay between AI's computational prowess and the nuanced expertise of orthopaedic practitioners promises to redefine the boundaries of possibility and reshape the trajectory of fracture management.

Large Language Models

The explosion of interest in generative AI⁴⁴—a class of machine learning models that are trained using large volumes of text, audio, image, or video data to generate new, natural, human-like content—holds the promise of clinical applications of AI in orthopaedics beyond the examples presented in this chapter.

The final AI example we will discuss involves the application of LLMs, also known as chatbots, with ChatGPT's output included in this chapter. LLMs have potential high-value applications in our outpatient clinics, including serving and acting as a virtual assistant, and linking triage and health advice to automated appointment scheduling.

One of few potential threats, however, is the so-called AI hallucination, which can fuel misinformation, such as the potential answers to a clinical question such as, “Is this distal radius fracture stable or will it lose threshold alignment during follow-up and become unstable?”—leading to intensified distress and unhelpful thinking through medical misinformation.

Generative AI and LLMs are expanding the reach of AI technology within health care and our daily lives. While such technology is poised to enable a variety of value-generating opportunities, stakeholders should be aware of potential threats and supportive strategies for data quality, model development and regulation, and legal and ethical guardrails.

CONCLUSION

AI in orthopaedic trauma shows great promise. There are many barriers to its adoption: overcoming bias in algorithms, incorporating new applications into clinical workflow, regulatory approval, and billing.⁹⁷

The ultimate goal is data-driven decision support to overcome human biases and to support shared decision making that focuses on patients' values and preferences.



PART II

**Advanced Imaging in Paediatric
Orthopaedic Trauma**



Triplane Ankle Fracture Patterns in Paediatric Patients

Extent of Physeal Closure Does Not Dictate Pathoanatomy

J. Prijs

J. Rawat

K. ten Duis

F.F.A. IJpma

J.N. Doornberg

B. Jadav

R.L. Jaarsma

ABSTRACT

Aims

Triplane ankle fractures are complex injuries typically occurring in children aged between 12 and 15 years. Classic teaching that closure of the physis dictates the overall fracture pattern, based on studies in the 1960s, has not been challenged. The aim of this paper is to analyze whether these injuries correlate with the advancing closure of the physis with age.

Methods

A fracture mapping study was performed in 83 paediatric patients with a triplane ankle fracture treated in three trauma centres between January 2010 and June 2020. Patients aged younger than 18 years who had CT scans available were included. An independent Paediatric Orthopaedic Trauma Surgeon assessed all CT scans and classified the injuries as n-part triplane fractures. Qualitative analysis of the fracture pattern was performed using the modified Cole fracture mapping technique. The maps were assessed for both patterns and correlation with the closing of the physis until consensus was reached by a panel of six surgeons.

Results

Fracture map grouped by age demonstrates that, regardless of age (even at the extremes of the spectrum), the fracture lines consolidate in a characteristic Y-pattern, and no shift with closure of the physis was observed. A second fracture map with two years added to female age also did not show a shift. The fracture map, grouped by both age and sex, shows a Y-pattern in all different groups. The fracture lines appear to occur between the anterior and posterior inferior tibiofibular ligaments, and the medially fused physis or deltoid ligament.

Conclusion

This fracture mapping study reveals that triplane ankle fractures have a characteristic Y-pattern, and acknowledges the weakness created by the physis, however it also challenges classic teaching that the specific fracture pattern at the level of the joint of these injuries relies on advancing closure of the physis with age. Instead, this study observes the importance of ligament attachment in the fracture patterns of these injuries.

INTRODUCTION

Paediatric patients often present with unique ankle fracture patterns that are different from the adults.¹ For example, transitional fractures of the distal tibia that occur in characteristic patterns are believed to follow the typical progressive closure of the physis. Generally, the distal tibial physis starts to close at the age of 12 years in females and 13 years in males,² and is completed over 18 months.³ The physis closes according to a characteristic pattern that starts in the (centro-)medial part of the tibia, and consequently continues laterally along the posterior tibia and finishes at the anterolateral tibia (Figure 1). The configuration of triplane fractures has been thought to follow the characteristic closure pattern, based on the radiological study by Kleiger and Mankin³ in 1964. However, to the best of our knowledge, this idea has not been challenged in the literature, and has become part of the established teaching.

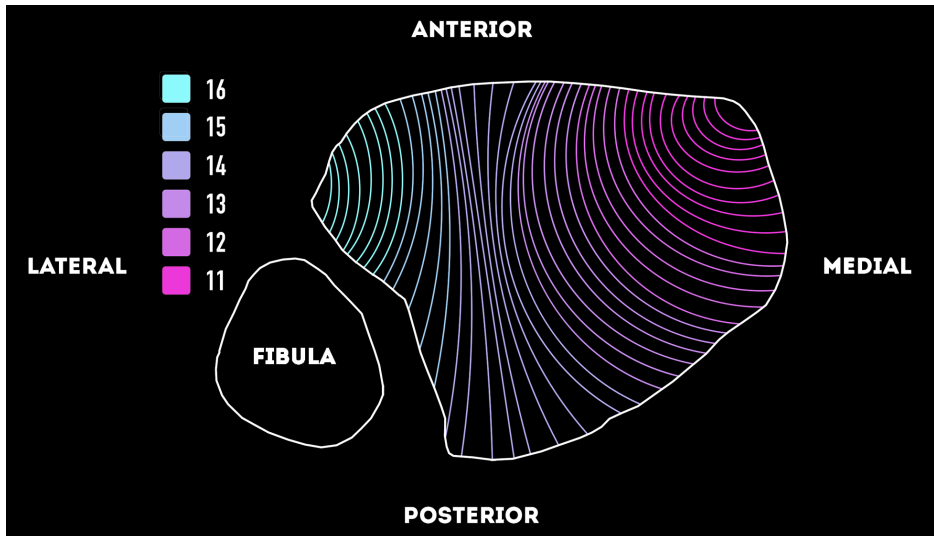


Figure 1. Fictive map presenting the expected fracture patterns by age, based on the characteristic closure of the physis as presented in the paper by Kleiger and Mankin in 1964.

Triplane and Tillaux fractures are transitional fractures of the tibia. Triplane fractures extend into three planes: coronal (metaphysis), sagittal (epiphysis), and axial (physis). Fracture pattern recognition of these injuries is often challenging,⁴ and unsatisfactory reduction can be the result when fracture characteristics are not fully appreciated. Tillaux fractures are antero-lateral avulsion fractures that are bordered proximally by, and extend into, the physis. These are thought to occur at a later age, as they are believed to come about due to advanced closure of the physis, where only the anterolateral part is still open.⁵

Fracture maps help to improve the understanding and recognition of fractures and patterns,⁶⁻¹² and are available to increase knowledge of triplane fractures in children. Based

on axial CT scan images, maps of superimposed fracture lines can help to define fracture patterns and location, especially in relation to physeal closure. Fracture maps have proved an asset in elucidating complex fracture patterns of scaphoid, radial head, tibial plateau, proximal humerus, and posterior malleolar fractures.⁹⁻¹⁴ To date, only one study has used fracture maps in triplane fractures,⁸ focusing on preoperative strategies for screw placement, but the authors did not explore the pathoanatomy of these injuries in any depth.

The purpose of this study is to explore the pathoanatomy of distal tibial triplane fractures, to correlate the classic teaching of advancing physeal closure with overall fracture pattern and answering the question: do triplane fracture patterns of the distal tibia correlate with the hypothesized characteristic closing of the physis (Figure 1) with age, at the level of the joint?

METHODS

This study was approved by the Institutional Review Boards of the participating centres, according to the Declaration of Helsinki.¹⁵

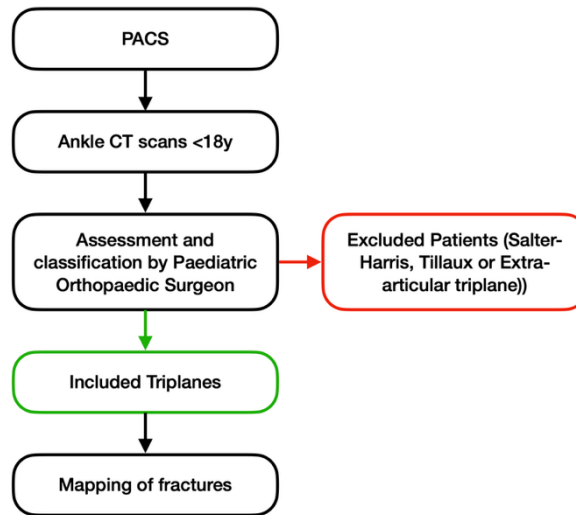
Patient selection.

For this mapping study, all consecutive ankle CTs from two level I trauma centres and one specialized paediatric hospital from two different countries (the Netherlands and Australia) between January 2010 and June 2020, in patients aged younger than 18 years, were retrospectively reviewed by a fellowship-trained paediatric orthopaedic surgeon (JR).

The criteria used for inclusion were: age between 10 and 18 years; availability of a preoperative CT scan with slices of 1 mm thickness or less; and presence of a transitional distal tibial fracture. Exclusion criteria were: presence of an old or pathological fracture; CT scan unavailable or only with slices thicker than 1 mm; and presence of associated conditions such as, but not limited to, osteomyelitis, joint infections, or other fractures (excluding fibula fractures).

Fracture classification.

Included CT scans were exported from various Picture Archiving and Communication Systems as Digital Imaging and Communications in Medicine (DICOM) files. Patient characteristics such as age at injury, sex, and side of injury were recorded. All CTs were assessed, and fractures were classified, by a fellowship-trained paediatric orthopaedic surgeon (JR), by Salter-Harris type¹⁶, or as triplane (AO/OTA 43t-E/6.1, defined as an epiphyseal fracture in the axial, coronal, and sagittal plane, including a posterior metaphyseal component)¹⁷ or Tillaux (AO/OTA 43t-E/5.1)¹⁷, using the axial, coronal, sagittal, and 3D CT reconstructions in Horos v. 3.3.6 (Nimble, USA). When all included patients were assessed and classified according to their respective fracture type, anything other than triplane fractures was excluded (Supplemental Figure 1).



Supplemental figure 1 — Inclusion Workflow

Fracture mapping.

Fractures were digitally recorded using standardized axial views, at 3 mm above the distal tibial subchondral surface, a reliable method as defined by Cole et al⁷ in the articular mapping of adult Pilon fractures. A healthy right paediatric ankle was chosen as the template. Recorded CT slices were (if needed) mirrored, resized, rotated, and normalized to fit the dimensions of the template tibia in Photoshop 2020 v. 21.1.0 (Adobe, USA), by aligning specific tibial landmarks: with the intact posteromedial tibia as the starting point. Subsequently, fractures with multiple displaced fragments were reduced digitally. Finally, for each recorded axial slice, the fracture lines were tracked by hand using the built-in pencil brush with a size of 4 pixels and superimposed on the template (Figure 2).

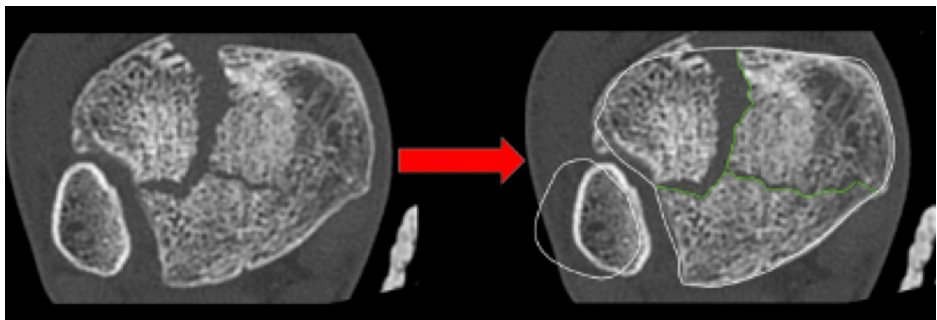


Figure 2. On the lefthand side an axial slice at 3mm above the plafond is fitted within the template and then the fracture line is recorded by hand.

Only fracture lines through cortical bone were considered major fracture lines, however minor fracture lines, which do not exit cortical bone, were also recorded for analysis in a separate fracture map. Physeal closure was recorded to create a fracture map for open physis versus partially closed. Complete versus partial closure was determined on the preoperative CT scan using the coronal, sagittal, and axial views by a fellowship-trained paediatric orthopaedic surgeon (JR).

Patient characteristics.

A total of 182 patients were identified, of whom 99 with a Salter-Harris II, III, and IV, Tillaux, and extra-articular triplane fractures were excluded. A total of 83 patients had a triplane fracture, and images of these patients were used for further fracture mapping analysis (Table I). Overall, 45 (54%) of these patients were male, and 38 (46%) were female. Mean ages were 14.4 years (11 to 18) for males and 12.5 years (11 to 15) for females ($p < 0.001$). In nine out of 83 patients (11%), the physis was completely open, in contrast to partial fusion detected in the other patients (Table I).

Table 1. Baseline demographics and fracture characteristics

Characteristics	Total
Male, n (%)	45 (54)
Female, n (%)	38 (46)
Mean age at trauma, yrs (range)	13.5 (11 to 18)
Male	14.4 (11 to 18)
Female	12.5 (11 to 15)
Side of ankle fracture, n (%)	
Left	28 (34)
Right	55 (66)
Fracture classification, n (%)	
II-Part	55 (66)
III-Part	25 (30)
IV-Part	3 (4)
Physis closure, n (%)	
Open	9 (11)
Partially Closed	74 (89)

Statistical analysis.

Patient demographic details were reported as means and ranges, or percentages. Comparison of means between two groups was by using the independent-samples *t*-test and, if non-normally distributed, the Mann-Whitney U test was used. Fracture maps were visually assessed and compared to the expected shift in fracture patterns based on physeal closure (Figure 1) by a panel of six fellowship-trained surgeons (JR, KtD, FIJ, JD, BJ, RJ), and consensus was reached for description of the fracture maps grouped per age, adjusted age (added two years to female biological age), sex, age and sex together, and closure of the physis (open versus partially closed).

RESULTS

Fracture patterns: comparisons with the expected fracture lines.

A fracture map was created showing the expected transition of fracture lines following the closure of the physis (Figure 1). The overview of all superimposed triplane fracture lines sorted by sex presents a characteristic fairly standard Y-pattern for both sexes (Figure 3).

Fracture map grouped by age at time of injury (Figure 4) demonstrates that regardless of age (even at the extremes of the spectrum), the fracture lines consolidate in the characteristic Y-pattern, and no colour shift, which would correspond to the physis closing from anteromedial to anterolateral with advancing age, was observed.

A second fracture map was produced with two years added to female age, the mean age difference between sexes found in our cohort, to try to negate the bone-age difference from males (Figure 5).¹⁸ This fracture map also did not reveal a colour shift along the supposed physeal closure with age (Figure 1). The fracture map grouped by both age and sex (Figure 6) again shows the Y-pattern in all different age groups and sex.

Even when mapping the minor fractures lines (i.e. those that do not go through cortical bone; Figure 7), the characteristic Y-pattern can still be observed. Like major fracture lines, the minor fracture lines seem to occur in the vector of the forces between the anterior and posterior inferior tibiofibular ligaments (Figure 7), and the medially fused physis or deltoid ligament.

When sorted by physeal closure (open versus partially closed, Figure 8), the Y-pattern can be appreciated for both open and partially closed. None of the produced fracture maps conformed to the expected transition in fracture lines as presented in Figure 1.



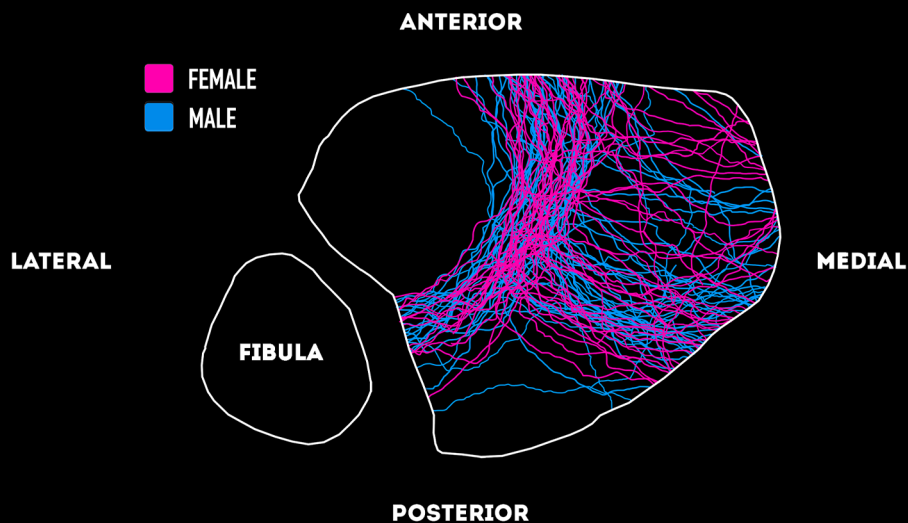


Figure 3. Fracture maps superimposed on the right-sided template. Girls are presented in pink, and boys in blue.

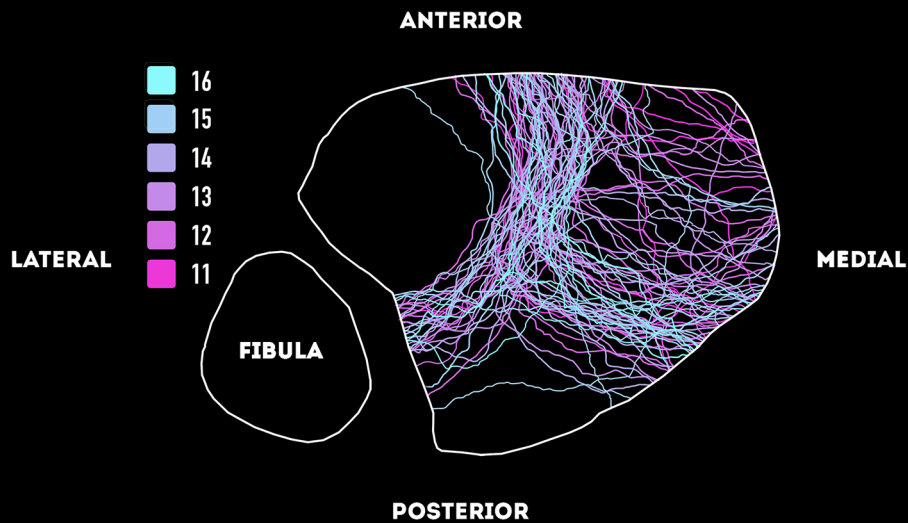


Figure 4. Triplane and Tillaux fracture maps sorted by age. Ages are presented from young to old (pink to light blue).

