ADVANCED DIAGNOSTIC STRATEGIES FOR WRIST TRAUMA

David W.G. Langerhuizen

Advanced Diagnostic Strategies for Wrist Trauma

David W.G. Langerhuizen

ISBN: 978-94-6361-651-5

Layout and printing: Optima Grafische Communicatie

This thesis was embedded within the Department of Orthopaedic Surgery, Amsterdam UMC, University of Amsterdam, the Netherlands and the Department of Orthopaedic & Trauma Surgery, Flinders Medical Centre, Flinders University, Adelaide, Australia.

The research described in this thesis was supported by grants from the Amsterdam UMC (AMC PhD Scholarship and AMC Young Talent Fund), Flinders University, Prins Bernhard Cultuurfonds, Prof. Michaël-van Vloten Fonds, Traumaplatform, Anna Fonds, and Amsterdam Universiteitsfonds.

Advanced Diagnostic Strategies for Wrist Trauma

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 11 maart 2022, te 10.00 uur

door

David Wilto Gerard Langerhuizen

geboren te Zevenaar

Promotiecommissie

Promotores:	prof. dr. G.M.M.J. Kerkhoffs	AMC-UvA
	prof. dr. R.L. Jaarsma	Flinders University
Copromotores:	prof. dr. J.N. Doornberg	Rijksuniversiteit Groningen
	dr. S.J. Janssen	AMC-UvA
Overige leden:	prof. dr. D. Eygendaal	AMC-UvA
	prof. dr. B.J. van Royen	Vrije Universiteit Amsterdam
	prof. dr. M. Maas	AMC-UvA
	prof. dr. F. Nollet	AMC-UvA
	prof. dr. P.C. Jutte	Rijksuniversiteit Groningen
	dr. F.F.A. IJpma	Rijksuniversiteit Groningen

Faculteit der Geneeskunde

Dit proefschrift is tot stand gekomen binnen een samenwerkingsverband tussen de Universiteit van Amsterdam en Flinders University, met als doel het behalen van een gezamenlijk doctoraat. Het proefschrift is voorbereid aan de Faculteit der Geneeskunde van de Universiteit van Amsterdam en aan het College of Medicine and Public Health van Flinders University.

The research project leading to this thesis was conducted under a Cotutelle arrangement between Flinders University and the University of Amsterdam. The thesis was prepared at the Faculty of Medicine of the University of Amsterdam and at the College of Medicine and Public Health of Flinders University.

ADVANCED DIAGNOSTIC STRATEGIES FOR WRIST TRAUMA

By

David W.G. Langerhuizen

Thesis
Submitted to Flinders University
for the degree of

Doctor of Philosophy

College of Medicine and Public Health

11 March 2022

This thesis has been written within the framework of the Cotutelle Program, with the purpose of obtaining a joint doctorate degree. The thesis was prepared at the College of Medicine and Public Health of Flinders University and at the Faculty of Medicine of the University of Amsterdam.

I certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Voor mijn ouders

TABLE OF CONTENTS

CHAPTER 1

General Introduction

INTRODUCTION

WRIST TRAUMA - INCIDENCE

In the Netherlands, approximately 34,000 patients with distal radius fractures -18% of all fractures- are treated annually, which makes it one of the most common fractures.[1] About 20% of patients sustaining wrist trauma after a fall on outstretched hand have a fracture of the distal radius.[2] The incidence of distal radius fractures is likely to increase due to an aging population together with more emphasis on sports-related activities among the elderly.[3]

Although the exact number is subject to debate, scaphoid fractures constitute 90% of carpal bone fractures and cumulate to 2-3% of all fractures.[4-6] The preponderance of these fractures peak predominantly in young and active men.[7] While non-union with avascular necrosis can lead to long term sequelae such as wrist arthritis and carpal collapse, accurate and early diagnosis is paramount for optimal treatment.[8]

Currently, distal radius- and scaphoid fractures cumulatively contribute to 20% of all fractures, leading to major healthcare and societal costs.[1,7,9]

ARTIFICIAL INTELLIGENCE

In 1959, Arthur Samuel coined AI as a field of study that enables a computer to learn without needing to be explicitly programmed.[10] Increasing computational power and processing speed resulted in development of AI algorithms across various fields of healthcare, such as dermatology, radiology, ophthalmology, internal medicine, and surgery.[11-16] According to McKinsey & Company, a value up to $100 billion annually could be generated based on efficient adoption of AI in clinical care.[17] Moreover, as a result of widespread implementation of electronic medical records, rapid accumulation of routinely collected data is becoming available.

Utilization of AI applications could potentially mitigate surgeon bias, physician fatigue, and might help overcome current diagnostic and treatment inconsistencies. However, at this stage, it is yet to be elucidated how AI will be integrated into the clinical workflow. Along with traditional analytics, additional study is merited to evaluate AI's potential utility for improving orthopaedic trauma care.

AIMS OF THESIS

This thesis is at a crossroads between the clinical subject of patients with wrist trauma and advanced diagnostic strategies that emerged in the field of artificial intelligence. In five parts, parallel to the clinical pathway of patients sustaining acute wrist trauma, we aim to evaluate clinical applications of artificial intelligence and 3D imaging strategies by encompassing six clinically relevant questions.

RISK STRATIFICATION IN THE EMERGENCY DEPARTMENT

In the emergency department, optimal and efficient use of resources is necessary since it reduces waiting time and associated costs. Criteria have been developed and implemented that indicate which patients suspected of having a fracture of the distal radius should be referred for radiographic evaluation.[18,19] Machine-learning (ML) derived algorithms may help to simplify these existing prediction tools as they are able to identify non-linear associations in data. As such, only the most relevant variables will be incorporated, while maintaining (or improving) diagnostic performance. The first study evaluates four machine learning algorithms to predict the probability of a fracture of the distal radius for patients presenting to the emergency department sustaining acute wrist trauma.