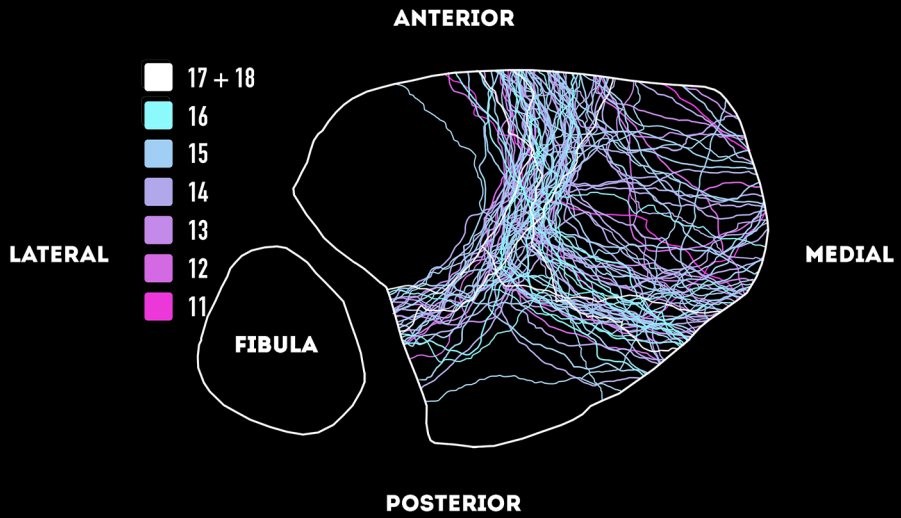


Figure 5. Fracture maps grounded by age, where girls' age was adjusted by two years to match the suggested two year difference in bone age.

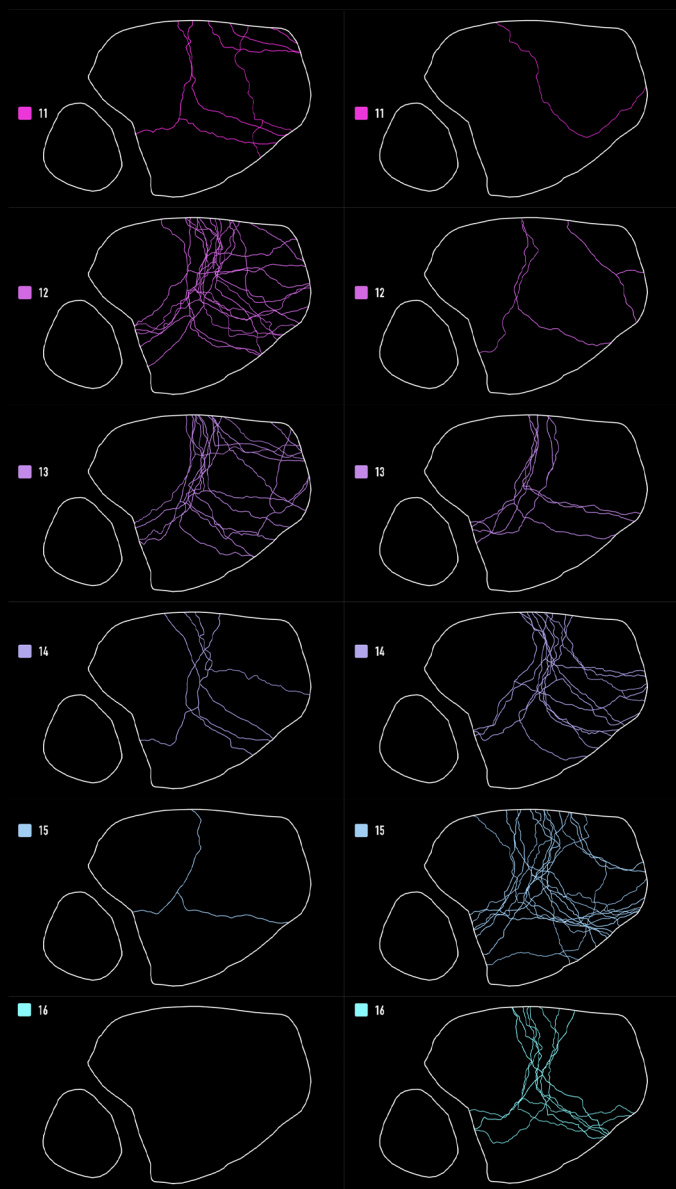


Figure 6. Fracture maps split by age and gender. On the left girls, on the right boys and from top to bottom from 11 to 16 years old.

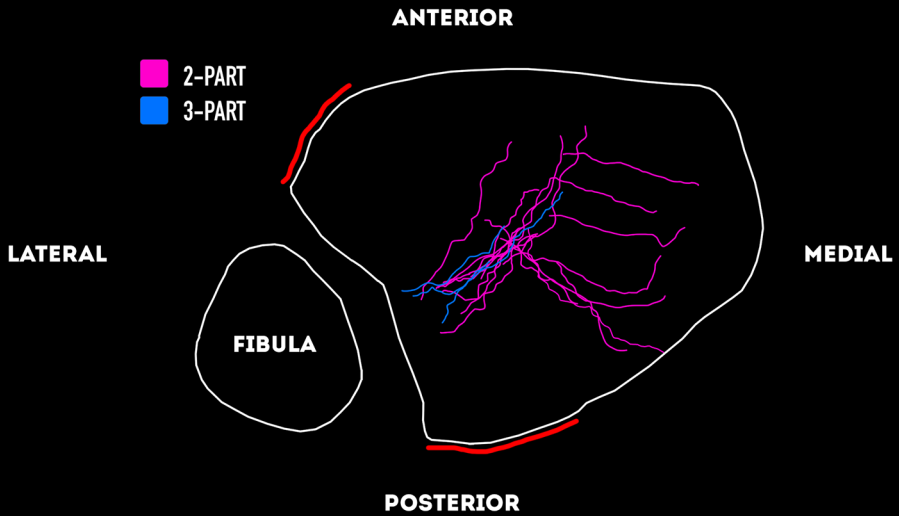


Figure 7. Minor fracture lines recorded from 2- (pink) and 3-part (blue) Triplane fractures, with in red the attachments (anteriorly) of the Anterior Tibiofibular Ligament and (posteriorly) the Posterior Tibiofibular Ligament.

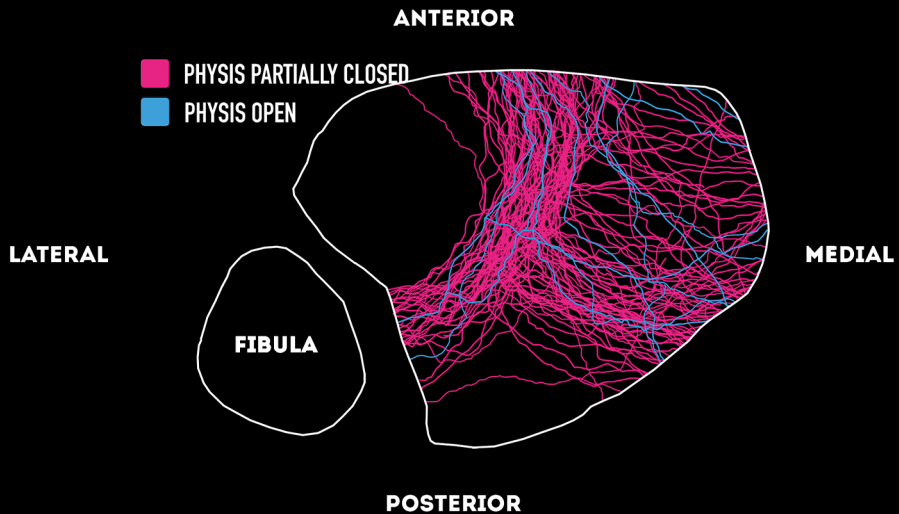


Figure 8. Fracture lines sorted by open physis (blue) versus partially closed (pink).

DISCUSSION

This retrospective multicentre cohort study set out to evaluate whether triplane fractures correlate with the asymmetric closing physis (Figure 1) and age, at the level of the joint. However, we did not find that the hypothesized shift or pattern of fractures following closure of the physis from anteromedial to anterolateral with age. In addition, 11% of triplane fractures occurred with a completely open physis. Thus, our study acknowledges the weakness created by the physis in the paediatric distal tibia; however, contrary to what was previously assumed, it challenges the theory that the closing physis is the principal determiner of the fracture pattern at the joint level in these injuries.

The major strengths of our study include the collection of thin-slice CT scans from two level I trauma centres and one specialized paediatric hospital covering two different continents, along with collaboration of an experienced paediatric orthopaedic surgeon to classify the different fracture types and qualitatively assess the fracture maps. Another strength is that all triplane fractures that present in the emergency departments of our centres undergo a CT scan, reducing the effect of a possible selection bias. Further strengths are the use of standardized axial views, and the largest cohort of patients presented in the literature. Limitations are that this study was retrospective, which reduces analysis of the imaging and the recorded parameters following the trauma. Another limitation is that it was not possible to quantify the level of physeal closure due to its complex 3D geometry, and thus we grouped partially closed physes of varying degrees together. There does not appear to be a reliable method to quantify closure on the CT scan of the fractured side, and, in addition, we did not have a CT scan of the contralateral unaffected side for comparison.

Our study of 83 patients is the largest dataset of triplane fractures worldwide. Hadad et al⁸ used a fracture map of triplane fractures to suggest optimal treatment for each respective II-, III-, IV-, and V-part fractures. The current study builds on this by including a larger dataset and fracture maps that are grouped not only by classification but also by sex, age, minor fracture lines, and extent of closure of the physis. Several studies in the literature have reported on the characteristic Y-pattern (or 'Mercedes-Benz star')^{4,19-21}, which was also found in our study. In comparison with other papers which present methods for optimal fixation or evaluate the use of CT scans in these injuries^{4,8,19,22}, the current study is the first to challenge the theory originated by Kleiger and Mankin in 1964³, and improves our understanding of the effect of the physis and the pathoanatomy of these complex injuries. In addition to other studies in literature²³⁻²⁶, our study reports a subset of patients with a completely open physis.

Overall, the fracture maps between females and males are very similar. Nonetheless, according to the established teaching, fracture lines in different age groups can be expected to follow the characteristic fusion of the physis, and thus produce a predictable pattern per individual age; however we did not observe this. Most importantly, one would expect there to be a difference between the fracture maps of 11 and 16 years. Interestingly, when fracture

maps were grouped by age, the distribution was random and no shift along the fusion path of the physis could be detected. This could be due to masking by the commonly held view that there is a two-year difference in bone age between males and females; however the distribution of the subsequent map with two years added to female biological age to overcome this did not reveal a pattern related to physeal closure.

Clement and Worlock²⁵ built on the findings of Kleiger and Mankin³ in 1964 and hypothesized in 1987 that presence of a medial hump in the epiphysis could be the leading factor in triplane fractures patterns with an open physis. This hump was posited to stabilize the open physis medially, similar to the early medial fusion.²⁵ Together with the deltoid ligament, medial stabilization due to a medial hump, and laterally with the anterior inferior tibiofibular ligament (AITFL) and posterior inferior tibiofibular ligament (PITFL) could explain the factors causing the Y-pattern. However, there are many fracture lines, including those with an open physis, that cross the medial hump, which questions the role of the medial hump in the pathoanatomy of these injuries, and instead perhaps medial fusion or the deltoid ligament are greater determinators. We believe that there are two reasons why the medial fracture line has more variation. First, the fibers of the deltoid ligament run in a vertical fashion (as opposed to the anterior/posterior TFL where the direction is horizontal) and it is easier to provoke a fracture in line with the fibers. More importantly though, the actual attachment of the deltoid ligament is below the level we have looked for fracture lines. Therefore, protection by the deltoid ligament is limited, if not lacking at that exact point, hence providing the opportunity for more variation in the fracture line. However, we believe that most medial fracture lines exit the cortex in line with the characteristic Y-pattern, which has been confirmed by other studies.^{4,19-21}

The anterolateral fragment seems to be a consistent component with minimal variation in size. It could be argued that there is a Tillaux fracture in each III-part triplane, which could be explained by the way the AITFL pulls on the anterolateral part of the distal tibia during external rotation. The ubiquity of the Tillaux fragment in triplane fractures is an argument against the physis as the primary reason for the configuration of these unique fractures. In addition, the convex Tillaux fragment fits well with a pulling force from the AITFL and seems to go against the expected concave fracture lines if they would follow closure of the physis as the main determiner. The fact that ligaments play an important role in these injuries is further strengthened by the superimposed map of the minor fracture lines, which are all located between the two tibiofibular ligaments and the medial deltoid ligament and follow the characteristic Y-pattern. However, our study is unable to differentiate between ligaments serving an 'active' role by pulling on the fragments or a 'passive' role in protecting those areas of the tibia from a fracture exiting; or, indeed, acting as a stress riser to the adjacent bone and therefore a common place for a fracture line to occur. In adults, anterolateral avulsions of the tibia (Tillaux-Chaput fracture) and posterolateral avulsions (Volkmann fracture) by the tibiofibular ligaments, albeit rare injuries have been reported.²⁷⁻³¹ These injuries are found to occur in isolation,^{28,30,31} and in exceptional cases simultaneously,^{27,29} indicating that even in



adults with a fused physis, these ligaments produce forces capable of avulsing parts of the tibia. This would support our theory that in a distal tibia weakened by the physis, the ligaments play an important role in the pathoanatomy of the injury. In addition, the Y-pattern described in our study is very similar to the pattern reported in adult tibial pilon fractures by Cole et al,⁷ which are injuries caused by high-energy axial impacts on the tibial plafond, in contrast to the lower-energy rotational type triplane fractures.

This study acknowledges the weakness that the physis creates in the paediatric distal tibia. However, it challenges the theory that the physis is the principal determiner of the configuration of the fracture pattern in triplane fractures at the joint level. Instead we advance the view that the role of the insertions or forces of the tibiofibular ligaments are more important determinators, rather than the actual extent of closure of the physis. Therefore, we conclude that age at time of injury and assumed extent of physeal closure with age are not the main determinants for the configuration of these injuries at the level of the joint.

Take home message

The physis creates a weakness in the paediatric distal tibia, and the role of the insertions or forces of the tibiofibular ligaments are the most important determinators of the triplanar fracture pattern.

Therefore, age at time of injury, and assumed extent of physeal closure with age, are not the main determinants for the configuration of these injuries at the level of the joint.

SUPPLEMENTARY MATERIAL

Visual overview of inclusion process of this study.

REFERENCES

1. Cancino, B., M. Sepulveda, and E. Birrer, Ankle fractures in children. *EFORT Open Rev*, 2021. 6(7): p. 593-606.
2. Kump, W.L., Vertical fractures of the distal tibial epiphysis. *Am J Roentgenol Radium Ther Nucl Med*, 1966. 97(3): p. 676-81.
3. Kleiger, B. and H.J. Mankin, Fracture of the Lateral Portion of the Distal Tibial Epiphysis. *J Bone Joint Surg Am*, 1964. 46: p. 25-32.
4. Schneidmueller, D., et al., Triplane fractures: do we need cross-sectional imaging? *Eur J Trauma Emerg Surg*, 2014. 40(1): p. 37-43.
5. Dias, L.S. and C.R. Giegerich, Fractures of the distal tibial epiphysis in adolescence. *J Bone Joint Surg Am*, 1983. 65(4): p. 438-44.
6. Armitage, B.M., et al., Mapping of scapular fractures with three-dimensional computed tomography. *J Bone Joint Surg Am*, 2009. 91(9): p. 2222-8.
7. Cole, P.A., et al., The pilon map: fracture lines and comminution zones in OTA/AO type 43C3 pilon fractures. *J Orthop Trauma*, 2013. 27(7): p. e152-6.
8. Hadad, M.J., B.T. Sullivan, and P.D. Sponseller, Surgically Relevant Patterns in Triplane Fractures: A Mapping Study. *J Bone Joint Surg Am*, 2018. 100(12): p. 1039-1046.
9. Hendrickx, L.A.M., et al., Incidence, Predictors, and Fracture Mapping of (Occult) Posterior Malleolar Fractures Associated With Tibial Shaft Fractures. *J Orthop Trauma*, 2019. 33(12): p. e452-e458.
10. Mellema, J.J., et al., Fracture mapping of displaced partial articular fractures of the radial head. *J Shoulder Elbow Surg*, 2016. 25(9): p. 1509-16.
11. Molenaars, R.J., L.B. Solomon, and J.N. Doornberg, Articular coronal fracture angle of posteromedial tibial plateau fragments: A computed tomography fracture mapping study. *Injury*, 2019. 50(2): p. 489-496.
12. Turow, A., et al., 3D mapping of scaphoid fractures and comminution. *Skeletal Radiol*, 2020. 49(10): p. 1633-1647.
13. Molenaars, R.J., et al., Tibial Plateau Fracture Characteristics: Computed Tomography Mapping of Lateral, Medial, and Bicondylar Fractures. *J Bone Joint Surg Am*, 2015. 97(18): p. 1512-20.
14. Hasan, A.P., et al., Fracture line morphology of complex proximal humeral fractures. *J Shoulder Elbow Surg*, 2017. 26(10): p. e300-e308.
15. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013;310(20):2191-2194.
16. Meinberg, E.G., et al., Fracture and Dislocation Classification Compendium-2018. *J Orthop Trauma*, 2018. 32 Suppl 1: p. S1-S170.
17. Brown, S.D., et al., Analysis of 51 tibial triplane fractures using CT with multiplanar reconstruction. *AJR Am J Roentgenol*, 2004. 183(5): p. 1489-95.
18. Rapariz, J.M., et al., Distal tibial triplane fractures: long-term follow-up. *J Pediatr Orthop*, 1996. 16(1): p. 113-8.
19. Schnetzler, K.A. and D. Hoernschemeyer, The pediatric triplane ankle fracture. *J Am Acad Orthop Surg*, 2007. 15(12): p. 738-47.
20. Gaudiani, M.A., D.M. Knapik, and R.W. Liu, Clinical Outcomes of Triplane Fractures Based on Imaging Modality Utilization and Management: A Systematic Review and Meta-analysis. *J Pediatr Orthop*, 2020. 40(10): p. e936-e941.
21. El-Karef, E., et al., Triplane fracture of the distal tibia. *Injury*, 2000. 31(9): p. 729-36.