DEEP LEARNING FOR FRACTURE DETECTION

Radiography remains the initial imaging modality for patients with a suspected scaphoid fracture.[20] However, among suspected scaphoid fractures, 1 in every 6 true scaphoid fracture is missed at first presentation on initial radiography.[20] Early and accurate diagnosis of scaphoid fractures is essential since it reduces the risk of long-term non-union and minimizes loss in productivity resulting from unnecessary cast immobilisation.[21-23]

AI algorithms could be developed to extract features from images to derive rules and patterns, thus enabling autonomous predictions with new sets of comparable data. For fracture care, AI might be a useful adjunct to aid certain diagnostic aspects, for example scenarios that are subtle and easily overlooked by humans or during secondary evaluation after complex trauma. To date, AI has focused mainly on commonly displaced and easy detectable fractures, such as proximal humerus and ankle fractures.[11,24] The clinical question to answer is whether application of AI will be beneficial for the detection of radiographically visible and occult scaphoid fractures on radiography?

CLINICAL PREDICTORS FOR SURGICAL DECISION MAKING

In patients with similar fracture patterns, substantial and unexplained variation among surgeons is observed in recommending operative or conservative treatment for distal radius fractures. For instance, about 75% of patients with a fracture of the distal radius in Australia are treated operatively, while only 20-30% in the Netherlands.[25,26] It is a truism that surgeons are biased on the decision whether to operate or not. Decision aids may have the potential to reduce bias by neutralizing the physicians influence as they provide more accurate estimates of an expected treatment outcome. However, patient factors that lead to practice variation among surgeons might first need to be identified before data-driven predictive models can be developed to facilitate optimal treatment recommendation. Via the Science of Variation Group (SOVG), factors are studied that explain the variation in treatment of our patients with a fracture of the distal radius.

3D PRINTING FOR PREOPERATIVE PLANNING

There is great variation among surgeons in reliably interpreting fracture patterns. For instance, the AO-classification for distal radius fractures--assessed on radiographs and computed tomography images--showed only moderate reliability for type C distal radius fractures.[27] Accurate perioperative fracture assessment will improve surgical reduction and fixation. To better visualize 3D aspects of fractured bones, 3D printing has been utilized.[28,29] For instance, addition of 3D hand-held models have shown to improve acetabular fracture classification (that is, the Judet and Letournel classification[30]) compared to using only radiographs and CT imaging.[31] On the contrary, for radial head, coronoid, and distal humerus fractures, evidence demonstrated no or only slight improvement among surgeons when evaluating certain fracture characteristics with an additional 3D hand held model.[32-34] With regards to distal radius fractures, the added preoperative value of 3D printed handheld models has been unclear. Given that these high-volume fractures can be challenging to treat due to its articular involvement, research is merited. The aim is to evaluate whether 3D printed hand-held modelling is a useful adjunct in preoperative planning of distal radius fractures?

3D FLUOROSCOPY FOR INTRAOPERATIVE ASSESSMENT

The 3D shape of the distal radius, mainly due to Lister's tubercle, complicates intraoperative detection of dorsal cortex penetrating screws after volar plating, especially when using conventional 2D fluoroscopy (that is, anteroposterior and elevated lateral projections). To overcome this iatrogenic pitfall, pre-clinical and clinical studies demonstrated the added value of the intraoperative 2D dorsal tangential imaging view (DTV), in which the forearm is placed in 75 degrees inclination with the wrist in flexion.[35-38] In addition, 3D fluoroscopy--used as an intraoperative diagnostic imaging strategy--demonstrated to reduce rates of postoperative revision procedures.[39-41] At the same time, prior evidence suggests that 3D fluoroscopy increased intraoperative screw exchange in patients who underwent operative treatment for distal radius fractures.[42,43] However, to date, the optimal intraoperative imaging strategy for patients who undergo volar plating for a fracture of the distal radius remains subject of debate. This part addresses whether intraoperative 3D fluoroscopy outperforms DTV to decrease the number of dorsal cortex penetrating screws post-operatively?

OUTLINE OF THESIS

PART I: RISK STRATIFICATION IN THE EMERGENCY DEPARTMENT

Chapter 2 aims to develop and validate a machine learning decision supportive tool to predict the probability for fracture of the distal radius in patients presenting to the emergency department after sustaining acute wrist trauma.

PART II: DEEP LEARNING FOR FRACTURE DETECTION

Chapter 3 is a systematic review of aggregated published orthopaedic trauma imaging studies examining the performance of machine-and deep learning algorithms. This review sheds light on the current anatomical locations that are being studied for automated fracture detection and classification. **Chapter 4** assesses utilization of a deep learning algorithm for automated detection of radiographically visible and occult fractures of the scaphoid. In addition, this study also aims to compare the algorithm's diagnostic performance with five human examiners.

PART III: CLINICAL PREDICTORS FOR SURGICAL DECISION MAKING

Chapter 5 is a case-vignette study surveying a large international group of surgeons specialized in fracture surgery to help understand factors that influence recommendation for operative treatment for fractures of the distal radius. This survey study aims to provide insights into predictive factors with relative consensus, disagreement, and areas where more evidence is needed.

PART IV: 3D PRINTING FOR PREOPERATIVE PLANNING

In addition to 2D and 3D CT images, it is reasonable to wonder whether 3D handheld models might help to better assess fracture characterization. **Chapter 6** evaluates the additional pre-operative clinical value of 3D handheld models for recognition of fracture characteristics and agreement on classification of intra-articular distal radius fractures.

PART V: 3D FLUOROSCOPY FOR INTRAOPERATIVE ASSESSMENT

Chapter 7 determines whether intraoperative 3D fluoroscopy is preferred over 2D dorsal tangential views to avoid dorsal screw penetration after volar plating of distal radius fractures. This study wishes to know whether 3D fluoroscopy adds value compared with dorsal tangential views.

REFERENCES

1. Bentohami A, Bosma J, Akkersdijk GJ, van Dijkman B, Goslings JC, Schep NW. Incidence and characteristics of distal radial fractures in an urban population in The Netherlands. *Eur J Trauma Emerg Surg.* 2014;40(3):357-361.
2. van den Brand CL, van Leerdam RH, van Ufford JH, Rhemrev SJ. Is there a need for a clinical decision rule in blunt wrist trauma? *Injury.* 2013;44(11):1615-1619.
3. Court-Brown CM, Clement ND, Duckworth AD, Aitken S, Biant LC, McQueen MM. The spectrum of fractures in the elderly. *Bone Joint J.* 2014;96-b(3):366-372.
4. Hey HW, Chong AK, Murphy D. Prevalence of carpal fracture in Singapore. *J Hand Surg Am.* 2011;36(2):278-283.
5. Hove LM. Epidemiology of scaphoid fractures in Bergen, Norway. *Scand J Plast Reconstr Surg Hand Surg.* 1999;33(4):423-426.
6. van Onselen EB, Karim RB, Hage JJ, Ritt MJ. Prevalence and distribution of hand fractures. *J Hand Surg Br.* 2003;28(5):491-495.
7. Duckworth AD, Jenkins PJ, Aitken SA, Clement ND, Court-Brown CM, McQueen MM. Scaphoid fracture epidemiology. *J Trauma Acute Care Surg.* 2012;72(2):E41-45.
8. Merrell GA, Wolfe SW, Slade JF, 3rd. Treatment of scaphoid nonunions: quantitative meta-analysis of the literature. *J Hand Surg Am.* 2002;27(4):685-691.
9. Kakar S. Clinical Faceoff: Controversies in the Management of Distal Radius Fractures. *Clin Orthop Relat Res.* 2015;473(10):3098-3104.
10. Samuel AL. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development.* 1959;3(3):210-220.
11. 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.
12. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115-118.
13. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA.* 2016;316(22):2402-2410.
14. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PloS one.* 2017;12(4):e0174708.
15. Karhade AV, Thio Q, Ogink PT, et al. Development of Machine Learning Algorithms for Prediction of 30-Day Mortality After Surgery for Spinal Metastasis. *Neurosurgery.* 2019;85(1):E83-E91.
16. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal.* 2017;35:303-312.
17. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning - part3.
18. Mulders MAM, Walenkamp MMJ, Sosef NL, et al. The Amsterdam Wrist Rules to reduce the need for radiography after a suspected distal radius fracture: an implementation study. *Eur J Trauma Emerg Surg.* 2019.
19. Walenkamp MMJ, Bentohami A, Slaar A, et al. The Amsterdam wrist rules: the multicenter prospective derivation and external validation of a clinical decision rule for the use of radiography in acute wrist trauma. *BMC Musculoskelet Disord.* 2015;16(389).
20. Suh N, Grewal R. Controversies and best practices for acute scaphoid fracture management. *J Hand Surg Eur Vol.* 2018;43(1):4-12.

21. Dorsay TA, Major NM, Helms CA. Cost-effectiveness of immediate MR imaging versus traditional follow-up for revealing radiographically occult scaphoid fractures. *AJR Am J Roentgenol.* 2001;177(6):1257-1263.
22. Karl JW, Swart E, Strauch RJ. Diagnosis of Occult Scaphoid Fractures: A Cost-Effectiveness Analysis. *J Bone Joint Surg Am.* 2015;97(22):1860-1868.
23. Langhoff O, Andersen JL. Consequences of late immobilization of scaphoid fractures. *J Hand Surg Br.* 1988;13(1):77-79.
24. 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):672-677.
25. Ansari U, Adie S, Harris IA, Naylor JM. Practice variation in common fracture presentations: a survey of orthopaedic surgeons. *Injury.* 2011;42(4):403-407.
26. Walenkamp MM, Mulders MA, Goslings JC, Westert GP, Schep NW. Analysis of variation in the surgical treatment of patients with distal radial fractures in the Netherlands. *J Hand Surg Eur Vol.* 2016.
27. Jayakumar P, Teunis T, Gimenez BB, Verstreken F, Di Mascio L, Jupiter JB. AO Distal Radius Fracture Classification: Global Perspective on Observer Agreement. *J Wrist Surg.* 2017;6(1):46-53.
28. Chen C, Cai L, Zheng W, Wang J, Guo X, Chen H. The efficacy of using 3D printing models in the treatment of fractures: a randomised clinical trial. *BMC Musculoskelet Disord.* 2019;20(1):65.
29. Misselyn D, Nijs S, Fieuws S, Shaheen E, Schepers T. Improved Interobserver Reliability of the Sanders Classification in Calcaneal Fractures Using Segmented Three-Dimensional Prints. *J Foot Ankle Surg.* 2018;57(3):440-444.
30. Judet R, Judet J, Letournel E. Fractures of the Acetabulum: Classification and Surgical Approaches for Open Reduction. Preliminary Report. *J Bone Joint Surg Am.* 1964;46:1615-1646.
31. Hurson C, Tansey A, O'Donnchadha B, Nicholson P, Rice J, McElwain J. Rapid prototyping in the assessment, classification and preoperative planning of acetabular fractures. *Injury.* 2007;38(10):1158-1162.
32. Brouwer KM, Lindenhovius AL, Dyer GS, Zurakowski D, Mudgal CS, Ring D. Diagnostic accuracy of 2- and 3-dimensional imaging and modeling of distal humerus fractures. *J Shoulder Elbow Surg.* 2012;21(6):772-776.
33. Guitton TG, Brouwer K, Lindenhovius AL, et al. Diagnostic accuracy of two-dimensional and three-dimensional imaging and modeling of radial head fractures. *J Hand Microsurg.* 2014;6(1):13-17.
34. Guitton TG, Kinaci A, Ring D. Diagnostic accuracy of 2- and 3-dimensional computed tomography and solid modeling of coronoid fractures. *J Shoulder Elbow Surg.* 2013;22(6):782-786.
35. Bergsma M, Doornberg JN, Duit R, et al. Volar plating in distal radius fractures: A prospective clinical study on efficacy of dorsal tangential views to avoid screw penetration. *Injury.* 2018;49(10):1810-1815.
36. Brunner A, Siebert C, Stieger C, Kastius A, Link BC, Babst R. The dorsal tangential X-ray view to determine dorsal screw penetration during volar plating of distal radius fractures. *J Hand Surg Am.* 2015;40(1):27-33.
37. Ganesh D, Service B, Zirgibel B, Koval K. The Detection of Prominent Hardware in Volar Locked Plating of Distal Radius Fractures: Intraoperative Fluoroscopy Versus Computed Tomography. *J Orthop Trauma.* 2016;30(11):618-621.
38. Haug LC, Glodny B, Deml C, Lutz M, Attal R. A new radiological method to detect dorsally penetrating screws when using volar locking plates in distal radial fractures. The dorsal horizon view. *Bone Joint J.* 2013;95-b(8):1101-1105.