22. van Laarhoven, C.J., R.S. Severijnen, and C. van der Werken, Triplane fractures of the distal tibia. *J Foot Ankle Surg*, 1995. 34(6): p. 556-9; discussion 594-5.
23. Clement, D.A. and P.H. Worlock, Triplane fracture of the distal tibia. A variant in cases with an open growth plate. *J Bone Joint Surg Br*, 1987. 69(3): p. 412-5.
24. von Laer, L., Classification, diagnosis, and treatment of transitional fractures of the distal part of the tibia. *J Bone Joint Surg Am*, 1985. 67(5): p. 687-98.
25. Kose, O., et al., Isolated Adult Tillaux Fracture Associated With Volkmann Fracture-A Unique Combination of Injuries: Report of Two Cases and Review of the Literature. *J Foot Ankle Surg*, 2016. 55(5): p. 1057-62.
26. Kumar, N. and M. Prasad, Tillaux fracture of the ankle in an adult: a rare injury. *J Foot Ankle Surg*, 2014. 53(6): p. 757-8.
27. Mansur, H., et al., Adult Tillaux Fracture in Association with Volkmann and Maisonneuve Fractures: A Case Report. *J Am Podiatr Med Assoc*, 2019. 109(6): p. 477-481.
28. Oak, N.R., et al., Isolated adult Tillaux fracture: a report of two cases. *J Foot Ankle Surg*, 2014. 53(4): p. 489-92.
29. Sharma, B., I.S. Reddy, and C. Meanock, The adult Tillaux fracture: one not to miss. *BMJ Case Rep*, 2013. 2013.



Lasseter Highway, Northern Territory, Australia



**Understanding the Mechanism of Injury and
Fracture Pattern of Paediatric Triplane Ankle
Fractures versus Adult Trimalleolar Fractures**

J. Prijs
J. Rawat
K. ten Duis
N. Assink
J. S. Harbers
J.N. Doornberg
B. Jadav
R.L. Jaarsma
F.F.A. IJpma

ABSTRACT

Aims

Paediatric triplane fractures and adult trimalleolar ankle fractures both arise from a supination external rotation injury. By relating the experience of adult to paediatric fractures, clarification has been sought on the sequence of injury, ligament involvement, and fracture pattern of triplane fractures. This study explores the similarities between triplane and trimalleolar fractures for each stage of the Lauge-Hansen classification, with the aim of aiding reduction and fixation techniques.

Methods

Imaging data of 83 paediatric patients with triplane fractures and 100 adult patients with trimalleolar fractures were collected, and their fracture morphology was compared using fracture maps. Visual fracture maps were assessed, classified, and compared with each other, to establish the progression of injury according to the Lauge-Hansen classification.

Results

Four stages of injury in triplane fractures, resembling the adult supination external rotation Lauge-Hansen stages, were observed. Stage I consists of rupture of the anterior syndesmosis or small avulsion of the anterolateral tibia in trimalleolar fractures, and the avulsion of a larger Tillaux fragment in triplanes. Stage II is defined as oblique fracturing of the fibula at the level of the syndesmosis, present in all trimalleolar fractures and in 30% (25/83) of triplane fractures. Stage III is the fracturing of the posterior malleolus. In trimalleolar fractures, the different Haraguchi types can be discerned. In triplane fractures, the delineation of the posterior fragment has a wave-like shape, which is part of the characteristic Y-pattern of triplane fractures, originating from the Tillaux fragment. Stage IV represents a fracture of the medial malleolus, which is highly variable in both the trimalleolar and triplane fractures.

Conclusion

The paediatric triplane and adult trimalleolar fractures share common features according to the Lauge-Hansen classification. This highlights that the adolescent injury arises from a combination of ligament traction and a growth plate in the process of closing. With this knowledge, a specific sequence of reduction and optimal screw positions are recommended.

INTRODUCTION

Triplane fractures are a unique type of transitional fracture that occur during the gradual closure of the physis in children, and often present in those aged 12 to 15 years. The term triplane was applied in 1972 by Lynn¹ and describes a fracture in three planes: sagittal through the epiphysis; axially splitting the physis; and finally exiting the metaphysis in a coronal plane. Closure of the physis is thought to follow a pattern starting centromedially, extending then posteriorly and laterally, before closing completely anterolaterally. It was generally accepted that the extent of physeal closure determined the fracture pattern, based on earlier studies using radiographs.¹⁻⁴ It is assumed that triplane fractures result from a supination external rotation injury.^{5,6} However, further clarification of the exact sequence of injuries that occurs in different planes of the bone is required. Understanding of the precise nature and displacement of these fractures is essential for optimal treatment.

In contrast to the limited studies describing supination external rotation injuries in paediatrics, the injury mechanism is well described in adults. In 1954, Lauge-Hansen⁷ described various trauma mechanisms in adult ankle fractures, including the supination external rotation injury. Trimalleolar ankle fractures will have completed the four stages of a supination external rotation injury as described by Lauge-Hansen: rupture of the anterior tibiofibular ligament or avulsion of the Tillaux fragment; oblique fracture of the fibula at the level of the joint; fracture of the posterior malleolus; and finally rupture of the deltoid ligament or a medial malleolar fracture. For adolescents, it has recently been described that ligaments around the ankle joint - rather than only the physis - are a major contributor to the pathophysiology of triplane fractures.⁸ Our study explores whether the fracture pattern of triplane fractures follows the supination external rotation sequence of the Lauge-Hansen classification.

Fracture maps may help to improve our understanding of fracture patterns. Fracture maps consist of a collection of fracture lines from a specific type of fracture, superimposed and fitted to a template to improve the process of comparison and analysis. This helps in elucidating fracture patterns and suggestions for fixation of several types of injuries such as triplane, tibial pilon, scapular, posterior malleolar, tibia plateau, and scaphoid fractures.⁸⁻¹⁵ Using fracture maps, it would be helpful to compare triplane fractures to trimalleolar fractures that are assumed to share an identical trauma mechanism. From this our knowledge regarding the pathophysiology of paediatric fractures can be expanded. From a treatment perspective, reversing this sequence of injuries could help in understanding how to perform fracture reduction and establish optimal screw positions for these fractures.

We hypothesize that the injury mechanism of triplane fractures in adolescents is comparable to that of trimalleolar fractures in adults. Hence, our research question was: what are the differences and similarities between the adolescent triplane and the adult trimalleolar fractures for each stage of the Lauge-Hansen classification?



METHODS

In this comparative diagnostic imaging study, imaging data of paediatric patients with triplane fractures and adult patients with trimalleolar fractures were collected, and their fracture morphology was compared using fracture maps.

All patients aged between ten and 18 years with an ankle CT scan who presented between January 2010 and June 2020 in two level 1 trauma centres (University Medical Centre Groningen, the Netherlands, and Flinders Medical Centre, Adelaide, Australia) and one specialized paediatric hospital (Women's and Children's Hospital Adelaide, Australia) were identified. The imaging data of these patients were then reviewed by a fellowship-trained paediatric orthopaedic surgeon (JR) and classified as triplane (AO/Orthopaedic Trauma Association (AO/OTA) 43t-E/6.1¹⁶, defined as an epiphyseal fracture in the axial, coronal, and sagittal plane, including a posterior metaphyseal component) or Tillaux (AO/OTA 43t-E/5.1)¹⁶.

Imaging data of 83 patients with triplane fractures were collected for further fracture-mapping analysis. A total of 45 (54%) were male and 38 (46%) were female. Mean ages were 14.4 years (11 to 18) for males and 12.5 years (11 to 15) for females ($p < 0.001$, independent-samples t-test).

Patients who sustained a trimalleolar ankle fracture were identified in our hospital in Australia between June 2015 and January 2021. These patients were subjected to the same inclusion and exclusion criteria as the patients with triplane fractures except that age >18 years was used. The fractures were classified as trimalleolar fractures by an experienced fellowship-trained trauma surgeon (JR) based on the preoperative CT scan.

The inclusion criteria for both datasets were availability of a preoperative CT scan with slices of 1 mm thickness or less, and presence of a transitional distal tibial/trimalleolar fracture. Exclusion criteria were: presence of an old or pathological fracture; CT scan unavailable or only with slices thicker than 1 mm; and presence of other disorders such as, but not limited to, osteomyelitis, joint infections, or other fractures (excluding fibula fractures).

Imaging data of 100 patients with a trimalleolar fracture were identified. A total of 29 (29%) were male and 71 (71%) were female. The mean age was 53 years (18 to 82).

Fracture classification

Patient characteristics such as age at injury and sex were recorded. CT scans were exported from the picture archiving and communication system (Vue; Carestream, USA) as DICOM files. All CTs were assessed, and fractures were classified using the axial, coronal, sagittal, and 3D CT reconstructions in Horos v. 3.3.6.

Fracture mapping

In line with our previous study's fracture-mapping methodology⁸, we digitally captured a series of paediatric triplane fractures as well as adult trimalleolar fractures. Standardized

axial views were created, located 3 mm above the distal tibial subchondral surface, a method described by Cole et al.¹⁰ in their work on articular mapping of adult Pilon fractures. As the reference template, a healthy right adult ankle was chosen. Recorded CT slices were adjusted as necessary, which included mirroring, resizing, rotating, and normalizing them to match the template tibia's dimensions using Photoshop 2023 v. 25.1.0 (Adobe, USA). This adjustment process began with aligning specific tibial landmarks, starting with the intact posteromedial tibia. For fractures with multiple displaced fragments, digital reduction was performed. Finally, the fracture lines on each recorded axial slice were manually traced using a built-in pencil brush with a four-pixel size and superimposed onto the template.

Statistical analysis

Patient characteristics were reported as means and ranges, or numbers and percentages. Comparison of means between two groups was performed by using the independent-samples t-test, and if data were not normally distributed the Mann-Whitney U test was used. A p-value <0.05 was considered statistically significant. The primary study goal was to assess the association between the paediatric triplane and adult trimalleolar fractures. To achieve this, fracture maps were visually assessed, compared with each other, and related step by step to the progression of injury according to the Lauge-Hansen classification. The fracture maps were assessed by a panel of five fellowship-trained surgeons, until consensus was reached regarding the description of the fracture pattern for each stage in the Lauge-Hansen classification. The surgeons individually received the fracture maps and were asked for each stage of the Lauge-Hansen classification about the entry and exit points of the fracture lines, and the size of the fragment, and to compare these between paediatric and adult fractures. The results were collected and then discussed in a consensus meeting.

RESULTS

Trimalleolar and triplane Lauge-Hansen sequence of injury

Supination external rotation injuries in adults following the progressions as described by Lauge-Hansen start with Stage I, the rupture of the anterior syndesmosis or small avulsion of the anterolateral tibia as depicted on the fracture map (Figure 1).

In paediatric patients, the external rotation force places stress on the anterior syndesmosis, causing a fracture of the distal tibial epiphysal plate anterolaterally, known as Tillaux fracture (Figure 2).

On the fracture map, the fracture line originates ventrally on the tibia, progresses towards the central area, and then extends towards the tibiofibular joint (Figure 1, 3, and 4).



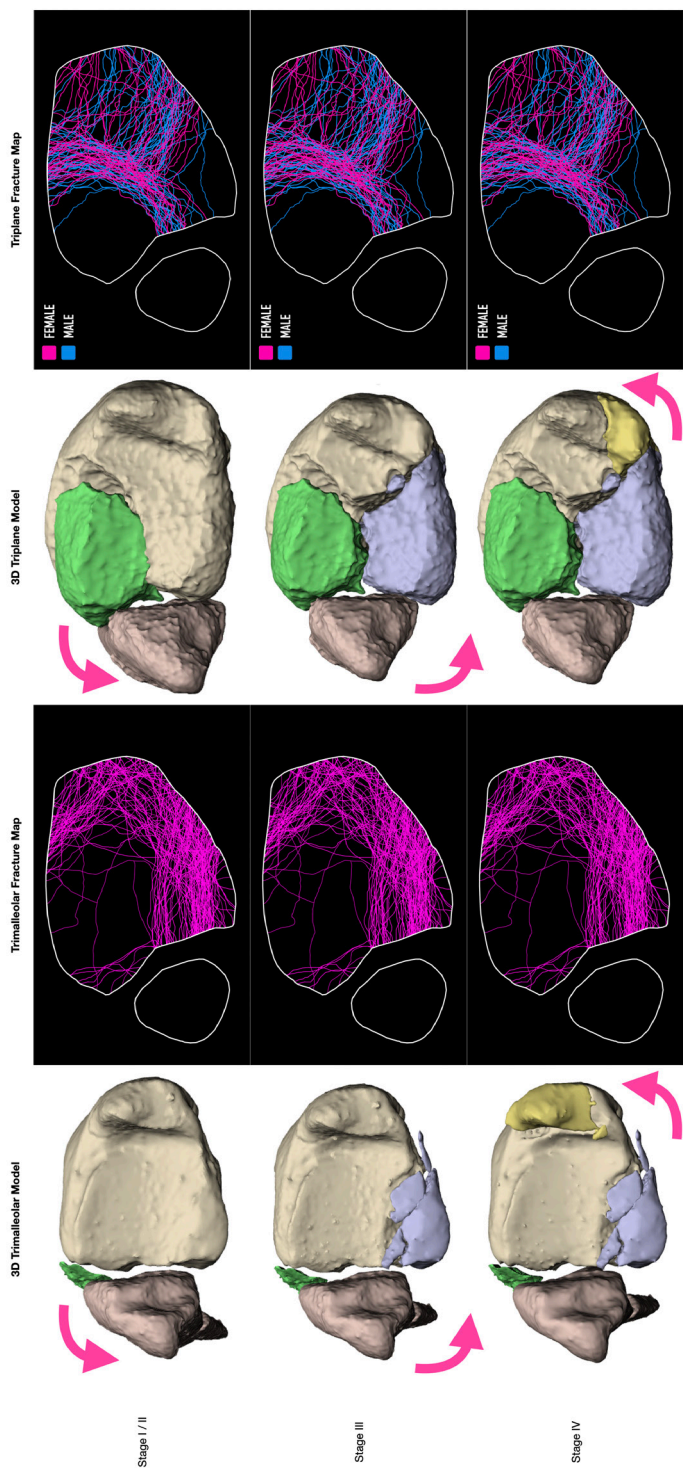


Fig. 1 The four stages of trimalleolar and triplane fractures according to the Lauge-Hansen classification.

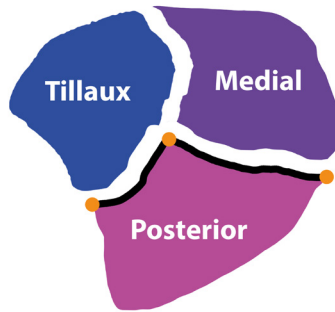


Fig. 2 Wave-like configuration of the posterior fragment in triplane fractures. The wave-like pattern of the posterior malleolar fracture line is part of the overall Y-pattern that is characteristic of triplane fractures.

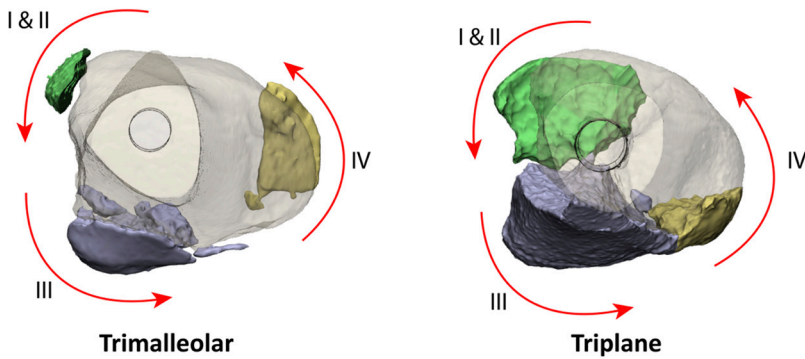


Fig. 3 Sequence of injuries in adult trimalleolar as compared to paediatric triplane fractures.

Stage II of the trauma mechanism is defined as oblique fracturing of the fibula at the level of the syndesmosis (OTA/AO type 44B)¹⁶ (Figure 1 and Figure 4). This was present in all trimalleolar fractures, and in 30% (25/83) of triplane fractures. Fibula fractures were recorded, but not mapped, since the fibula fracture lines were proximal to the level of the fracture maps, and may pass variably through the distal fibular growth plate.

In adults, stage III represents an external rotation force advanced to the point where either the posterior syndesmosis is ruptured or a fracture of the posterior malleolus occurs (Figure 3). On the trimalleolar fracture map, many fracture lines signifying a posterior malleolar fracture can be observed (Figure 1). If these lines are closely analyzed, the three different Haraguchi types can be discerned: Type I, characterized by one large isolated fragment; Type II, consisting of two fragments that overlap and of which one extends anteromedially; and Type III, which consists of small shell-like fractures.¹⁷

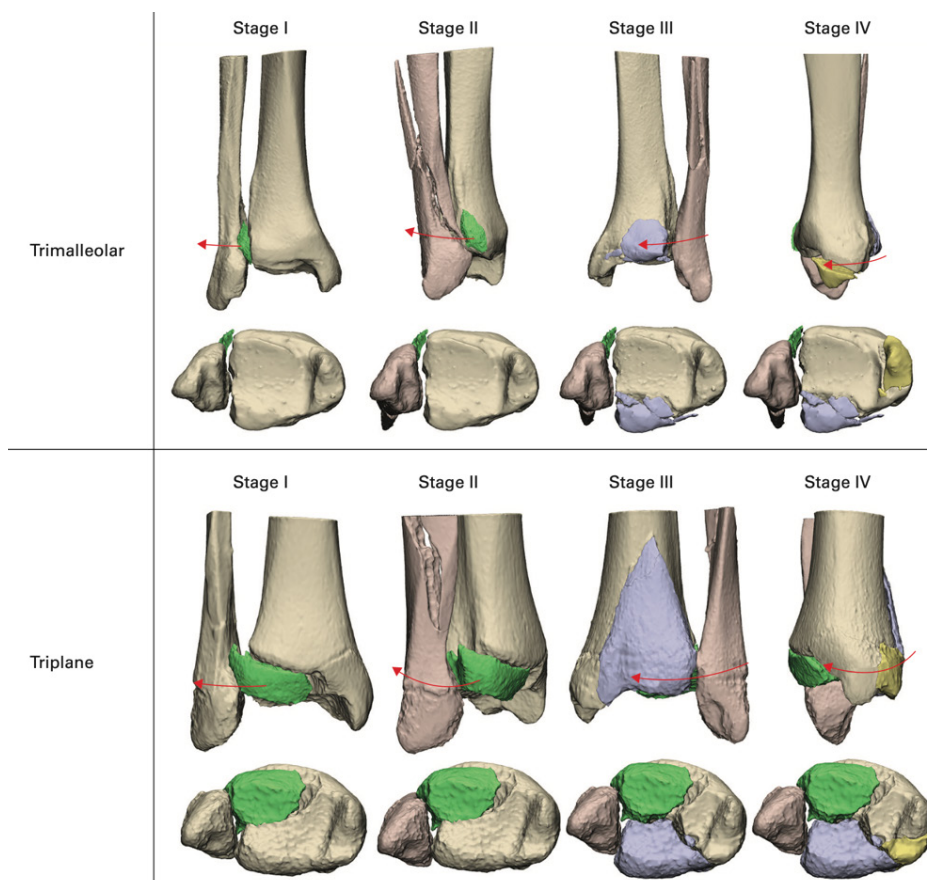


Fig. 4 3D representation of the sequence of injuries in trimalleolar versus triplane fractures.