39. Hufner T, Stubig T, Gosling T, Kendoff D, Geerling J, Krettek C. [Cost-benefit analysis of intraoperative 3D imaging]. *Der Unfallchirurg.* 2007;110(1):14-21.
40. Richter M, Geerling J, Zech S, Goesling T, Krettek C. Intraoperative three-dimensional imaging with a motorized mobile C-arm (SIREMOBIL ISO-C-3D) in foot and ankle trauma care: a preliminary report. *J Orthop Trauma.* 2005;19(4):259-266.
41. Wich M, Spranger N, Ekkernkamp A. [Intraoperative imaging with the ISO C(3D)]. *Der Chirurg; Zeitschrift fur alle Gebiete der operativen Medizen.* 2004;75(10):982-987.
42. Selles CA, Beerekamp MSH, Leenhouts PA, et al. The Value of Intraoperative 3-Dimensional Fluoroscopy in the Treatment of Distal Radius Fractures: A Randomized Clinical Trial. *J Hand Surg Am.* 2020;45(3):189-195.
43. Mehling I, Rittstieg P, Mehling AP, Kuchle R, Muller LP, Rommens PM. Intraoperative C-arm CT imaging in angular stable plate osteosynthesis of distal radius fractures. *J Hand Surg Eur Vol.* 2013;38(7):751-757.

PART I

Risk Stratification in the Emergency Department

CHAPTER 2

Machine Learning to Estimate the Probability for Fracture of the Distal Radius in Patients Presenting to the Emergency Department with Sustained Wrist Trauma

D.W.G. Langerhuizen
L.A.M. Hendrickx
G.M.M.J. Kerkhoffs
R.L. Jaarsma
J.N. Doornberg
M.M.J. Walenkamp
M.A.M. Mulders
J.C. Goslings
N.W.L. Schep

Submitted

ABSTRACT

Objectives

Only one third of patients presenting to the emergency department (ED) with wrist pain following trauma have a fracture of the distal radius. However, the majority are referred for radiographic evaluation. Artificial-intelligence derived algorithms may simplify existing prediction tools for risk stratification as only the most relevant variables are incorporated, thereby enhancing ease of utilization in clinical practice. The primary aim was to develop and externally validate four machine learning algorithms to predict the probability of a fracture of the distal radius for patients presenting to the ED.

Methods

We included 854 patients who were prospectively enrolled at five hospitals EDs; 488 patients in the derivation cohort and 366 in the validation cohort. Missing data were imputed using the missForest method. Among nineteen clinical predictors, random forest algorithm identified four variables most influential: age, swelling of the wrist, visible deformation, and distal radius tender to palpation. Four ML-algorithms were developed on the derivation cohort: boosted decision tree, support vector machine, neural network and Bayes point machine. Each algorithm's performance for selection of patients with a suspected distal radius fracture in the validation cohort was assessed according to the following metrics: (1) c-discrimation; (2) calibration; and (3) Brier-score.

Results

All models showed nearly similar performance: c-statistics ranged between 0.86 and 0.88, while the Brier scores was 0.16 for all models. Calibration slopes ranged between 0.72 and 0.84, while calibration intercepts ranged between -0.05 and -0.21. Bayes point machine was the best-fit algorithm. At a threshold of 0.05, the sensitivity and specificity were 0.98 and 0.24 respectively. We incorporated Bayes point machine into an open access web-based application (accessible: https://traumaplatform.shinyapps.io/distal-radius_ed).

Conclusion

We developed an online decision tool that can accurately predict the probability of a fracture of the distal radius after injury to the wrist. Clinicians could use the generated low and high probabilities to identify distal radius fractures, while using an intermediate probability to decide whether further radiographic evaluation is needed.

INTRODUCTION

Only one third of patients presenting to the emergency department (ED) after sustaining wrist trauma have a fracture of the wrist; however, the majority of patients will undergo radiographic evaluation.[1-3] This may lead to higher direct medical costs and prolongation of ED waiting time. For patients presenting to the ED with ankle trauma, reliable criteria have previously been developed to determine which patients need additional radiographic evaluation.[4] In addition, a recently developed and implemented clinical decision rule for wrist trauma (Amsterdam Wrist Rules) showed a safe reduction of wrist radiographs in patients suspected of having a distal radius fracture.[5]

More recently, Horng et al. developed machine learning decision supportive tools to automatically detect patients suspected of having a sepsis.[6] It is expected that complex and rapid accumulation of datasets offer unprecedented opportunities to apply artificial-intelligence (AI) algorithms--probability estimators that can iteratively learn to derive rules and patterns from data--to develop individualized prediction tools to optimally enhance shared decision-making.

Although criteria have been developed and implemented that indicate which patients suspected of having a fracture of the wrist should be referred for radiographic evaluation, there may be additional value in re-using the data for AI algorithms as they have the potential to identify potential non-linear associations.[5,7,8] Also, AI-derived algorithms may ease the utilization in clinical practice by only incorporating the most relevant variables. As such, machine-learning (ML) algorithms may simplify already existing prediction tools by incorporating less variables, while simultaneously maintaining the diagnostic performance and improving stewardship of resources when these tools are shifted towards earlier use in the clinical workflow (i.e. from physician to triage nurse).

Therefore, we aimed 1) to develop and externally validate four ML-algorithms to predict the probability of a fracture of the distal radius for patients presenting to the ED after acute wrist trauma; 2) to evaluate the diagnostic accuracy of the best-fit machine learning algorithm; and 3) to deploy the best-fit algorithms as an open-access freely available application for clinical use.

PATIENTS AND METHODS

We secondarily used prospectively collected data of 854 patients that were included between November 11, 2010 and June 25, 2014 at five hospitals EDs (one academic and four regional hospitals).[7] In the index study, all consecutive adult patients presenting to the ED with pain or tenderness after sustaining wrist trauma were included with the intent to develop criteria that indicate which patient should be referred for additional

radiographic evaluation. Upon presentation, attending physicians collected nineteen clinical predictors (including patient demographics, physical examination, and functional testing) prior to potential radiographic evaluation (Table 1).

The derivation cohort consisted 488 patients enrolled in the academic hospital, while the validation cohort--having similar demographics (e.g. age and sex)--consisted 366 patients enrolled in the four regional hospitals.[7]

For the purpose of this study, we used the presence or absence of a fracture of the distal radius as the outcome of interest. This was assessed on radiographic evaluation by the attending radiologist (i.e. reference standard for machine learning algorithms).

This study was conducted according to the Reporting Machine Learning Models in Biomedical Research and transparent reporting of a multivariable prediction model for individual prognosis or diagnosis statement.[9,10]

Table 1. Clinical predictors in both cohorts evaluated at the emergency department

	Missing (*n*, %)
Age	
Sex	
Swelling wrist	2 (0.2%)
Swelling anatomical snuffbox	5 (0.6%)
Visible deformation	11 (1.3%)
Tenderness distal radius to palpation	3 (0.4%)
Tenderness ulna to palpation	9 (1.1%)
Tenderness anatomical snuffbox to palpation	4 (0.5%)
Tenderness scaphoid tubercle to palpation	41 (4.8%)
Active mobility painful	
Dorsiflexion	5 (0.6%)
Palmar flexion	10 (1.2%)
Supination	6 (0.7%)
Ulnar deviation	8 (0.9%)
Radial deviation	11 (1.3%)
Functional tests painful	
Radioulnar ballottement test	33 (3.9%)
Axial compression forearm	28 (3.3%)
Axial compression thumb	44 (5.2%)
Pinch grip test	64 (7.5%)

Statistical analysis

We present categorical variables as frequencies and percentages and continuous variables with median and interquartile range (IQR).

Missing data for all variables was low. We applied the missForest method to impute missing values, which can be used for both continuous and categorical data.[11] Using data from the derivation cohort, random forest algorithms were utilized to determine clinical variable hierarchy and to eventually reduce the number of clinical predictors incorporated in the final algorithms.[12] As such, we included the following four predictors: age, swelling of the wrist, visible wrist deformation, and distal radius tender to palpation.

Based on prior work for binary classification tasks[13-17], we selected the following four supervised machine learning algorithms: (1) boosted decision tree; (2) support vector machine; (3) neural network; and (4) Bayes point machine (Figure 1). We used the derivation cohort as a dataset for training of each algorithm. Subsequently, we tested the performance on the validation cohort to predict the probability of a patient having any wrist fracture as well as having a fracture of the distal radius only.

Performance of each algorithm was assessed on the validation cohort according to the following metrics: (1) discrimination (C-statistic); (2) calibration; and (3) overall performance score (i.e. Brier-score).[18] A C-statistic (also known as area under the curve) of 1.0 indicates perfect discrimination, whereas 0.5 indicates discrimination is no better than a flip-of-a-coin.[19] Calibration of the models was evaluated by plotting calibration curves. The slope and intercept of these calibration plots are indicative of the agreement between the predicted probability of an outcome and the observed outcome. A calibration slope of 1 and a calibration intercept of 0 indicate perfect agreement, whereas smaller or larger slopes or intercepts indicate predictions are too extreme, too high, not extreme enough, or too low. The Brier score was calculated as the mean of the squared differences between the actual outcomes and predicted probabilities. A Brier score of 0 indicates perfect prediction, whereas 0.25 indicates an uninformative algorithm.

According to the initial Amsterdam Wrist Rules, we set the diagnostic cut-off point of our best performing algorithm at a value that would maintain a sensitivity of 98%, however, at the cost of a limited specificity.[7] Sensitivity reflects to the proportion of positives that are correctly determined as such, whereas specificity applies to the negatives that are correctly identified. Using bootstrapping (number of resamples: 10,000), we calculated 95% confidence intervals. The negative predictive value indicates the proportion of true negatives among all negatively tested patients.

For data analysis and algorithm creation, we used Stata 15.0 (StataCorp LP, College Station, TX), Microsoft Azure (Redmond, WA, USA), and RStudio (Version 1.1.463; Boston, MA, USA) with the packages CalibrationCurves and ggplot2.