When appreciating the comparative fracture lines on the triplane fracture maps, the configuration is slightly different from trimalleolar fractures in adults, as a large anterolateral fragment (Tillaux) is already avulsed during stage I in these paediatric patients (Figure 1). The fracture line of the Tillaux fragment, extending from the tibiofibular joint to the central region of the tibia, marks the onset of the posterior malleolus fracture. At stage III of the triplane fracture, this fracture line is extended posteromedially. This results in a wave-like pattern of the fracture line, initiated at the posterolateral end of the Tillaux fragment and exiting posteromedially (Figure 2 and Figure 4). This wave-like pattern of the posterior malleolar fracture line is part of the characteristic Y-pattern of triplane fractures.⁸

Finally, stage IV presents full external rotation and fracture of the medial malleolus or rupture of the medial deltoid ligament (Figure 3). In the trimalleolar fracture map, this is evident by fracturing of the medial malleolus in two ways. One way the fracture lines exit the medial malleolus on the fracture map is as part of the Haraguchi Type 2, where the fracture

line extends medially and splits the medial malleolus in the sagittal plane. The second way consists of fracture lines that run from anteromedially to posteromedially, which split the medial malleolus in the coronal plane instead of the sagittal plane, and exit in a similar location to the aforementioned fracture lines.

The medial malleolar fractures as part of trimalleolar fractures are more variable in fracture pattern than those that are part of triplane fractures (Figure 1). Stage IV of triplane fractures consists of a fracture line that mostly originates at the posterior fragment, extends medially, and splits the medial malleolus in the sagittal plane (Figure 4).

DISCUSSION

Recently, we questioned the role of the physis as the main determiner of triplane fractures, and instead showed that a partially closed physis is the requisite while ligaments determine the fracture pattern.⁸ We found that both the paediatric triplane and adult trimalleolar fractures follow the stages as described by Lauge-Hansen, highlighting that the distinctive fracture pattern in adolescents arises from a combination of ligament traction and a growth plate in the process of closing. With this new knowledge, together with existing literature, optimal screw positions and sequence of fixation can be recommended.

This study has some strengths and some limitations. Important strengths of this study are the inclusion of patients with thin-slice CT scans, the use of the largest cohort of triplane fractures in the literature, and the assessment of the fractures by several experienced surgeons including a fellowship-trained paediatric orthopaedic surgeon. Limitations include the retrospective design of this study, making it difficult to verify the trauma mechanism of these patients. Another limitation is the lack of MRI scans as a standard of care for enrolled patients, preventing us from directly assessing ligamentous injuries through imaging.

Hadad et al.⁹ were the first to employ the fracture map technique with 33 triplane fractures, and described the most common fracture patterns of triplane fractures. Our paper builds on their findings by assessing 83 triplane fractures and comparing these to fracture patterns of trimalleolar fractures. We explored the involvement of ligaments in triplane injuries, and challenged the commonly taught theory that physis is the main determinant of fracture pattern in triplane fractures. While acknowledging the weakness associated with the physis, we propose that the role of ligaments exceeds that of the physis in these injuries.⁸ Although this concept might offer an innovative perspective on fractures in adolescents, the literature on fractures in adults already indicates that ligaments play a crucial role in intra-articular fractures, contributing to the development of specific fracture patterns in proximal humerus and distal radius fractures.^{18,19} However, it is still unclear if that is an 'active' role pulling off the fragments, or 'passive' in protecting the bone where it attaches to the bone.



If ligaments play a crucial role in triplane injuries, the pattern for supination external rotation injuries should be reasonably predictable and consistent, independent of the presence of a partially open physis. When compared with the adult trimalleolar fractures that result from the same trauma mechanism, we found similarities across all stages of the Lauge-Hansen classification.⁷

Stage I represents a rupture of the anterior syndesmosis or avulsion of the anterolateral tibia. As ligamentous ruptures cannot be reliably detected on CT scans, these were not captured in the fracture maps. For both fractures avulsion of the anterolateral tibia (the Tillaux fragment) can be appreciated. The fragments of the triplane fractures are larger in size than in the trimalleolar fractures, most likely caused by the weakness created by the physis where the balance between the opposing forces (posterior syndesmosis, and medial fusion or deltoid ligament) is closer to the centre of the tibia.

Stage II shows similarity between the two injuries, as all trimalleolar fractures have an oblique fracture of the fibula, and a substantial number of the triplane injuries have a similar fracture pattern. However, the fibula is more likely to remain intact in paediatric patients compared to adults for a few reasons: 1) we observed that most paediatric patients have a certain degree of plastic deformation in the fibula without fracturing it; and 2) in paediatric patients, a lot of the energy is already absorbed by the formation of a large Tillaux fragment. This supports the notion that trimalleolar as well as triplane fractures follow a supination external rotation pattern.

Stage III results in fracturing of the posterior malleolus in both adult trimalleolar and adolescent triplane fractures, albeit with a slightly different configuration. The trimalleolar posterior fragment has a concave fracture line, which starts between the middle and posterior thirds of the tibia and exits posteromedially in the middle of the tibia. For the triplane, weakness created by the physis allows fracturing of a larger anterolateral Tillaux fragment, as there is less inherent strength of the bone in the epiphysis.^{2,20} Hence, fragment size is determined by the counter forces of the opposing ligaments, such as the posterior-inferior tibiofibular and the medial deltoid ligaments.

Stage IV is the final step of the sequence of injuries, where full external rotation has occurred, with fracturing of the medial malleolus or rupture of the deltoid ligament. The attachment site and fibre orientation of the deltoid ligament probably have a considerable influence on the fracture pattern in the medial malleolus.⁸ In general, the fracture pattern of the medial malleolus, as part of trimalleolar fractures, demonstrated greater variability compared to those occurring within triplane fractures. Our findings are in line with the extensive fracture maps provided by Liu et al.²¹ who demonstrated great variation in fractures of the medial malleolus in adults. Moreover, it is likely that triplane fractures do not solely result from rotational forces but also involve an adduction component, which could contribute to the variability in the fracture pattern of the medial malleolus.^{22,23} However, some identical fracture lines can be appreciated in trimalleolar and triplane fractures. In paediatric triplane fractures,

medial malleolar fractures usually arise as a medial extension of the posterior fragment. In the adult trimalleolar fractures, this extension is also observed in the Haraguchi type 2 fractures (i.e. medial-extension type).

The similarities between triplane and trimalleolar fractures found in the current study reflect the importance of injury mechanism and ligamentous forces in these injuries. The insights gained from our study regarding the mechanism and sequence of injuries may help to understand the manoeuvres which would enable reduction of these fractures. We propose that minimally displaced triplane fractures should undergo anterior directed force to reduce the medial and posterior fragment (i.e. reversing stages 3 and 4), followed by reduction of the anterior fragment using internal rotating forces (i.e. reversing stages I and II). Subsequently, a cast should be applied to preserve internal rotation and prevent secondary displacement.^{24,25} Further, the findings assist in understanding appropriate screw positions for surgical treatment. Hadad et al.⁹ described different screw locations for fixing triplane fractures based on fracture mapping. In addition, Sheth et al.²⁶ performed a similar study, albeit without a fracture map, where they suggest screw angles based on axial CT slices. These studies suggest several screw trajectories, but this is not substantiated by a detailed analysis of the sequence of injuries around the ankle. Moreover, those studies do not suggest in what order to stabilize these injuries. We recommend reversing the sequence of injuries, first reducing with anterior directed force and fixing the posterior (medial) fragment using an anteroposterior screw (i.e. reversing stages III and IV), usually leaving the fibula fracture (stage 2), and then reducing with internal rotation and fixing the Tillaux fragment with an anterolateral to anteromedial screw (or vice versa, i.e. reversing stage I) (Figure 5).

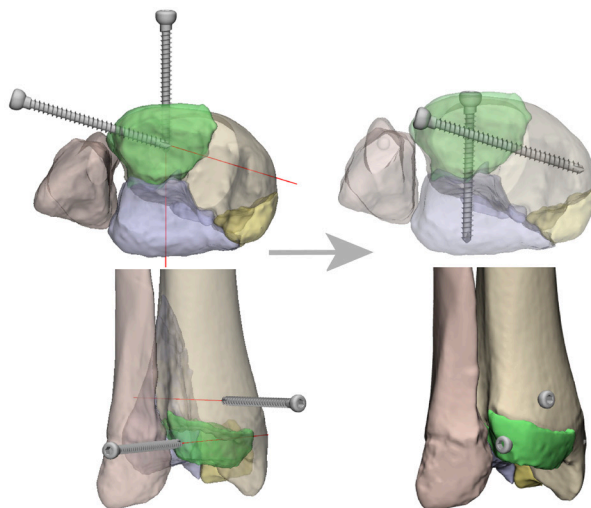


Fig. 5 Recommended screw placement for treatment of triplane fractures reversing the sequence of injuries.

This study shows major similarities in supination external rotation injuries, as described by Lauge-Hansen⁷, between adults with trimalleolar fractures and adolescents with triplane fractures at the level of the joint. Based on the injury mechanism, we identified a sequence of injury consisting of four stages in triplane fractures. The characteristic fracture pattern results from a combination of pulling and stabilizing forces of ligaments and weakness due to a partially open growth plate.



PART III

Prospective Analysis of Ankle Fractures

The Twelve Apostels, Great Ocean Road, Victoria, Australia



**Subtle Factors Associated with Outcome
of Ankle Fractures: Do Not Forget
Intra-Articular Loose Bodies**

J. Prijs

D.T.M. Meijer

R.P. Blom

R-J de Muinck Keizer

T. Schepers

I. Sierevelt

J.C. Goslings

F.F.A. IJpma

J.N. Doornberg

R.L. Jaarsma

EF3X-trial Study Group

ABSTRACT

Background

Despite the ubiquity of ankle fractures in daily practice and the following surgical experience and comfort with these injuries, outcomes do not match the expectations. It is still unclear why, although some studies revealed Haraguchi classification and fracture severity to be important. In this study we explored the effect of 'subtle' factors on outcome, for example loose bodies and malpositioned hardware.

Methods

This study performed a retrospective review of 102 prospectively collected anonymous postoperative Computed Tomography (CT) scans of adult patients with ankle fractures from the randomized clinical EF3X-trial. All included patients completed the Foot and Ankle Outcome Score (FAOS) at two years post-operatively. Subtle factors hypothesised to affect outcome were: 1) loose bodies in the ankle joint or syndesmosis; 2) (marginal) medial and/or lateral impaction; 3) posterior malleolar edge impaction and/or comminution with incarcerated fragments; 4) talar osteochondral defects; 5) quality of reduction of tibia, fibula, and syndesmosis, and 6) malpositioned hardware. These were manually recorded and a majority consensus of three independent observers was used. These factors were used to evaluate if there was an association with the FAOS domains.

Results

There are no significant differences for all FAOS domains between AO/OTA types A (n=6), B (n=74) and C (n=22). Univariate analysis for the 'subtle' factors revealed that Haraguchi classification, sex, loose bodies, posterior malleolar incarcerated fragments and malpositioned hardware significantly affected two or more FAOS domains. Specifically, loose bodies resulted in significantly lower outcomes in: Pain, Activities of Daily Living and Quality of Life. Multivariate analysis with adjusting for confounders revealed 'Haraguchi classification' and 'loose bodies' to be significant contributing factors.

Conclusion

In AO type 44 ankle fractures, loose bodies in the joint or syndesmosis, significantly affect patient reported Pain and Quality of Life at two years postoperatively. During surgery, effort should be made to identify and remove loose bodies.

Clinical relevance of the paper / Take Home Message

- During surgery, effort should be made to identify and remove intra-articular loose bodies using the pre-operative CT scan.

INTRODUCTION

Ankle fractures are one of the most common fractures with an incidence of 179 per 100.000 person-years¹. Most of these fractures are low-energy rotational types as described by Lauge-Hansen in 1950². A small percentage are high-energy axial tibial pilon fractures with extensive soft-tissue damage—which can be considered a different type of injury from what is commonly considered an ankle fracture, and often requires a multi-disciplinary approach³⁻⁶. In contrast to the high volume of rotational type ankle fractures, and thus surgeons with a lot of experience and comfort treating these injuries, outcomes following operative fixation are still suboptimal. For example, common short-term and long-term consequences such as reduced mobility, chronic pain, and early posttraumatic osteoarthritis⁷⁻⁹ are often reported. Large cohorts are scarce, but multiple smaller studies report 21-64% poor or fair outcome¹⁰⁻¹⁴. The landmark study by Egol et al.⁷, reports up to 37% of patients suffering from a form of limitation one year postoperatively after an AO/OTA type 44¹⁵ ankle fracture.

It is still unclear why such a common fracture, in which surgeons have a lot of experience and confidence in their abilities to optimally treat these injuries, has relatively poor outcome. There have been a few studies reporting specific characteristics that are associated with poor outcome, such as posterior malleolar involvement, particularly Haraguchi type 2 fractures extending medially¹⁶. Another long term outcome study¹⁷ reports fracture severity, trimalleolar involvement and Body Mass Index (BMI) to be significant factors associated with poorer clinical outcome. The authors, however, do not report any correlation between outcomes, fracture morphology and articular involvement as quantified with postoperative computed tomography (CT). Besides pathophysiological theories, two studies^{9, 18} advocate for the role of mental health as one of the main factors associated with long term recovery.

Elucidating prognostic factors that affect outcome is crucial to adequately inform patients following their trauma and improve surgical decision making. So far, common factors like age, fracture severity and morphology have not been able to explain the variability in outcomes and treatment of these common fractures. Hence, treatment is often based on the experience of the treating surgeon where one would expect more evidence-based (shared) decision making in this era. Perhaps 'subtle' surgical and pathoanatomic factors such as quality of postoperative reduction¹⁶, loose bodies or malpositioned hardware can predict outcomes more accurately following rotational type ankle fractures.

Therefore, we aimed to retrospectively evaluate factors associated with functional outcome in a prospective consecutive cohort of patients with rotation type ankle fractures with pre- and postoperative CT scans. The current study aimed to answer the following questions: 1) What are Foot and Ankle Outcome Scores (FAOS) for simple ankle fractures two years postoperatively; and 2) What are (subtle) factors associated with functional outcome in these injuries, including advanced imaging analysis of pre- and postoperative CT scans?



METHODS

In accordance with the Declaration of Helsinki¹⁹, our Institutional Review Board (IRB) approved a retrospective review of prospectively collected anonymous Computed Tomography (CT) images of adult patients with ankle fractures from the randomized clinical EF3X-trial²⁰, (Dutch Trial Register NTR 1902). The primary aim of the EF3X-trial was to evaluate the effect of intraoperative use of 3D-fluoroscopy compared to 2D-fluoroscopy alone on the quality of fracture reduction, in operatively treated calcaneal, wrist and ankle fractures.²⁰ Our group previously reported on quantification of three-dimensional CT (Q3DCT) in posterior malleolar ankle fractures by secondary retrospective analysis of this prospective cohort of consecutive patients.^{16, 21, 22}

Similar methodology, as described earlier in our previous study that identified factors of outcome of posterior malleolar fractures, was used.¹⁶ In short, in the current study all patients with an ankle fracture who were prospectively included in the EF3X-trial, were retrospectively included for secondary analysis.²⁰ All included patients completed the Foot and Ankle Outcome Score (FAOS)²³ at two years postoperatively. As part of the EF3X-trial, all patients underwent postoperative CT scanning (Somatom Definition AS+; Siemens, Erlangen, Germany) within one week after surgery and resulting in thin-slice (<1 mm) reconstructions of the injured leg. These were saved as anonymous Digital Imaging and Communications in Medicine (DICOM) files. Post-operative CT scans were used as reference standard in the EF3X-trial²⁰ and the secondary analysis of EF3X study data with advanced imaging analysis was used for the current study.

Research Questions

For the primary research question the cohort of malleolar ankle fractures (i.e., AO/OTA type 44¹⁵) was analysed per fracture type (A, B or C) by using the respective domain scores of the FAOS (i.e., Symptoms, Pain, Activities of Daily Living (ADL), Sports and Quality of Life (QoL)) at two years postoperative. For the second research question, the potential 'subtle' factors associated with outcome in rotational type ankle fractures were correlated to the same FAOS domain scores.

Patients

From the EF3X-trial total of 176 patients²⁰, the current study retrospectively included a total of 102 patients with a 'simple' rotational type ankle fracture (AO/OTA Fracture Classifications¹⁵; 44-A, -B and -C) with a completed FAOS at two years postoperative. The 74 excluded patients all had a Pilon tibial fracture; AO/OTA type 43¹⁵. Patients were all treated according to the conventional AO principles of open reduction and internal fixation according to surgeons' discretion. Mean age at surgery was 46.5 ($\pm$ 14.7) years, and of all patients: 46 (45%) were male and in 51 (50%) patients the right ankle was injured. A total of 23 (23%) patients smoked during the period of inclusion. The trauma mechanism ranged from: low energetic trauma (LET) from fall (n=76), high energetic trauma (HET) from fall (n=3), HET motor vehicle accident (n=6), and 'other' or unknown (n=17). Baseline characteristics are reported in Table I.

Table I. Baseline demographics and fracture characteristics

Characteristics	Total
Sex. n (%)	
Male	(45)
Female	(55)
Side of ankle fracture. n (%)	
Left	(50)
Right	(50)
Mean age at surgery. yrs (SD)	46.5 (14.7)
Smoking. n (%)	
Yes	(22)
No	(68)
Unknown	(10)
Trauma Mechanism. n (%)	
LET - Fall	(75)
HET - Fall from height	3 (3)
HET - MVA	6 (6)
Other or unknown	(16)
AO/OTA Fracture Classification. n (%)	
A	6 (6)
B	(72)
C	(22)
Revision Surgery within 6 weeks of operation. n (%)	5 (5)
Non-satisfactory post-operative result	3
Infection	2
Revision Surgery within 12 weeks of operation. n (%)	5 (5)
Removal of Syndesmotic Screw(s)	3
Removal of Gentamycine beads	1
Infection	1
Revision Surgery within one year of operation. n (%)	(18)
Complaints of osteosynthesis material	12
Removal of Syndesmotic Screw(s)	4
Infection	1
Malunion corrective osteotomy	1



Table I. Baseline demographics and fracture characteristics (continued)

Characteristics	Total
Foot and Ankle Outcome Scores, Means (0-100), Standard Deviations	
Symptoms	(±16)
Pain	(±24)
Sports and Recreation	(±29)
Activities of Daily Living	(±22)
Quality of Life	(±25)
Subtle Predictors. n (%)	
Loose body in joint	(22)
Loose body in syndesmosis	(40)
Medial tibial impaction	(14)
Lateral tibial impaction	1 (1)
Posterior malleolar edge impaction	(36)
Posterior malleolar comminuted with incarcerated fragment(s)	(37)
Talar osteochondral defect	1 (1)
Quality of reduction fibula. n (%)	
Adequate	(88)
Poor	(12)
Quality of reduction tibia. n (%)	
Adequate	(88)
Poor	7 (7)
N/A	5 (5)
Quality of reduction syndesmosis. n (%)	
Adequate	(75)
Poor	(20)
N/A	6 (6)
Malpositioned hardware. n (%)	
Screw in ankle joint	4 (4)
Screw in syndesmosis	(15)
Fibula screw grinding tibia	(24)
Screw through fibula and tibia	3 (3)
Tibia plate overlapping fibula	1 (1)
Protruding screw extra-articular	6 (6)

Within the first six weeks post-operatively, five patients (5%) had revision surgery for non-satisfactory postoperative result (n=3), and infection (n=2). Between 6 and 12 weeks, in three patients the syndesmotic screws were removed, and between 12 weeks and one year post-operatively, 18 patients (18%) had subsequent surgery, of which the majority (n=12) was elective implant removal (Table I).