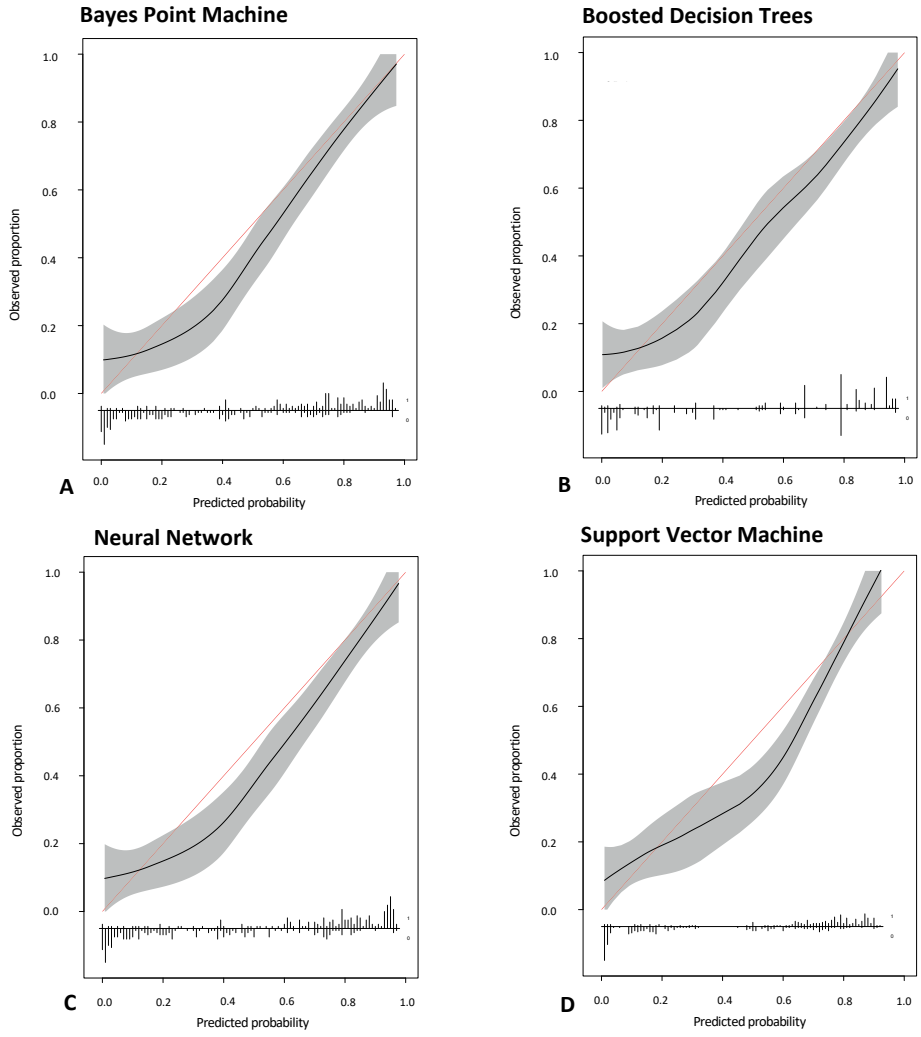


Figure 1: This figure depicts the calibration plots including 95% CI (in grey) for algorithms developed to determine the probability for fracture of the distal radius.

Patient Characteristics

Of 488 patients in the derivation cohort, median age at diagnosis was 48 years (interquartile range [IQR] 29-61 years); and 211 patients were men (43%). A fracture of the distal radius was present in 204 patients (42%).

In the validation cohort, the median age of 366 patients was 52 years (IQR 34-69 years); and 124 patients were men (34%). In 172 patients (47%), a fracture of the distal radius was detected.

RESULTS

Performance for Machine Learning Algorithms

All algorithms showed nearly similar performance (Table 2). C-statics ranged between 0.85 and 0.86 (Figure 2). Calibration slopes ranged from 0.72 to 0.84 and calibration intercepts ranged from -0.05 to -0.21. The overall algorithm performance as assessed by the Brier scores was 0.16.

Table 2. Performance for machine learning models for radiography use in distal radius fracture patients

Method	Performance Measure	Bayes Point Machine	Boosted Decision Trees	Neural Network	Support Vector Machine
Discrimination	C-statistic	0.86	0.86	0.86	0.85
Calibration	Slope	0.84	0.72	0.8	0.81
	Intercept	-0.13	-0.11	-0.21	-0.05
Overall	Brier score	0.16	0.16	0.16	0.16

A C-statistic (also known as area under the curve [AUC]) of 1.0 indicates perfect discrimination, whereas 0.5 indicates discrimination is no better than a flip-of-a-coin. The slope and intercept of calibration plots are indicative of the agreement between the predicted probability of an outcome and the observed outcome. A calibration slope of 1 and a calibration intercept of 0 indicate perfect agreement. Smaller or larger slopes or intercepts indicate predictions are too extreme, too high, not extreme enough, or too low. A Brier score of 0 indicates perfect prediction, whereas 0.25 indicates an uninformative algorithm.

Diagnostic Performance Characteristics

The best performing algorithm was Bayes Point Machine (c-statistic: 0.86, calibration slope: 0.84, calibration intercept: -0.13, and Brier score: 0.16). At a threshold of 0.05, the sensitivity of the algorithm was 98% (95% CI, 95-100) and the specificity was 24% (95% CI 18-30). The number of true positive, false positive, true negative and false negative cases were 168, 147, 47 and 4 respectively.

Application Development

To allow users to calculate the probability of having a distal radius fracture, the Bayes Point Machine was incorporated into an online application: https://traumaplatform.shinyapps.io/distalradius_ed

DISCUSSION

Data-driven predictive analytics--commonly referred to as AI or ML--can be used to develop decision supportive tools that calculate patient-tailored probabilities of various outcomes of interest. In the clinical scenario of a patient presenting to the ED following wrist trauma, we found that machine learning algorithms can accurately determine