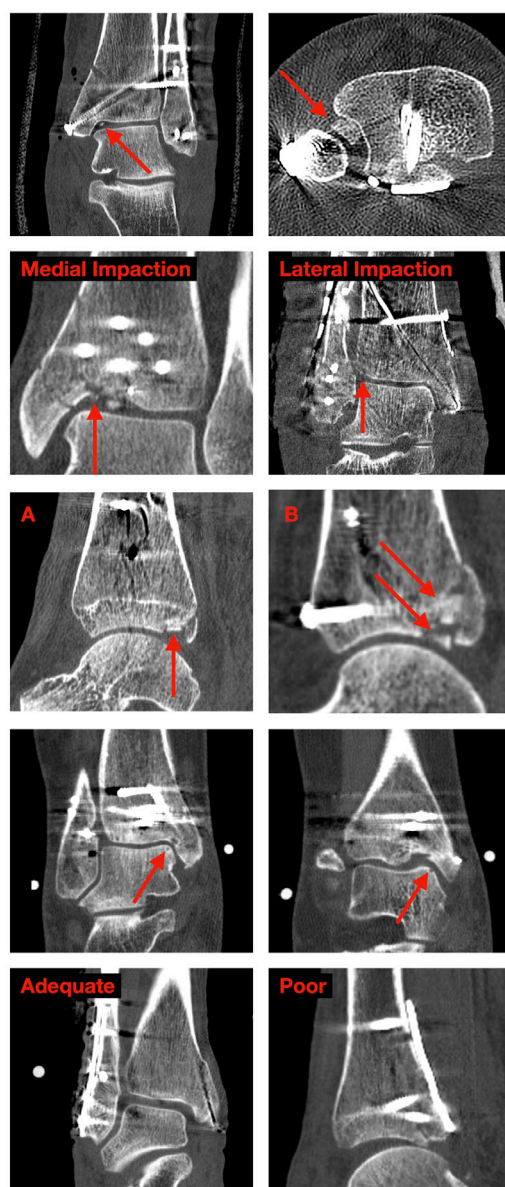


Figure 1. Examples of potential subtle factors that may affect outcomes

Factors Associated with Clinical Outcome

Potential 'subtle' factors associated with clinical outcome analysed in this study were: 1) loose bodies in the ankle joint or syndesmosis (Supplemental Figure 1); 2) (marginal) medial and/or lateral impaction; 3) posterior malleolar edge impaction and/or comminution with incarcerated fragments; 4) talar osteochondral defects; 5) quality of reduction of tibia, fibula, and syndesmosis, and 6) malpositioned hardware. The majority consensus of three independent observers (DM, RB, JP), who individually evaluated the presence of the above-mentioned factors on the included postoperative CT scans with the use of standardized axial, sagittal and coronal reconstructions, was used.

Definitions of the above-mentioned factors were as follows (Figure 1): 1) loose bodies were defined as any visible fragments—from small specks of bone to larger fragments—of bone in the ankle joint or syndesmosis on the postoperative CT in axial, coronal or sagittal planes; 2) (marginal) impaction was defined as a depression ($>2\text{mm}$) of the articular surface in the medial or lateral tibial plafond as observer on the sagittal plane in the middle one third of the tibia; 3) posterior malleolar edge impaction was defined as any impaction of the posterior malleolus as observer in the sagittal plane in the posterior 1/3 of the tibial plafond, whereas comminution with incarcerated fragments was scored regardless of impaction; 4) talar osteochondral defects were defined as a defect of any detectable size of the subchondral bone on the CT scan; 5) adequate quality of reduction of the tibia was defined as: 5a) an adequate articular reduction meaning a congruent ankle joint and intra-articular fractures with a step-off and gap of less than 2 millimetres each in the coronal, sagittal and axial planes; 5b) for the fibula this was defined as having the correct length as measured of the distal talar facet; 5c) whereas for the syndesmosis this was defined when the lateral border of the fibula was in line with the anterolateral border of the tibia; 6) Malpositioned hardware (Figure 2) was defined as: a) screws protruding into the joint or syndesmosis; b) fibular lag screw heads grinding anteriorly against the tibia (Figure 2); c) incidental screws into other bones—i.e. fibula screws that protrude into the tibia—but not protruding into the joint (excluding syndesmotic fixation screws); d) overlapping plates—i.e. a posterolateral tibia plate that also sits on the posterior fibula—or screws protruding into soft tissues.

Due to clinical overlap of syndesmotic and intra-articular loose bodies, these combined as 'loose bodies'. The specific malpositioned hardware subtypes were combined into one variable—'Malpositioned Hardware'—due to low case numbers individually.

Statistical analysis

IBM SPSS software for Macintosh (version 27; IBM Corp., Armonk, NY, USA) was used for the statistical analysis. Baseline characteristics were reported as frequencies and percentages in case of categorical variables, and as means and standard deviations in case of continuous variables. Variables with less than a total of five cases were excluded from analysis. The one-way ANOVA was used to determine statistically significant differences between AO/OTA

types A, B and C for each respective FAOS domain score. First, to identify the association between the potentially clinically important factors and the FAOS domain scores at two years follow-up, the independent T-test was used for univariate analysis. A p-value of <0.05 was used for statistical significance. Second, a multivariate analysis was performed to adjust for possible confounders (age, sex, AO/OTA fracture type, Haraguchi fracture type and smoking) and determine the effect and significance of the potentially subtle factors that determine outcome. For the multivariate, factors that had a significance level of $p < 0.1$ were included in the model, and with a backward stepwise selection procedure the significant ($p < 0.05$) subtle factors were identified.



Figure 2. Examples of Mal-Positioned Hardware

RESULTS

Simple (AO 44) Ankle Fractures Outcomes

The FAOS domains symptoms, sports and recreation, and quality of life have mean scores between 60 and 70 at two years postoperative. In addition, the FAOS domains pain and activities of daily living report higher mean scores, 78 and 84, respectively (Table I).

There were no significant differences between AO/OTA fracture types 44 A (n=6), B (n=74) and C (n=22) for all FAOS domains; Symptoms ($p=0.275$), Pain ($p=0.129$), Sports ($p=0.141$), Activities of Daily Living ($p=0.270$) and Quality of Life ($p=0.255$)(Figure 3). Comparison of each respective FAOS domain between those with a posterior malleolar fragment (PMF, n=79) and those without (n=23) also did not reveal statistically significant differences (Table II). However, there is a significant difference between outcomes of the Haraguchi types on the 2-year postoperative outcomes for all five domains (Table II), where Haraguchi type 2 has the lowest mean FAOS domain scores.

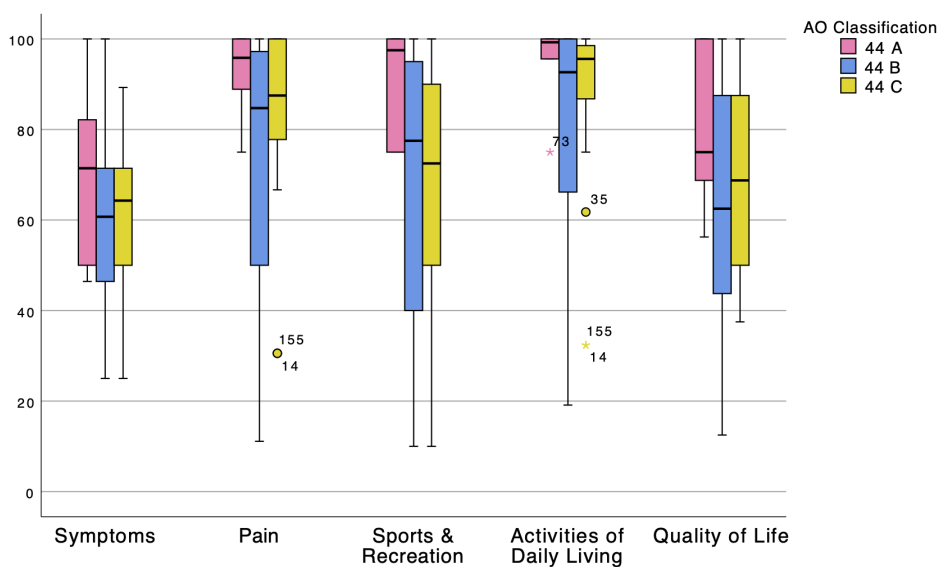


Figure 3. FAOS domains Boxplot clustered by AO classification

Factors Associated with Outcome in Simple Ankle Fractures

The factors 'Lateral Tibial Impaction', 'Talar Osteochondral Defect' were excluded from analysis due to low number of cases (n<5).

The presence of potential subtle factors on postoperative CT scans are reported in Table I. Univariate analysis for the subtle factors, showed that Haraguchi classification, sex, loose bodies, posterior malleolar incarcerated fragments on the postoperative CT, and malpositioned hardware significantly affect two or more FAOS domains (Table II).

Table II. Univariate analysis of confounders and subtle predictors for FAOS at two years post-operative in simple ankle fractures

Confounders	FAOS symptoms	FAOS pain	FAOS ADL	FAOS sports	FAOS QoL
Posterior Malleolar Fragment	.170	.201	.301	.442	.862
Age	.981	.716	.245	.790	.952
Sex	.002*	.031*	.009*	.070	.078
AO Classification	.275	.129	.270	.141	.255
Smoking	.753	.669	.677	.935	.966
Haraguchi Classification	.002	<0.001	<0.001	.005	.017
Subtle Predictors	FAOS symptoms	FAOS pain	FAOS ADL	FAOS sports	FAOS QoL
Loose bodies	.306	.001*	.004*	.053	.002*
Medial impaction	.251	.244	.109	.677	.050*
Posterior malleolar edge impaction	.135	.293	.367	.902	.574
Posterior malleolar incarcerated fragments	.008*	.005*	.014*	.140	.041*
Quality of reduction Fibula	.212	.869	.466	.137	.890
Quality of reduction Tibia	.678	.579	.941	.116	.525
Quality of reduction Syndesmosis	.405	.583	.483	.614	.882
Mal-positioned Hardware	.020*	.068	.047	.052	.059

Specifically, presence of loose bodies (Figure 4) results in significantly lower outcomes in three of the five FAOS domains (Table II); Pain 69 versus 85, ADL 77 versus 89, and QoL 57 versus 72.

The majority of patients with loose bodies were AO¹⁵ type 44B3 66% (31/47) compared to 16% (10/55) in patients without loose bodies, where AO¹⁵ type B1 36% (20/55) formed the majority. When significant (with adjusted p<0.1) factors were concatenated in a multivariate analysis, 'Haraguchi classification', 'loose bodies', 'posterior malleolar comminution with incarcerated fragments' and 'sex' were significant factors (Table III).



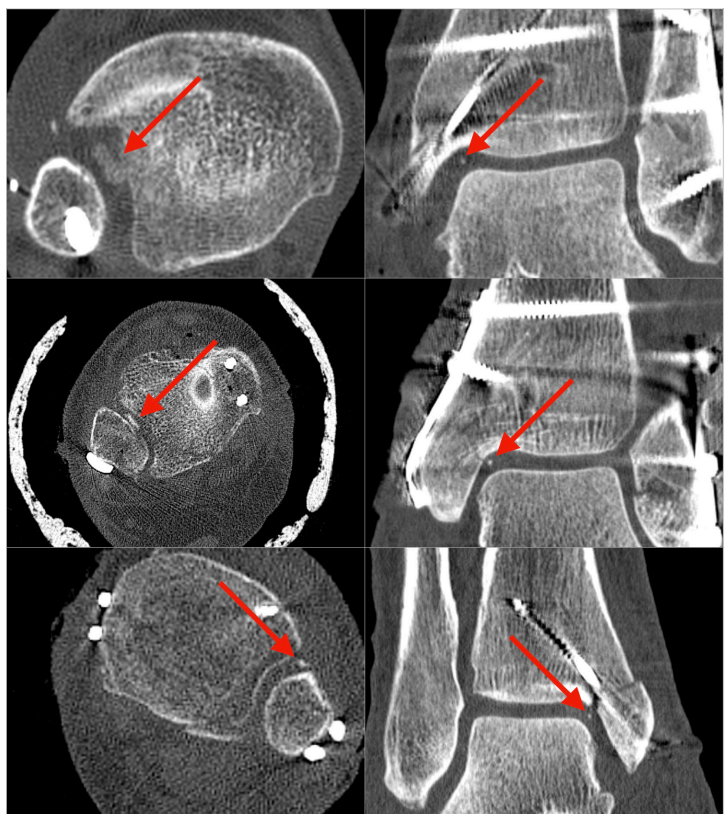


Figure 4. Examples of Loose Bodies in Syndesmosis (left column) and Joint (right column)

Table III. Multivariate analysis of subtle predictors for FAOS at two years post-operative in posterior malleolar ankle fractures

Subtle Predictors	FAOS symptoms	FAOS pain	FAOS ADL	FAOS sports	FAOS QoL
Sex	.006	-	-	-	-
Posterior Malleolar Comminuted with Incarcerated Fragment	.024	-	-	-	-
Loose bodies	-	.025	-	-	.002
Haraguchi					
Type I	-	.166	.113	.323	-
Type II	-	<.001	<.001	.002	-
Type III	-	.867	.549	.094	-

'Haraguchi type 2' and 'Loose bodies' are most significantly associated with outcome in these patients. Malpositioned hardware, age and AO fracture type classification, did not affect FAOS scores at two years postoperatively.

DISCUSSION

This study found loose bodies—that were defined as any visible loose bone, ranging from small specks to larger fragments—in the ankle joint and syndesmosis to be significant factors, even when adjusting for confounders such as, Haraguchi classification, age, AO classification, gender and smoking. The presence of loose bodies (either in the joint or in the syndesmosis; Figure 4) resulted in significantly lower FAOS domain scores Pain ($p=0.025$) and Quality of Life ($p=0.002$).

The current study should be interpreted considering its strengths and limitations. Strengths include a unique high-quality subsequently prospectively collected data of simple ankle fractures with postoperative CT scans with long term follow-up up to two years postoperatively. Secondly, three independent observers evaluated all postoperative CT scans. Limitations include a relatively small number of patients introducing the possibility for a selection bias. Another limitation is that surgical treatment was not standardized during data collection, and instead relied on the individual preference of the treating surgeon. Finally, detection of loose bodies is limited to what can be appreciated on the CT scan and thus those consisting of just a cartilage component were not included in this study.

Building on the study findings of Blom et al.¹⁶, that reported Haraguchi type to be an important factor of outcome in posterior malleolar fractures, this study also explored other factors that could explain the relatively poor outcomes for common simple rotational type ankle fractures. Rajan et al.¹⁷ reported that regardless of Haraguchi type, trimalleolar fractures have worse outcomes compared to bimalleolar. However, in the current study Haraguchi type was a significant factor and neither AO/OTA fracture type classification¹⁵ nor presence of a posterior malleolar fragment, significantly affect outcomes. The current study reports an important effect of loose bodies on the two-year postoperative outcomes. Although this study corrected for the AO classification, it is unclear if 'loose bodies' have an active role in outcome or if they signify the extent of damage. However, 'loose bodies' playing an active role is supported by two previously published studies^{24,25} that advocate arthroscopy for that reason in ankle fractures. Smith et al.²⁵ reported improved outcomes following arthroscopic removal of loose bodies during operative fixation, especially in patients whom suffered an ankle dislocation injury. Kim et al.²⁴ advocate the use of arthroscopy during hardware removal when patients suffer from chronic pain and a problem such as a loose body or ankle impingement (bony or soft-tissue) has been objectified. In contrast to the beforementioned studies, Fuchs et al.²⁶ reported no improved outcome following arthroscopy in unstable ankle fracture treated with open reduction and internal fixation (ORIF). All three studies mention 10 to 15 minutes of



added operating time, and no added complications by employing arthroscopy during ORIF of these ankle fractures. Considering the minimal risk of using arthroscopy in these injuries, the reported effect of intra-articular loose bodies on outcomes, it may be a worthwhile technique to employ.

In conclusion, in 'simple' ankle fractures (AO/OTA fracture type 44) a loose body in the joint or syndesmosis negatively impacts patient reported FAOS domain scores Pain and Quality of Life at two years postoperatively. During surgery, effort should be made to identify and remove loose bodies by thoroughly washing the ankle joint or performing an arthroscopy.

REFERENCES

1. Juto H, Nilsson H, Morberg P. Epidemiology of Adult Ankle Fractures: 1756 cases identified in Norrbotten County during 2009-2013 and classified according to AO/OTA. *BMC Musculoskelet Disord*. 2018;19(1):441.
2. Lauge-Hansen N. Fractures of the ankle. II. Combined experimental-surgical and experimental-roentgenologic investigations. *Arch Surg* (1920). 1950;60(5):957-85.
3. Olszewski N, Tornetta P, 3rd. Open Reduction and Internal Fixation of a Partial Articular (OTA-43B) Pilon Fracture Through a Direct Anterior Approach. *J Orthop Trauma*. 2022;36(Suppl 3):S25-S6.
4. Stamatou NJ, Ostrowski TJ, Mori BV, Fiscella K, Anoushiravani AA, Rosenbaum A. Team Approach: Perioperative Management of Pilon Fractures. *JBJS Rev*. 2023;11(3).
5. Lou Z, Wang Z, Liu C, Tang X. Outcomes of tibial pilon fracture fixation based on four-column theory. *Injury*. 2023;54 Suppl 2:S36-S42.
6. Middleton SD, Guy P, Roffey DM, Broekhuysen HM, O'Brien PJ, Lefavre KA. Long-Term Trajectory of Recovery Following Pilon Fracture Fixation. *J Orthop Trauma*. 2022;36(6):e250-e4.
7. Egol KA, Teiwani NC, Walsh MG, Capla EL, Koval KJ. Predictors of short-term functional outcome following ankle fracture surgery. *J Bone Joint Surg Am*. 2006;88(5):974-9.
8. Sung KH, Kwon SS, Yun YH, Park MS, Lee KM, Nam M, et al. Short-Term Outcomes and Influencing Factors After Ankle Fracture Surgery. *J Foot Ankle Surg*. 2018;57(6):1096-100.
9. Meijer DT, Gevers Deynoot BDJ, Stufkens SA, Siersevelt IN, Goslings JC, Kerkhoffs G, et al. What Factors Are Associated With Outcomes Scores After Surgical Treatment Of Ankle Fractures With a Posterior Malleolar Fragment? *Clin Orthop Relat Res*. 2019;477(4):863-9.
10. Nilsson G, Jonsson K, Ekdahl C, Eneroth M. Outcome and quality of life after surgically treated ankle fractures in patients 65 years or older. *BMC Musculoskelet Disord*. 2007;8:127.
11. Ponzer S, Nasell H, Bergman B, Tornkvist H. Functional outcome and quality of life in patients with Type B ankle fractures: a two-year follow-up study. *J Orthop Trauma*. 1999;13(5):363-8.
12. Donken CC, Goorden AJ, Verhofstad MH, Edwards MJ, van Laarhoven CJ. The outcome at 20 years of conservatively treated 'isolated' posterior malleolar fractures of the ankle: a case series. *J Bone Joint Surg Br*. 2011;93(12):1621-5.
13. Stufkens SA, van den Bekerom MP, Kerkhoffs GM, Hintermann B, van Dijk CN. Long-term outcome after 1822 operatively treated ankle fractures: a systematic review of the literature. *Injury*. 2011;42(2):119-27.
14. Weening B, Bhandari M. Predictors of functional outcome following transsyndesmotom screw fixation of ankle fractures. *J Orthop Trauma*. 2005;19(2):102-8.
15. Meinberg EG, Agel J, Roberts CS, Karam MD, Kellam JF. Fracture and Dislocation Classification Compendium-2018. *J Orthop Trauma*. 2018;32 Suppl 1:S1-S170.
16. Blom RP, Hayat B, Al-Dirini RMA, Siersevelt I, Kerkhoffs G, Goslings JC, et al. Posterior malleolar ankle fractures. *Bone Joint J*. 2020;102-B(9):1229-41.
17. Rajan L, Eble S, Kim J, Kukadia S, Kumar P, Day J, et al. Risk Factors Associated With Worse Clinical Outcomes of Ankle Fractures Involving the Posterior Malleolus. *Foot Ankle Orthop*. 2023;8(1):24730114231154217.
18. Lambers KT, van den Bekerom MP, Doornberg JN, Stufkens SA, van Dijk CN, Kloen P. Long-term outcome of pronation-external rotation ankle fractures treated with syndesmotom screws only. *J Bone Joint Surg Am*. 2013;95(17):e1221-7.
19. World Medical A. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013;310(20):2191-4.



20. Beerekamp MS, Ubbink DT, Maas M, Luitse JS, Kloen P, Blokhuis TJ, et al. Fracture surgery of the extremities with the intra-operative use of 3D-RX: a randomized multicenter trial (EF3X-trial). *BMC Musculoskelet Disord*. 2011;12:151.
21. de Muinck Keizer RO, Meijer DT, van der Gronde BA, Teunis T, Stufkens SA, Kerkhoffs GM, et al. Articular Gap and Step-off Revisited: 3D Quantification of Operative Reduction for Posterior Malleolar Fragments. *J Orthop Trauma*. 2016;30(12):670-5.
22. Meijer DT, de Muinck Keizer RO, Stufkens SAS, Schepers T, Sierevelt IN, Kerkhoffs G, et al. Quantification of Postoperative Posterior Malleolar Fragment Reduction Using 3-Dimensional Computed Tomography (Q3DCT) Determines Outcome in a Prospective Pilot Study of Patients With Rotational Type Ankle Fractures. *J Orthop Trauma*. 2019;33(8):404-10.
23. Sierevelt IN, Beimers L, van Bergen CJA, Haverkamp D, Terwee CB, Kerkhoffs G. Validation of the Dutch language version of the Foot and Ankle Outcome Score. *Knee Surg Sports Traumatol Arthrosc*. 2015;23(8):2413-9.
24. Kim HN, Park YJ, Kim GL, Park YW. Arthroscopy combined with hardware removal for chronic pain after ankle fracture. *Knee Surg Sports Traumatol Arthrosc*. 2013;21(6):1427-33.
25. Smith KS, Drexelius K, Challa S, Moon DK, Metzl JA, Hunt KJ. Outcomes Following Ankle Fracture Fixation With or Without Ankle Arthroscopy. *Foot Ankle Orthop*. 2020;5(1):2473011420904046.
26. Fuchs DJ, Ho BS, LaBelle MW, Kelikian AS. Effect of Arthroscopic Evaluation of Acute Ankle Fractures on PROMIS Intermediate-Term Functional Outcomes. *Foot Ankle Int*. 2016;37(1):51-7.





General Discussion
English Summary
Nederlandse Samenvatting

GENERAL DISCUSSION

PART I EXPLORING CONVOLUTIONAL NEURAL NETWORKS IN ANKLE FRACTURES

An Increasing Number of Convolutional Neural Networks for Fracture Recognition and Classification in Orthopaedics: Are these Externally Validated and Ready for Clinical Application?

Chapter 2 showed that, despite the surge in studies that report great accuracy in detecting and classifying fractures on radiographs with Convolutional Neural Networks (CNNs), only a few (4/36) use external validation (EV) in evaluating performance. In addition, this review highlights that geographical (by location) EV is superior to temporal (separate from training set by time) EV. The importance of and need for EV is highlighted by many studies; however, this review shows that EV of CNNs remains scarce. In addition, there is a lack of uniformity in the method of conducting and reporting of EV. For example, some studies fail to clearly define the ground truth—the benchmark to which the model is trained. Readers are advised to be cautious in interpreting performance when evaluation is limited to an internal or temporal validation set—as geographical EV should be used to assess 'true' performance and generalizability. When conducting a study using a form of Artificial Intelligence (AI), the use of standardized methodology such as the Clinical Artificial Intelligence Research (CAIR) checklist, Standard Protocol Items: Recommendations for Interventional Trials—Artificial Intelligence (SPIRIT-AI), and CONSolidated Standard for Reporting Trials—Artificial Intelligence (CONSORT-AI) are highly recommended to improve methodological rigor, quality of models, and facilitate eventual implementation in clinical practice. These checklists also require an external validation of the results. Since the publication of this study, the field has evolved, and the recommendations presented in this chapter are being used (standardized methodology) and often required by journals (external validation). In my opinion, it is not a surprise that due to various challenges, external validation is scarcely performed. For example, privacy laws and legal departments of the different hospitals often require many discussions and lengthy documents to be drafted before sharing of data is allowed. The lengthy legal documents are often a barrier to sharing large datasets, thereby hampering real progress in the EV of promising algorithms. Therefore, I expect the field to move towards federated learning—send the model to the data instead of sending data to the model—and improve the external validity of models and their performance.

Artificial Intelligence Fracture Recognition on Computed Tomography: Review of Literature and Recommendations

In **Chapter 3**, we explored the current state and benefits of CNNs in computed tomography (CT) scans and concluded based on 17 studies that for radiologists, CNNs can reduce the time to diagnosis—such as ruling out rib fractures, useful in the acute trauma setting—and the number of missed diagnoses on CT scans while increasing diagnostic accuracy. The omission of EV is rife in studies developing a CNN for orthopaedic trauma, as in line with **Chapter 2**, only a minority of studies performed an EV (6/17). We found great heterogeneity in how many patients were included per study, ranging from 39 to 8529. The phrase ‘the more the better’ often applies to the quantity of data used for training a CNN. However, in this study, there was no clear correlation between the number of patients and performance. It is likely that as datasets grow, the effort to create quality data by experts becomes too great, and quality is sacrificed for quantity. This study does not recommend a minimum of patients to include but rather advocates for a progressive approach where data is added to the training set until sufficient performance or a plateau is reached. Three-dimensional models are rare in the field of computer science, and the use of CNNs for CT and MRI scans has been limited to simple tasks such as detecting fractures. Once computer scientists can develop advanced models that can process three-dimensional data, I expect that more studies will explore predictions of CNNs for clinically important decisions. This could for example mean that in the future patients are seen in the Emergency Department where imaging is performed, and initial decision-making is done with support of an AI prediction model. Meaning less wait time in the ER, and hopefully less ‘ramping’ of ambulances or crowded waiting rooms.

Development and External Validation of Automated Detection, Classification, and Localization of Ankle Fractures: Inside the Black Box of a Convolutional Neural Network

In **Chapter 4**, we developed and externally validated the first convolutional neural network (CNN) that can accurately detect, classify, and localize fibula fractures, with an accuracy of 89%. Detection and classification of fractures is not new; however, combining these with segmented (highlighted) localization of fracture lines is. Compared to colleagues Olczak et al., our model needed only one fourth of the data to achieve a better performance, underlining the efficiency of training a preliminary CNN to select optimal cases to learn from. One of the great challenges of these Machine Learning models is that these are essentially a ‘black box’—it is unclear how they arrive at their predictions. The localization used in this study provides a visual presentation as to where the CNN predicts a fracture to be, which can be validated by us—humans. We found that the model—despite its good performance—has difficulties in differentiating between ‘Weber B’ and oblique ‘Weber C’ fractures. These two



types look alike in their configuration but are discerned based on their location, at the level of the syndesmosis or above the syndesmosis, respectively. This also remains a clinical challenge, where the classification of these two injuries is often ambiguous. As this study progressed, it became clear that the computer struggles with tasks that humans also find challenging. In daily practice, this situation is a common occurrence, where patients are regularly informed that it is unclear if their ankle fracture is stable or unstable, and the need for syndesmotic fixation can only be assessed intraoperatively with certainty. The localization was initially meant to improve the accuracy of the model by limiting it to the tibia and fibula. However, a newer model that we have trained using the same data proved that pointing the model in the general direction of the fractures using a 'scribble' line—akin to drawing a circle on paper to highlight something of interest—results in better performance, and segmentation turns out to be superfluous. In addition, this study again iterated that 'garbage in = garbage out', as fractures where surgeons immediately agreed on for its classification were classified with higher accuracy by the algorithm than fractures that warranted extra discussion to reach a consensus. It must be noted that the 'Weber A' and 'Weber C' classifications have a specificity of 100%. Especially the 'Weber C' is interesting, as these often require further assessment for stability and often require surgical fixation. Thus, this model may have a place in the Emergency Department to aid junior doctors in recognizing potentially unstable ankle fractures. But may also improve agreement between surgeons, for example in classifying fractures that are plagued by great interobserver variability—reducing human bias. The question remains, should we teach the computer how to classify certain fractures, or once these models are outpacing us, should they teach us? I foresee a future where the computer is much more equipped to consistently sort fractures according to a computer-designed classification (i.e. complex fracture pattern recognition) and aids us in linking different groups of fractures with specific treatments and outcomes—facilitating data-driven shared decision-making with patients.

External Validation of an Artificial Intelligence Multilabel Deep Learning Model Capable of Ankle Fracture Classification

In **Chapter 5**, we collaborated with Swedish colleagues to externally validate their multi-label CNN that classifies ankle fractures according to the AO/OTA system. The model performed well on our external data, despite a different distribution of types of fractures and fewer radiographic views per patient. This is one of the few studies that externally validates its data, a crucial step towards more robust models. This study is exemplary for external validation, using an international collaboration and the challenges they entail. What I discovered during this study was the number of challenges for labeling data with experts in three countries. Luckily, our Swedish colleagues developed an online labelling platform where we could log in and label the data. However, I realize not every research group can build its own platform. If needed, there are commercial alternatives available. With a small introduction and 'supervised'

labelling to smooth out the initial learning curve, this proved to be an efficient approach. Another valuable take-away was how crucial the consensus meetings were—especially as there were 40 possible classifications—to ensure consistency across experts. In the future, I expect international collaborations to build increasingly advanced models, which will likely use federated learning (i.e. transfer of the algorithm instead of the data itself) to reduce roadblocks such as dealing with complicated and time-costly legislation, such as ‘data transfer agreements’.

Artificial Intelligence and 3D-Guided Surgery in Orthopaedic Trauma: Why, How and What—as a book chapter in Rockwood and Green

Chapter 6 is a book chapter that we wrote for Rockwood and Green, describing Artificial Intelligence (AI) in orthopaedic trauma in detail, why it is used, how it is employed, and what the next steps are. The focus of this chapter is on the computer aiding decision-making, providing probabilities based on data, and reducing the tedious part of the workload of doctors, versus replacing us entirely. In Orthopaedic Trauma specifically, treatment often depends on where and by whom you are treated. If an AI model is trained well, the ground truth is well defined and bias mostly removed, this can help us to create uniform treatment strategies. The computer does not tire, does not change its decisions based on isolated negative experiences—such as adverse events, or personal circumstances—and is consistent. However, doctors need to oversee the computer, not vice versa. Interestingly perhaps, is the question what will happen when doctors deviate from AI predictions, and turn out to be wrong? Or will these questions and statements age like the one I was repeatedly told in middle school?; “Learning to quickly add, subtract, or divide numbers is important as an adult, you cannot bring a calculator everywhere you go”. Another important discussion is “who should develop these models?”. I expect a shared approach in medicine, open-source models developed by scientists and new stakeholders such as technology companies leveraging their expertise and sources of data. In the end, I believe the goal should be to use computers to overcome human biases and to support shared decision-making for individual patients.

PART II Advanced Imaging in Paediatric Ankle Fracture Trauma

Triplane Ankle Fracture Patterns in Paediatric Patients: Extent of Physeal Closure Does Not Dictate Pathoanatomy

In **Chapter 7**, we challenged a theory that has been part of classic teaching since 1964, using advanced imaging in the form of fracture maps. The extent of closure of the physis was thought to determine the fracture configuration of triplane fractures. In contrast to this long taught posit, we concluded in this chapter that not the physis, but ligaments play a pivotal role



in the configuration of these fractures. The corresponding characteristic Y-shaped fracture pattern at the level of the joint is ubiquitous and independent of sex or age, also when adjusted for the difference in bone age between boys and girls. It remains a challenge to determine whether the ligaments serve as a passive protection of the parts where they attach, passively in increased stress adjacent to their attachments, or play an active role in pulling off the fragments. However, when combining the findings in this study with the trauma mechanism, an active role is tenable. Especially as in adults, isolated avulsions of the anterior and posterior tibia are reported, like the Tillaux-Chaput and Volkmann fractures. If ligaments play an active role in triplane fractures, they must share similarities with the adult equivalent of a supination external rotation injury—the trimalleolar fracture. We explored this theory in **Chapter 8**.

Understanding the Mechanism of Injury and Fracture Pattern of Paediatric Triplane Ankle Fractures versus Adult Trimalleolar Fractures

Chapter 8 builds on the findings in **Chapter 7**, where we concluded that the principal determinant for the triplane fracture pattern is not the extent of physis closure but the ligaments. Following this conclusion, triplanes should then correspond to adult ankle fractures that are caused by an identical trauma mechanism, namely a supination external rotation injury—the trimalleolar fracture. In this chapter, with the use of fracture maps and 3D analysis, we introduced the concept that triplane fractures consist of four stages in their sequence of injury, following the description by Lauge-Hansen of Supination External Rotation injuries. We found that triplane and trimalleolar fractures both follow the stages described by Lauge-Hansen, emphasizing that paediatric triplane and adult trimalleolar fractures share a similar injury mechanism. However, there are some differences due to the closure of the physis in triplane fractures, resulting in a larger anterolateral avulsed fragment (Tillaux) and a subsequent different configuration of the posterior fragment. Following the Lauge-Hansen classification in triplanes, there are arguments to suggest the optimal sequence of reduction and fixation techniques—namely, reversing the sequence of the trauma mechanism. For minimally displaced fractures, first, one should apply anteriorly directed force to reduce the posterior fragment, and subsequently internally rotate the foot to reduce the anterolateral fragment, before applying a cast. Ideally, these patients should receive a cast in slight internal rotation to prevent secondary displacement. When internally fixing more displaced injuries, first, fixate the posterior fragment by placing an anteroposterior screw above the physis, remove any periosteum stuck between fracture fragments and then apply internal rotation to fixate the anterolateral fragment with a lateral to medial (or vice versa) screw. The concepts introduced in this chapter, together with previous work in **Chapter 7**, can be a guide for surgeons, emergency, and junior doctors for fracture reduction and fixation of triplane fractures.

PART III Prospective Analysis of Ankle Fractures

Subtle Factors Associated with Outcome of Ankle Fractures: Do Not Forget Intra-Articular Loose Bodies

Chapter 9 explored possible 'subtle' predictors associated with the outcome of ankle fractures using post-operative CT scans and Foot and Ankle Outcome Scores (FAOS) at two years postoperative. Despite the high volume of these injuries, surgical expertise and comfort in treating them, and diverse studies, ankle fractures suffer from relatively poor outcomes, and the exact cause remains elusive. This chapter uncovers at least part of the puzzle, as we found that—aside from the Haraguchi classification—postoperative loose bodies of any size are associated with worse FAOS scores at two years postoperative. However, in contrast to other studies, AO/OTA fracture classification nor posterior malleolar fractures did not significantly affect outcome at two years. The conclusion of this study is supported by two other studies in the literature, and in addition, these also recommend the routine use of arthroscopy when a loose body can be appreciated on the pre-operative CT scan. Studies report an added 10-15 minutes to operating time and no significant added complications when used together with internal fixation. This can be the right approach, allowing careful consideration of loose bodies formed from just cartilage—as these will be missed on CT and can still cause issues in these patients. An extra effort should be taken to remove loose bodies and look for potential osteochondral lesions. Even if an ankle fracture appears simple, there could be more happening under the surface. For example, more cartilage damage than is expected based on the fracture type on the radiographs or CT scan. In patients that remain in pain without new findings on repeat radiographs, MRI can be valuable, as it may reveal small cartilage loose bodies or osteochondral defects. Based on these findings, routine use of arthroscopy deserves a place in the treatment of ankle fractures.



SUMMARY

Diagnostics for ankle fractures mostly relies on imaging. Conventionally with radiographs and more frequently combined with Computed Tomography (CT) scans. Given the availability of large datasets of radiographs and CT scans, advanced imaging analyses can be performed by using Artificial Intelligence (AI) applications and 3D technology. We hypothesized that advancements in imaging modalities can improve diagnostics, understanding, and treatment of ankle fractures. Therefore, this thesis explores ankle fractures in adults and children, using advanced imaging such as Convolutional Neural Networks (CNNs), fracture maps, and 3D reconstructions. **Chapter 1** introduces the why, how and what of AI in trauma surgery.