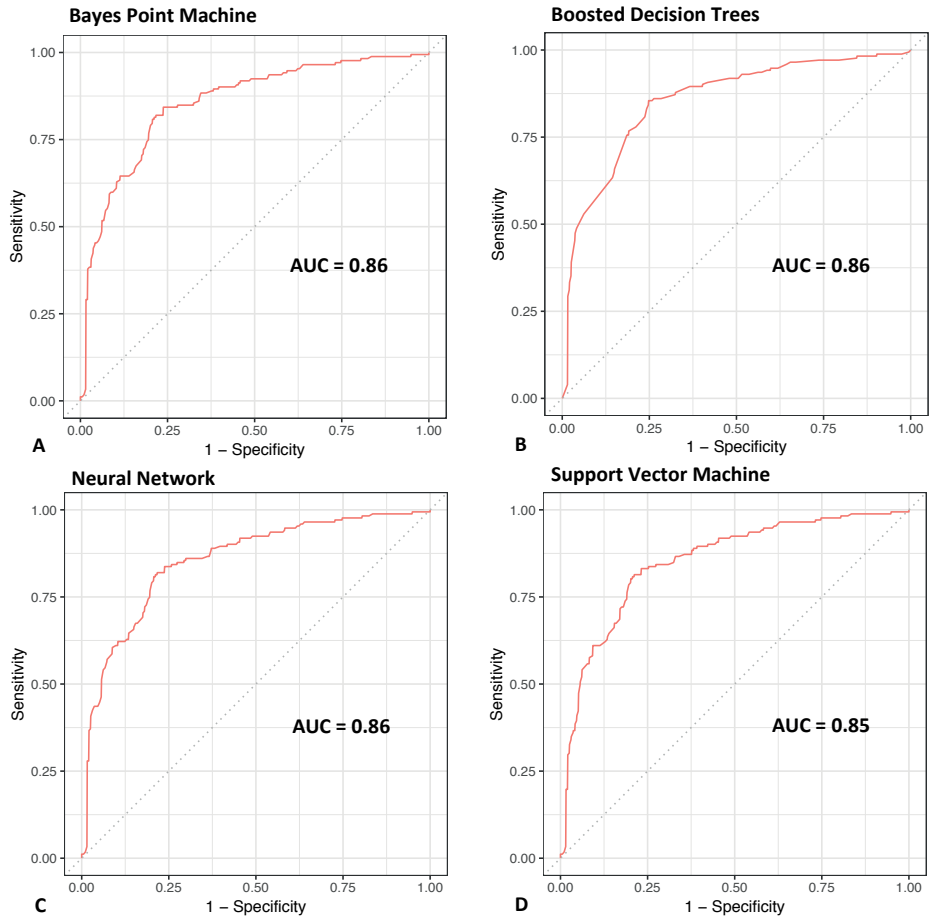


Figure 2: Receiver operating curves for (A) Bayes point machine, (B) Boosted decision trees, (C) Neural network, and (D) Support vector machine to determine the probability for having a fracture of the distal radius.

which patients suspected of having a fracture of the distal radius should be referred for additional radiographic evaluation.

This study has several limitations. First, we accounted for missing data by applying the missForest method, an accurate and robust technique that can handle mixed-type data (i.e. categorical and continuous variables).[11] This may have led to biased analyses. Given that missing data for all variables was low, we only regard this as a minor limitation. Second, as already addressed in the initial study, clinical variables were only assessed by one attending clinician at the ED. Consistency of the clinical variables was not determined, as it was regarded unethical to comprehensively examine patients twice. Third, given that the dataset of the initial study mainly consists distal radius fractures, we decided not to develop machine learning models for all wrist fractures (e.g. carpal or

distal ulna fractures). A dataset---encompassing more ulnar and carpal fractures--might be more suitable in the future for clinical purposes.

We found that Bayes Point machine showed best performance. It may go without saying that the marginally observed differences in performance are due to varying used algorithms. Based on the assumptions that some models work best for a specific data set, but may not hold in other, we decided to test four commonly used algorithms for binary classification tasks.[20] As so, we intended to only focus on algorithms that are most likely useful a-priori.[16] Our best performing algorithm showed similar discriminating capabilities as well as sensitivity and specificity compared to a previously developed, externally validated, and implemented clinical decision rule based on a logistic regression model.[5,7] For ease of utilization in clinical practice, we carefully considered to only incorporate the most relevant variables while maintaining the diagnostic performance. Except for tenderness of the distal radius, we incorporated two variables that are visually assessable (swelling of the wrist and visible deformation), while age is already part of the interview. Although other variables are also potentially associated with a fracture of the distal radius, our data analysis demonstrated that they may not add incremental value to the four variables we incorporated in our model. Instead, the initially developed clinical decision rule for a suspected fracture of the distal radius incorporated eight variables, of which four are considered burdensome for the patient (e.g. pain on radioulnar ballottement test).[7] We speculate our algorithm might potentially improve stewardship of resources when shifted towards earlier use in the clinical workflow from physician to triage nurse. However, our model has not been implemented in clinical practice yet, while the AWR has proven to safely reduce the number of requested radiographs at the ED.[5]

At a threshold of 0.05 for our model, true positive, false positive, true negative and false negative cases were 168, 147, 47, and 4, respectively. In line with prior studies, this indicates that 315 out of 366 patients would have been referred for additional radiographic imaging instead of 100%, reducing the need for further radiographic evaluation with 14%.[5,7] Today, our model can only estimate the probability of having a distal radius fracture following trauma, but cannot discern other wrist fractures. This may explain why 7 patients with an ipsilateral carpal fracture amongst 47 true negatives are missed.

Nevertheless, we decided to deploy Bayes point machine as an open-source web-based prediction tool because it may still give clinicians a useful insight about the specific risk of a distal radius for their individual patients. This tool is a first step in the development of prediction tools with implemented feedback loops enabling continuous data collection to improve the probable outcome. It is important to note that the end-user should always consider that the complex statistical back-end model is not intuitive and medicolegal regulation has yet to be established.[21] We believe that the potential

clinical value may lie in using the prediction tool in conjunction with ML-algorithms trained in predicting carpal fractures.

In conclusion, we developed a decision supportive tool with only four clinical variables that can reliably predict the probability of having a distal radius fracture after sustaining wrist trauma. Clinicians could use the generated low and high probabilities to identify distal radius fractures, while using an intermediate probability to decide whether further radiographic evaluation is needed. However, our decision tool is not able to accurately detect concomitant ipsilateral fractures simultaneously (e.g. carpal or ulnar styloid process fractures). Further research should evaluate whether its predictive capability will hold in practice as well as developing an overarching all wrist fracture decision supportive tool.