PART I EXPLORING CONVOLUTIONAL NEURAL NETWORKS IN ANKLE FRACTURES

An Increasing Number of Convolutional Neural Networks for Fracture Recognition and Classification in Orthopaedics: are these Externally Validated and Ready for Clinical Application?

Chapter 2 revealed that although numerous studies have demonstrated impressive accuracy in detecting and categorizing fractures on radiographs using CNNs, only a few incorporate external validation (EV) into their performance evaluation process. Furthermore, this analysis underscores that geographical validation yields better results compared to temporal validation, which involves separating the validation set based on time. Despite the widespread acknowledgment of the significance and necessity of EV in CNN studies, this review underscores the persistent scarcity of such validation practices.

Artificial Intelligence Fracture Recognition on Computed Tomography: Review of Literature and Recommendations

Chapter 3 explored CNNs in analyzing CT scans. Drawing insights from 17 studies, we concluded that CNNs hold the potential to improve diagnostics in clinical practice, particularly in scenarios like the rapid assessment of rib fractures in acute trauma cases. Additionally, CNNs exhibit the capacity to diminish the occurrence of missed diagnoses on CT scans while concurrently enhancing diagnostic precision. The omission of EV remains rife in studies focusing on CNN development for orthopaedic trauma, mirroring the findings of **Chapter 2**, where only a minority of studies incorporated EV. Furthermore, our investigation unveiled substantial variability in the sample sizes across these studies, spanning from 39 to 8529 patients.

Development and External Validation of Automated Detection, Classification, and Localization of Ankle Fractures: Inside the Black Box of a Convolutional Neural Network

In **Chapter 4**, we developed and externally validated a convolutional neural network (CNN) capable of accurately detecting, categorizing, and pinpointing fibula fractures, achieving an impressive accuracy rate of 89%. While fracture detection and classification are not novel, the integration of these tasks with segmented localization of fracture lines represents a significant advancement. Notably, when compared to other models with similar tasks, ours demonstrated superior performance while requiring only one-fourth of the data, underscoring the efficiency of training an initial CNN to select optimal learning cases. A major obstacle in utilizing these machine learning models is their inherent “black box” nature, leaving the process behind their predictions unclear. However, the localization technique employed in our study offers a visual representation of the CNN’s fracture predictions, enabling validation by human experts, thus bridging the interpretability gap.

External Validation of an Artificial Intelligence Multilabel Deep Learning Model Capable of Ankle Fracture Classification

In **Chapter 5**, our collaboration with Swedish colleagues led to the external validation of their multi-label CNN, designed to classify ankle fractures based on the AO/OTA system. Remarkably, the model exhibited strong performance on our external dataset, despite variations in the distribution of fracture types and fewer radiographic views per patient. This study stands out as one of the rare instances where external validation of data was conducted, representing a pivotal stride toward fostering more resilient models in the field. This study is exemplary for external validation using an international collaboration and the challenges they entail.

Artificial Intelligence and 3D-Guided Surgery in Orthopaedic Trauma: Why, How and What—as a book chapter in Rockwood and Green

Chapter 6, authored by Rockwood and Green, offers an exhaustive exploration of the clinical implementation of AI in Orthopaedic Trauma. Delving into the rationale behind its utilization, the chapter elucidates how AI is harnessed and outlines the forthcoming endeavors in the field. Central is the role of AI in augmenting decision-making processes, with probabilistic insights grounded in data, and alleviating the burdensome aspects of doctors’ workloads. In Orthopaedic Trauma, treatment outcomes often hinge on the specifics of where and by whom the patient is treated. A well-trained AI model, characterized by clearly defined ground truths and minimal bias, holds the potential to streamline and standardize treatment approaches, thereby fostering uniformity in clinical strategies.



PART II Advanced Imaging in Paediatric Ankle Fractures

Triplane Ankle Fracture Patterns in Paediatric Patients: Extent of Physeal Closure Does Not Dictate Pathoanatomy

In **Chapter 7**, we challenged a theory that has been part of traditional teaching since 1964, leveraging advanced imaging in the form of fracture maps. Historically, the extent of physeal closure has been believed to dictate the configuration of triplane fractures. Contrary to this longstanding notion, our findings in this chapter revealed a paradigm shift: while the physeal closure does have some influence, it is primarily the ligaments (PITFL, AITFL, and medial deltoid) that exert a pivotal role in shaping these fractures. We observed a characteristic Y-shaped fracture pattern at the joint level, a hallmark feature that proved consistent across sexes and ages, even after accounting for differences in bone age between boys and girls.

Understanding the Mechanism of Injury and Fracture Pattern of Paediatric Triplane Ankle Fractures versus Adult Trimalleolar Fractures

Chapter 8 extends upon **Chapter 7**, where we challenged the conventional belief regarding triplane fracture patterns, emphasizing the importance of ligaments over physeal closure. Hence, we posited that triplane fractures should correspond to adult ankle fractures resulting from a similar trauma mechanism, specifically the supination external rotation injury—such as a trimalleolar fracture. Through the utilization of fracture maps and 3D analysis, we introduced the concept that triplane fractures unfold with the stages delineated in the Lauge-Hansen Supination External Rotation injury classification. Our investigation unveiled a striking parallel between triplane and trimalleolar fractures, both adhering to the injury sequence elucidated by Lauge-Hansen, thus underscoring the shared underlying injury mechanism between pediatric triplane and adult trimalleolar fractures.

PART III Prospective Analysis of Ankle Fractures

Subtle Factors Associated with Outcome of Ankle Fractures: Do Not Forget Intra-Articular Loose Bodies

Chapter 9 delved into potential “subtle” predictors linked to the outcomes of ankle fractures by leveraging post-operative CT scans and Foot and Ankle Outcome Scores (FAOS) assessed at the two-year postoperative mark. Despite the prevalence of these injuries, the abundance of surgical expertise, and the diverse array of studies, ankle fracture fixation often results in suboptimal outcomes. This chapter sheds light on a portion of this, revealing that, beyond the Haraguchi classification, the presence of postoperative loose bodies of any size—as a hallmark of osteochondral injury—correlates with poorer FAOS scores two years postoperative.

NEDERLANDSE SAMENVATTING

De diagnostiek van enkelbreuken vindt plaats naast een gedegen anamnese en lichamelijk onderzoek, met beeldvorming. Doorgaans met röntgenfoto's, maar tegenwoordig ook steeds vaker gecombineerd met Computed Tomography (CT)-scans. Gezien de beschikbaarheid van grote datasets van röntgenfoto's en CT-scans, kunnen geavanceerde analyses gedaan worden met behulp van kunstmatige intelligentie (AI) en 3D-technologie. Dit proefschrift onderzocht de hypothese dat vooruitgangen in beeldvorming de diagnostiek, de kennis en de behandeling van enkelbreuken kunnen verbeteren. Daarom verkent dit proefschrift enkelbreuken bij zowel volwassenen als kinderen, met gebruik van geavanceerde beeldvorming zoals Convolutional Neural Networks (CNN's), 'fracture maps', en 3D-reconstructies. **Hoofdstuk 1** introduceert waarom, hoe en waarmee AI gebruikt wordt binnen de traumachirurgie.

PART I EXPLORING CONVOLUTIONAL NEURAL NETWORKS IN ANKLE FRACTURES

An Increasing Number of Convolutional Neural Networks for Fracture Recognition and Classification in Orthopaedics: Are these Externally Validated and Ready for Clinical Application?

Hoofdstuk 2 toont aan dat hoewel talloze studies indrukwekkende nauwkeurigheid laten zien in het detecteren en classificeren van breuken op röntgenfoto's met behulp van CNN's, slechts enkele studies externe validatie (EV) toepassen. Deze systematische review benadrukt verder dat geografische validatie een betere methode is om de kwaliteit van een model te evalueren in vergelijking met temporele validatie—het scheiden van de validatie- en trainingdataset op basis van tijd. Ondanks de brede erkenning van het belang en de noodzaak van EV in de studies in deze systematische review, blijft er een gebrek aan het gebruik ervan. Dit betekent dat van veel gepubliceerde modellen er nog niet genoeg bekend is over hoe goed deze werken in verschillende klinische omgevingen, met wisselende patiënt populaties.

Artificial Intelligence Fracture Recognition on Computed Tomography: Review of Literature and Recommendations

Hoofdstuk 3 onderzocht het gebruik van CNN's in de analyse van CT-scans. Op basis van 17 studies concludeerden we dat CNN's de potentie hebben om de diagnostiek in de klinische praktijk te verbeteren, bijvoorbeeld voor de snelle beoordeling van gebroken ribben bij de opvang van acute ongevallen. Bovendien blijken CNN's in staat te zijn om het aantal gemiste diagnoses op CT-scans te verminderen, terwijl tegelijkertijd het aantal juiste diagnoses toeneemt. Echter, net zoals in **Hoofdstuk 2**, was er een gebrek aan externe validatie. Ons



onderzoek toonde aanzienlijke variabiliteit in het aantal patiënten dat werd gebruikt voor het trainen van de CNN's in deze studies, variërend van 39 tot 8529.

Development and External Validation of Automated Detection, Classification, and Localization of Ankle Fractures: Inside the Black Box of a Convolutional Neural Network

In **Hoofdstuk 4** ontwikkelden we een convolutional neural network (CNN) dat zeer nauwkeurig fibulafracturen (kuitbeenbreuken) kan detecteren, classificeren en lokaliseren, met een indrukwekkende nauwkeurigheid van 89%. Hoewel breukdetectie en -classificatie niet nieuw zijn, is de integratie van deze taken met gelokaliseerde segmentatie—dat wil zeggen het aanduiden van de specifieke plek—van breuklijnen een belangrijke vooruitgang. Opmerkelijk is dat ons model, vergeleken met andere modellen, superieure prestaties vertoonde, terwijl het getraind is met slechts één vierde van het aantal patiënten. Dit benadrukt het nut van het trainen van een initiële CNN, die patiënten selecteert waar het uiteindelijke CNN zo efficiënt mogelijk van kan leren. Een groot obstakel bij het gebruik van deze machine learning-modellen is de beruchte “zwarte doos”, waardoor het proces achter hun voorspellingen onduidelijk blijft. De segmentatie die in ons onderzoek werd gebruikt, biedt echter een visuele representatie van het CNN, waardoor validatie door dokters gefaciliteerd wordt en de kloof in interpretatie overbrugd.

External Validation of an Artificial Intelligence Multilabel Deep Learning Model Capable of Ankle Fracture Classification

In **Hoofdstuk 5** leidde onze samenwerking met Zweedse collegae tot de externe validatie van hun multi-label CNN, ontworpen om enkelbreuken te classificeren op basis van de AO/OTA-classificatie. Opmerkelijk is dat het model goede prestaties vertoonde op onze externe dataset, ondanks variaties in de verdeling van breuktypen en minder röntgen views—in onze dataset miste de mortise view—per patiënt. Deze studie valt op als een van de zeldzame gevallen waarin externe validatie werd uitgevoerd, en vormt een belangrijke stap naar ontwikkelen van robuustere modellen. Deze studie is een voorbeeld van externe validatie door internationale samenwerking en de uitdagingen die daarmee gepaard gaan.

Artificial Intelligence and 3D-Guided Surgery in Orthopaedic Trauma: Why, How and What—as a book chapter in Rockwood and Green

Hoofdstuk 6, als deel van het toonaangevende boek 'Rockwood and Green: Fractures in Adults', biedt een uitgebreide verkenning van de klinische implementatie van AI in traumatologie. Het hoofdstuk gaat in op de vraag waarom AI gebruikt wordt, hoe AI wordt ingezet en schetst de toekomstige inspanningen op dit gebied. Centraal staat de rol van AI bij het versterken van

besluitvormingsprocessen, met voorspellende inzichten gebaseerd op kansen, en het verlichten van de werkdruk van artsen. Binnen de traumatologie hangt de behandeling regelmatig af van specifieke omstandigheden—namelijk waar en door wie de patiënt wordt behandeld. Een goed getraind AI-model, gekarakteriseerd door duidelijk gedefinieerde 'Ground Truth' en zo min mogelijke vooroordelen, heeft potentie om behandelmethoden te stroomlijnen en te standaardiseren, met als uiteindelijke doel; uniformiteit in de behandeling van specifieke letsels.

PART II Advanced Imaging in Paediatric Ankle Fractures

Triplane Ankle Fracture Patterns in Paediatric Patients: Extent of Physeal Closure Does Not Dictate Pathoanatomy

In **Hoofdstuk 7** onderzochten we een theorie die sinds 1964 als klassiek gedachtegoed is opgenomen binnen de traumatologie. We hebben deze theorie getest door gebruik te maken van geavanceerde beeldvorming in de vorm van 'fracture maps'—een afbeelding met daarop een verzameling van breuk patronen op een specifieke anatomische locatie. Historisch werd aangenomen dat de mate van sluiting van de groeischijf bepalend was voor de configuratie van triplane breuken. In tegenstelling tot dit klassieke gedachtegoed, zorgen onze bevindingen in dit hoofdstuk voor een paradigma verschuiving: hoewel de groeischijf enige invloed heeft, zijn het vooral de ligamenten (posterieure-inferieure tibiofibulaire ligament, anterieure-inferieure tibiofibulaire ligament en lig. deltoideum) die een cruciale rol spelen bij het vormen van deze fracturen. We observeerden een kenmerkend Y-vormig fractuurpatroon ter hoogte van het gewricht, een kenmerk dat consistent was ongeacht leeftijd, en bij zowel jongens als meisjes, zelfs nadat we corrigeerden voor de verschillen in botleeftijd—een verschil van ongeveer twee jaar.

Understanding the Mechanism of Injury and Fracture Pattern of Paediatric Triplane Ankle Fractures versus Adult Trimalleolar Fractures

Hoofdstuk 8 gaat verder waar **Hoofdstuk 7** stopt, namelijk waar we de traditionele opvatting over triplane fractuur patronen testten en het belang van ligamenten boven groeischijf sluiting benadrukten. Gezien ligamenten een minstens zo belangrijkere rol spelen in deze letsels als de groeischijf, stelden we dat triplane breuken daarom overeen moeten komen met volwassen enkelbreuken die het gevolg zijn van een soortgelijk trauma-mechanisme—de trimalleolaire enkelfractuur. Ons onderzoek onthult een opvallende overeenkomst tussen triplane en trimalleolaire fracturen, die beiden het supinatie externe-rotatie patroon volgen zoals beschreven door Lauge-Hansen, wat het onderliggende gemeenschappelijke letsel mechanisme tussen de triplane fracturen bij kinderen en volwassen trimalleolaire fracturen benadrukt. Op basis van deze bevinding, suggereren we de optimale manier om triplane fracturen gesloten en open te reduceren en fixeren.



PART III Prospective Analysis of Ankle Fractures

Subtle Factors Associated With Outcome of Ankle Fractures: Do Not Forget Intra-Articular Loose Bodies

In **Hoofdstuk 9** onderzochten we mogelijke “subtiele” voorspellers die verband houden met de uitkomsten van enkelfracturen door gebruik te maken van postoperatieve CT-scans en de Foot and Ankle Outcome Scores (FAOS) tweejaar postoperatief. Ondanks de prevalentie van deze letsels, de overvloed aan chirurgische expertise en de diverse reeks aan studies, leidt de behandeling van enkelfracturen vaak tot suboptimale uitkomsten. Dit hoofdstuk werpt licht op een gedeelte hiervan, door aan te tonen dat, naast de Haraguchi-classificatie, de aanwezigheid van postoperatieve corpora aliena onafhankelijk van hun grootte—als kenmerk van osteochondrale letsels—geassocieerd is met slechtere functionele uitkomsten (FAOS-scores) twee jaar na de operatie.





Appendix

Bibliography and Presentations
Parameters of Esteem
List of Co-Authors and Affiliations
Dankwoord
Curriculum Vitae

BIBLIOGRAPHY

Chapter 1 J. Prijs, Z. Liao, S. Ashkani-Esfahani, J. Olczak, M. Gordon, P. Jayakumar, P.C. Jutte, R.L. Jaarsma, F.F.A. IJpma, J.N. Doornberg and Consortium for Machine Learning (2022). "Artificial intelligence and computer vision in orthopaedic trauma: the why, how, and what." *Bone Joint J* 104-B(8): 911-914.

PART I EXPLORING CONVOLUTIONAL NEURAL NETWORKS IN ANKLE FRACTURES

Chapter 2 E.C.L. Oliveira, A. van den Merkhof, J. Olczak, M. Gordon, P.C. Jutte, R.L. Jaarsma, F.F.A. IJpma, J.N. Doornberg, J. Prijs (2021). "An increasing number of convolutional neural networks for fracture recognition and classification in orthopaedics: are these externally validated and ready for clinical application?" *Bone Jt Open* 2(10): 879-885.

Chapter 3 L.H.M. Dankelman, S. Schilstra, F.F.A. IJpma, J.N. Doornberg, J.W. Colaris, M.H.J. Verhofstad, M.M.E. Wijffels, J. Prijs (2023). "Artificial intelligence fracture recognition on computed tomography: review of literature and recommendations." *Eur J Trauma Emerg Surg* 49(2): 681-691.

Chapter 4 J. Prijs, Z. Liao, M.S. To, J. Verjans, P.C. Jutte, V. Stirlor, J. Olczak, M. Gordon, D. Guss, C.W. DiGiovanni, R. L. Jaarsma, F.F.A. IJpma, J.N. Doornberg and C. Machine Learning (2023). "Development and external validation of automated detection, classification, and localization of ankle fractures: inside the black box of a convolutional neural network (CNN)." *Eur J Trauma Emerg Surg* 49(2): 1057-1069.

Chapter 5 J. Olczak, J. Prijs, F.F.A. IJpma, F. Wallin, E. Akbarian, J.N. Doornberg and M. Gordon (2024). "External validation of an artificial intelligence multi-label deep learning model capable of ankle fracture classification." *BMC Musculoskelet Disord* 25(1): 788.

Chapter 6 Artificial Intelligence and 3D-Guided Surgery in Orthopaedic Trauma: Why, How, and What—a book chapter in *Rockwood and Green J.N. Doornberg, F.F.A. IJpma, P. Jayakumar, V.M.A. Stirlor, N. Assink, J. Prijs Rockwood and Green's Fractures in Adults, 10e (2024). Paul Tornetta, III; William M. Ricci, Robert F. Ostrum, Michael D. McKee, Benjamin J. Ollivere, Vincent A. de Ridder*

PART II Advanced Imaging in Paediatric Ankle Fracture Trauma

Chapter 7 J. Prijs, J. Rawat, K. Ten Duis, F.F.A. IJpma, J.N. Doornberg, B. Jadav and R.L. Jaarsma (2023). "Triplane ankle fracture patterns in paediatric patients." *Bone Joint J* 105-B(11): 1226-1232.

Chapter 8 J. Prijs, J. Rawat, K. Ten Duis, N. Assink, J.S. Harbers, J.N. Doornberg, B. Jadav, R.L. Jaarsma and F.F.A. IJpma (2024). "Understanding the mechanism of injury and fracture pattern of paediatric triplane ankle fractures versus adult trimalleolar fractures." *Bone Joint J* 106-B(9): 1008-1014.

PART III Prospective Analysis of Ankle Fractures

Chapter 9 J. Prijs, D.T.M. Meijer, R.P. Blom, R-J de Muinck Keizer, T. Schepers, I. Sierevelt, J.C. Goslings, F.F.A. IJpma, J.N. Doornberg, R.L. Jaarsma. Subtle Factors Associated with Outcome of Ankle Fractures: Do Not Forget Intra-Articular Loose Bodies. *Submitted to Foot and Ankle International*



PRESENTATIONS

Annual Australian Orthopaedic Association Scientific Meeting, *Melbourne, Australia*

Australian Orthopaedic Association Scientific February Meeting, *Adelaide, Australia*

Traumadagen, *Amsterdam, Nederland*

Orthopaedic Trauma Association Meeting, *Tampa, USA*

The European Federation of National Associations of Orthopaedics and Trauma-tology Congress,
Lisbon, Portugal

Chirurgendagen (*abstract accepted, congres geannuleerd wegens COVID*)

Traumadagen, *Amsterdam, The Netherlands*

Symposium Experimenteel Onderzoek Heelkundige Specialismen, *Groningen, The Netherlands*

Orthopaedic Trauma Association Meeting, *Fort Worth, USA*

The European Federation of National Associations of Orthopaedics and Trauma-tology Congress,
Vienna, Austria

Australian Orthopaedic Association February Meeting, *Adelaide, Australia*

Australian Orthopaedic Association August Meeting, *Adelaide, Australia*

Australian Orthopaedic Association February Meeting, *Adelaide, Australia*

PARAMETERS OF ESTEEM

RJ Bauze Price for best paper and abstract, Australian Orthopaedic Association

Featured Author of the Bone and Joint Open Journal February

Prins Bernhard Cultuurfonds

Flinders PhD Orthopaedic Trauma Scholarship

ZonMw Translationeel Onderzoek Reisbeurs

Prof. Michaël van Vloten fonds



LIST OF CO-AUTHORS AND AFFILIATIONS

E. Akbarian, MD PhD

Danderyd University Hospital, Karolinska Institute, Stockholm, Sweden.

S. Ashkani-Esfahani, MD PhD

Department of Orthopaedic Surgery, Massachusetts General Hospital, Boston, USA.

N. Assink, PhD

Department of Trauma Surgery, University of Groningen, University Medical Center Groningen, Groningen, Netherlands.

R.P. Blom, MD PhD

Department of Surgery, Amsterdam, The Netherlands, Amsterdam University Medical Centre.

J.W. Colaris, MD PhD

Department of Orthopedics, Erasmus University Medical Centre, Rotterdam, The Netherlands.

L.H.M. Dankelman, BSc

Department of Surgery, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands.

C.W. DiGiovanni, MD PhD

Department of Orthopaedic Surgery, Massachusetts General Hospital, Boston, USA.
Harvard Medical School, Boston, USA.

J.N. Doornberg, MD PhD

Department of Surgery and Orthopaedic Surgery, University Medical Center Groningen, University of Groningen, Groningen, Netherlands.
Department of Orthopaedic & Trauma Surgery, Flinders University, Flinders Medical Centre, Adelaide, Australia.

K. ten Duis, MD

Department of Surgery, University Medical Center Groningen, University of Groningen, Groningen, Netherlands.

M. Gordon, MD PhD

Danderyd University Hospital, Karolinska Institute, Stockholm, Sweden.

J.C. Goslings, MD PhD

Department of Surgery, Amsterdam, The Netherlands, Amsterdam University Medical Centre.

D. Guss, MD PhD

Department of Orthopaedic Surgery, Massachusetts General Hospital, Boston, USA.
Harvard Medical School, Boston, USA.

F.F.A. IJpma, MD PhD

Department of Surgery, University Medical Center Groningen, University of Groningen, Groningen, Netherlands.

R.L. Jaarsma, MD PhD

Department of Orthopaedic & Trauma Surgery, Flinders University, Flinders Medical Centre, Adelaide, Australia.

B. Jadav, MD

Department of Orthopaedic & Trauma Surgery, Flinders University, Flinders Medical Centre, Adelaide, Australia.

P. Jayakumar, MD PhD

The University of Texas at Austin, Dell Medical School, Austin, Texas, USA.

P.C. Jutte, MD PhD

Department of Orthopaedic Surgery, University Medical Center Groningen, University of Groningen, Groningen, Netherlands.

Z. Liao, PhD

Australian Institute for Machine Learning, Adelaide, Australia.

D.T.M. Meijer, MD PhD

Department of Surgery, Amsterdam, The Netherlands, Amsterdam University Medical Centre.

R-J de Muinck Keizer, MD PhD

Department of Surgery, Amsterdam, The Netherlands, Amsterdam University Medical Centre.

A. van den Merkhof, MD

Department of Orthopaedic & Trauma Surgery, Flinders University, Flinders Medical Centre, Adelaide, Australia.

J. Olczak, MD

Danderyd University Hospital, Karolinska Institute, Stockholm, Sweden.

L. Oliveira e Carmo, MD

Department of Orthopaedic Surgery, University Medical Centre, University of Groningen, Groningen, Groningen, Netherlands.



J. Rawat, MD

Department of Orthopaedic & Trauma Surgery, Flinders University, Flinders Medical Centre, Adelaide, Australia.

Department of Orthopaedics, Women's and Children's Hospital, Adelaide, Australia.

T. Schepers, MD PhD

Department of Surgery, Amsterdam, The Netherlands, Amsterdam University Medical Centre.

S.R.A. Schilstra, MD

Department of Orthopedic Surgery, Groningen University Medical Centre, Groningen, The Netherlands.

I. Sierevelt, PhD

Specialized Centre for Orthopedic Research and Education (SCORE), Xpert Orthopedics, Amsterdam, The Netherlands.

V.M.A. Stirler, MD PhD

Department of Surgery, Groningen University Medical Centre, Groningen, The Netherlands.

M-S To, MD PhD

College of Medicine and Public Health, Flinders University, Adelaide, Australia.

Department of Neurosurgery, Flinders Medical Center, Adelaide, Australia.

M.H.J. Verhofstad, MD PhD

Department of Surgery, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands.

J.W. Verjans, MD PhD

Australian Institute for Machine Learning, Adelaide, Australia.

F. Wallin, MD

Danderyd University Hospital, Karolinska Institute, Stockholm, Sweden.

M.M.E. Wijffels, MD PhD

Department of Surgery, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands.

DANKWOORD

Geachte lezer,

Dit proefschrift is tot stand gekomen door middel van een internationale samenwerking tussen Nederland en Australië. Dat is niet alleen tekenend voor dit proefschrift, maar ook voor mij als persoon. Het is een kenmerkende gebeurtenis in mijn leven, en ik koester de ervaringen die ik op heb mogen doen op zowel wetenschappelijk, medisch als persoonlijk vlak. De kaft van dit proefschrift combineert beide landen met het onderwerp van dit proefschrift—namelijk enkelfracturen. De roze lijn geeft de route weer die ik afgelegd heb tijdens mijn reizen door Australië en symboliseert de reis van dit proefschrift.

Zonder samen te werken was dit proefschrift nooit gelukt. Het heeft geleid tot onvergetelijke avonturen in Australië en op congressen, hechte vriendschap en natuurlijk veel nieuwe en inspirerende wetenschappelijke inzichten en vaardigheden. Naast alle patiënten die mee hebben gedaan in de onderzoeken in dit proefschrift wil ik graag de volgende personen in het bijzonder bedanken:

Het promotieteam in volgorde van kennismaking:

Dr. F.F.A. IJpma

Beste Frank, waar moet ik beginnen? Onze ontmoeting is de basis van de reis die we samen hebben afgelegd sinds 2018. Wat begon als een project zodat ik naar Boston zou kunnen, is uiteindelijk uitgegroeid tot een hechte samenwerking en vriendschap. Jij was bij elke stap, van beginnende coassistent, promovendus, beginnende dokter en nu in opleiding tot chirurg. Ik ben je enorm dankbaar voor je vertrouwen ('het komt wel goed') en de steun die je vanaf het begin geboden hebt. Ik heb altijd op je kunnen rekenen, als het onderzoek niet zo soepel liep of als er weer eens een deadline op het nippertje gehaald moest worden. Daarnaast natuurlijk de mooie momenten, assisteren op OK in de dienst, daarna CT-scans bekijken tot diep in de nacht, of de stad in, onvergetelijke momenten tijdens de traumadagen, en recent samen opereren! Je bent altijd vol goede ideeën en hebt veel studies beter gemaakt met jouw scherpe blik. Het was een prachtige reis, en ik kijk ernaar uit nog veel van je te leren.

Prof. dr. J.N. Doornberg

Beste Job, jij bent de katalysator geweest in Australië in mijn ontwikkeling van beginnende onderzoeker tot ervaren promovendus. Onze ontmoeting begon in Amsterdam in 2019, en al snel met succes. Samen hebben we al vlot een eerste beurs binnengesleept zodat ik naar Australië kon verhuizen, met nog vele die volgden. Je hebt mij de vrijheid en ondersteuning gegeven om in korte tijd een enorme ontwikkeling door te maken. Jouw creativiteit en energie



zijn aanstekelijk. Je hebt mij gemotiveerd om voor grote (inter)nationale podia ons onderzoek te delen, ervaringen waar ik erg van genoten heb. In het onderzoek was niets te gek of onhaalbaar. Bij uitdagingen denk ik nog regelmatig aan jouw adagium: 'if it was easy, everybody would do it'. Daarnaast vergeet ik natuurlijk nooit meer de gezellige barbecues aan Shore Court of het bouwen van de koffie-kar aan de Onnerweg.

Prof. dr. R.L. Jaarsma

Beste Ruurd, zoals elk nieuw begin bij orthopaedics in Flinders begon onze ontmoeting met een 'latte' bij Theo's, iets wat we daarna nog vaak herhaald hebben. Niet alleen voor cruciale 'roadblocks' in het onderzoek, maar ik waardeer het ook enorm dat je altijd tijd had om over persoonlijke zaken te praten en over doelen in het leven te sparren. Daarnaast zorg jij samen met Pinky voor een warm thuis voor alle onderzoekers en fellows in Adelaide. Een voorbeeld daarvan waren de barbecues rond kerst met surfen in Middleton, die altijd druk bezocht werden! Bij het onderzoek dacht je altijd pragmatisch mee wat er mogelijk was, en hoe dat dan het beste aangevlogen kon worden. Je hebt mij gemotiveerd om zelf initiatief te nemen in verschillende projecten, van een workstation bouwen voor de AI studies, tot samen microfoons in het plafond schroeven voor beter geluid bij de ochtendoverdracht. Door jou zal Australië altijd als een tweede thuis voelen.

Prof. dr. P.C. Jutte

Beste Paul, jij hebt ervoor gezorgd dat ik na een succesvol jaar in Flinders uiteindelijk in Groningen de eerste Adelaide-Groningen MD-PhD'er kon worden. Zonder jouw hulp zou het onmogelijk zijn geweest te navigeren tussen twee universiteiten. De samenwerking tussen Adelaide en Groningen heeft veel betekend voor mijn proefschrift, maar ook voor mijzelf. Als MD-PhD'er kreeg ik de kans opnieuw af te reizen naar Adelaide, op naar een tweede avontuur! Ik wil je bedanken voor je waardevolle adviezen onderweg en de warme ontvangst vanuit Australië op de afdeling orthopedie in het UMCG.

Co-auteurs

Beste coauteurs, ik dank jullie voor de bijdrage aan dit proefschrift. Ik heb erg genoten van jullie kritische noten en harde werk. Ik ben trots op de studies die we samen gepubliceerd hebben, en hoop in de toekomst onze vruchtbare wetenschappelijke samenwerking voort te zetten! Ik wil in het bijzonder nog **dr. Sierevelt** bedanken voor de cruciale hulp bij de statistiek in het laatste hoofdstuk.

Mr. Bhavin Jadav

Dear Bhavin, you are always in a good mood and bursting with clinically relevant research questions! Your observation that the adolescent triplane fracture looks like an adult pilon fracture has led to one of the pearls of this thesis. You inspired me with your drive to educate

during morning handover, providing sharp clinical comments on papers and thorough explanations of anatomy and your thoughts on trauma mechanisms and resulting injuries.

Mr. Jaideep Rawat

Dear Jai, thank you so much for sitting down with me many times in the orthopaedic office and going through the countless CT-scans of children with transitional ankle fractures. I have learned a lot exploring these scans with you. Your expertise in diagnosing and treating these injuries has been essential for two landmark papers on triplane fractures.

Drs. K. ten Duis

Beste Kaj, heel erg bedankt voor de gezellige en leerzame samenwerking aan twee mooie studies in dit proefschrift. Ik waardeer de tijd die je hebt genomen om samen door alle scans te gaan met triplane fractures. Je klinische blik en kritische opmerkingen waren een zeer waardevolle toevoeging. Ik hoop als chirurg in opleiding nog veel van je te leren.

Sylvia McAndrew

Dear Sylvia, you are the glue of the department of orthopaedics in Flinders. Your optimistic mood and humor always made sure it was a pleasure dropping in. I enjoyed our coffees together and treasure the barbecue that you organized in your home (Jerry setting the grill on fire will always be a great memory). I appreciate the advice that you have given me during my time at Flinders. I look forward to sharing a glass of champagne after my defense!

Overige leden van de promotiecommissie: **prof. dr. J.P.P.M. de Vries, prof. dr. M. Maas, prof. dr. P.M.A. van Ooijen and associate prof. dr. T. White.**

Geachte leden van de promotiecommissie, ik dank jullie voor de tijd, kritische vragen en bovenal interesse in dit proefschrift.

Fellows en onderzoekers in Australië

Bedankt voor de fijne tijd die we met elkaar hadden, terwijl we allemaal ver van huis waren. Zowel tijdens het onderzoek, als daarnaast met de barbecues op het strand tijdens kerst, of de uitdagende partijen op de tennisbaan. In het bijzonder wil ik **Reinier Spek** en **Marouska van Boxel** bedanken voor de hulp bij het onderzoek, de gezellige etentjes en leerzame research meetings gedurende een pandemie in een land afgesloten van de rest van de wereld.

Collegae Chirurgie Martini Ziekenhuis

Beste collegae uit het Martini Ziekenhuis, bedankt voor de gezelligheid in de kliniek, vele borrels en jullie aanwezigheid bij het vieren van belangrijke momenten tijdens de afronding van dit proefschrift.



Paranimf Thijs Vaartjes

Beste Thijs, van samen op de basisschool naar samen promoveren--wie had dat ooit gedacht? Het voelt bijzonder om dit belangrijke hoofdstuk in ons leven samen te doorlopen. Naast het onderzoek waardeer ik onze gesprekken over de essentie.

Paranimf Angad Bedi

Beste Angad, onze maandagochtenden in het UMCG om 07.00 uur na de researchmeeting hebben naast een slaapte kort ook geleid tot een waardevolle vriendschap. Ontspanning vind ik niet alleen in onze gesprekken, maar ook samen op de golfbaan.

Ik wil jullie allebei bedanken voor jullie onvermoeibare steun, maar vooral voor jullie vriendschap. Voor de avonden bij de Uurwerker, het nachtelijke ouwehoeren in de Poelestraat, en de onvergetelijke herinneringen die we hebben gemaakt op de traumadagen. De laatste fase van mijn PhD is onlosmakelijk met jullie verbonden. Ik kijk ernaar uit om nog lang van jullie vriendschap te genieten.

Floris

Lieve Floris, tijdens mijn reis met dit proefschrift heb jij je gestort op heel complexe materie tijdens je studie sterrenkunde. Ik ben trots op je dat het je gelukt is de studie zonder moeite af te ronden. Daarnaast geniet ik van onze schaakpartijen, al verlies ik regelmatig door jouw sterke analytische denkwijze. Ik hoop dat je met net zo veel plezier als ik nog een academisch pad zal bewandelen.

Pap en Mam

Lieve Papa en Mama, ik ben jullie erg dankbaar voor de zorgeloze jeugd die jullie ons hebben gegeven. Jullie staan altijd voor Floris en mij klaar, en ik waardeer jullie vrijgevigheid en warmte. Ondanks jullie niet-medische achtergrond, zijn jullie altijd geïnteresseerd en nieuwsgierig geweest naar de studies en uitdagingen van dit proefschrift. Ik hoop nog vele jaren van jullie adviezen en gezelligheid te kunnen genieten.

Anke

Lieve Anke, jij staat altijd aan mijn zij. Samen begonnen met de studie geneeskunde, samen coschappen gelopen, samen naar Australië en allebei als dokter aan het werk. Je bent een enorme steun geweest vanaf het begin. Samen zijn we een waanzinnig avontuur aangegaan 'Down Under'. Ondanks de grote spinnen ging jij vol enthousiasme mee. De herinneringen die we daar hebben gemaakt ga ik nooit vergeten. We hebben zelfs samen een paper geschreven! Daarnaast zorgde je voor de nodige balans door roadtrips naar 'the Outback' te plannen. Ik ben heel dankbaar voor de onvoorwaardelijke liefde en steun die je mij nog dagelijks geeft. Ook ben ik heel trots op de route die jij als dokter bewandelt, en weet zeker dat je een fantastische huisarts zal zijn.

CURRICULUM VITAE

Jasper Prijs was born on 17th of April 1996, in Emmen, the Netherlands. Where he grew up as the oldest son of Peter and Vera Prijs, and brother to Floris Prijs. After obtaining his gymnasium diploma from the Esdal college in Emmen in 2014, he started medical school at the University of Groningen. During his medical studies, he gained an early interest in trauma surgery. Chasing his goal of becoming a surgeon he started conducting research in 2018 with dr. F.F.A. IJpma. During his clinical rotations as a medical student, an interest grew to travel abroad. In 2020 he travelled to Adelaide, Australia and developed his research skills further together with prof. J.N. Doornberg and prof. R.L. Jaarsma. Which formed the basis for this thesis. Together with the team Jasper successfully applied for grants to fund his research and stay in Australia. After a year of performing research, he was accepted into the MD/PhD program of the University of Groningen (under guidance of prof. P.C. Jutte). Finally in 2022 he completed his Master's degree in Medicine, before travelling back to Australia a second time. Where he completed most of his projects before submitting his thesis in May of 2024. Chasing his passion of becoming a surgeon he started as a surgical intern in the Martini Hospital, Groningen (dr. Kelder and drs. Zwaving). After working for a year and a half, he was accepted into the resident training program for general surgery in Groningen and Almelo in 2025 (dr. Lange and dr. Lutke Holzik).