REFERENCES

1. Gleadhill DN, Thomson JY, Simms P. Can more efficient use be made of x ray examinations in the accident and emergency department? *Br Med J (Clin Res Ed).* 1987;294(6577):943-947.
2. Radiography of injured arms and legs in eight accident and emergency units in England and Wales. Royal College of Radiologists Working Party. *Br Med J (Clin Res Ed).* 1985;291(6505):1325-1328.
3. van den Brand CL, van Leerdam RH, van Ufford JH, Rhemrev SJ. Is there a need for a clinical decision rule in blunt wrist trauma? *Injury.* 2013;44(11):1615-1619.
4. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. *Ann Emerg Med.* 1992;21(4):384-390.
5. Mulders MAM, Walenkamp MMJ, Sosef NL, et al. The Amsterdam Wrist Rules to reduce the need for radiography after a suspected distal radius fracture: an implementation study. *Eur J Trauma Emerg Surg.* 2019.
6. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PloS one.* 2017;12(4):e0174708.
7. Walenkamp MMJ, Bentohami A, Slaar A, et al. The Amsterdam wrist rules: the multicenter prospective derivation and external validation of a clinical decision rule for the use of radiography in acute wrist trauma. *BMC Musculoskelet Disord.* 2015;16(389).
8. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMA. Surgical Risk Is Not Linear: Derivation and Validation of a Novel, User-friendly, and Machine-learning-based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) Calculator. *Ann Surg.* 2018;268(4):574-583.
9. Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res.* 2016;18(12):e323.
10. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ (Clinical research ed).* 2015;350:g7594.
11. Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics.* 2012;28(1):112-118.
12. Zou RY, Schonlau M. The Random Forest Algorithm for Statistical Learning with Applications in Stata. *The Stata Journal.* 2018.
13. Karhade AV, Thio Q, Ogink PT, et al. Development of Machine Learning Algorithms for Prediction of 30-Day Mortality After Surgery for Spinal Metastasis. *Neurosurgery.* 2018.
14. Staartjes VE, de Wispelaere MP, Vandertop WP, Schroder ML. Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar diskectomy: feasibility of center-specific modeling. *Spine J.* 2018.
15. Thio Q, Karhade AV, Ogink PT, et al. Can Machine-learning Techniques Be Used for 5-year Survival Prediction of Patients With Chondrosarcoma? *Clin Orthop Relat Res.* 2018;476(10):2040-2048.
16. Wainer J. Comparison of 14 different families of classification algorithms on 115 binary dataset. *Available at:* http://arxivorg/abs/160600930 *Accessed July 13, 2018.* 2016.
17. Karhade AV, Ogink PT, Thio Q, et al. Development of machine learning algorithms for prediction of prolonged opioid prescription after surgery for lumbar disc herniation. *Spine J.* 2019;19(11):1764-1771.

18. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. *New York, NY: Springer.* 2009.
19. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters.* 2006;27:861-874.
20. Wolpert DH. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation.* 1996;8(7):1341-1390.
21. Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation". Available at: https://arxiv.org/abs/1606.08813. Accessed June 28, 2016.

PART II

Deep Learning for Fracture Detection

CHAPTER 3

What Are the Applications and Limitations of Artificial Intelligence for Fracture Detection and Classification in Orthopaedic Trauma Imaging? A Systematic Review

D.W.G. Langerhuizen
S.J. Janssen
W.H. Mallee
M.P.J. van den Bekerom
D. Ring
G.M.M.J. Kerkhoffs
R.L. Jaarsma
J.N. Doornberg

Clinical Orthopaedics & Related Research 2019 Nov;477(11):2482-2491.

Commentary by:
J.D. Michelson in Clinical Orthopaedics & Related Research 2019 Nov; 477(11):2492-2494.

ABSTRACT

Background

Artificial-intelligence algorithms derive rules and patterns from large amounts of data to calculate the probabilities of various outcomes using new sets of similar data. In medicine, artificial intelligence (AI) has been applied primarily to image-recognition diagnostic tasks and evaluating the probabilities of particular outcomes after treatment. However, the performance and limitations of AI in the automated detection and classification of fractures has not been examined comprehensively.

Question/purposes

In this systematic review, we asked (1) What is the proportion of correctly detected or classified fractures and the area under the receiving operating characteristic (AUC) curve of AI fracture detection and classification models? (2) What is the performance of AI in this setting compared with the performance of human examiners?

Methods

The PubMed, Embase, and Cochrane databases were systematically searched from the start of each respective database until September 6, 2018, using terms related to "fracture", "artificial intelligence", and "detection, prediction, or evaluation." Of 1221 identified studies, we retained 10 studies: eight studies involved fracture detection (ankle, hand, hip, spine, wrist, and ulna), one addressed fracture classification (diaphyseal femur), and one addressed both fracture detection and classification (proximal humerus). We registered the review before data collection (PROSPERO: CRD42018110167) and used the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA). We reported the range of the accuracy and AUC for the performance of the predicted fracture detection and/or classification task. An AUC of 1.0 would indicate perfect prediction, whereas 0.5 would indicate a prediction is no better than a flip-of-a-coin. We conducted quality assessment using a seven-item checklist based on a modified methodologic index for non-randomized studies instrument (MINORS).

Results

For fracture detection, the AUC in five studies reflected near perfect prediction (range, 0.95-1.0), and the accuracy in seven studies ranged from 83% to 98%. For fracture classification, the AUC was 0.94 in one study, and the accuracy in two studies ranged from 77% to 90%. In two studies AI outperformed human examiners for detecting and classifying hip and proximal humerus fractures, and one study showed equivalent performance for detecting wrist, hand and ankle fractures.

Conclusions

Preliminary experience with fracture detection and classification using AI shows promising performance. AI may enhance processing and communicating probabilistic tasks in medicine, including orthopaedic surgery. At present, inadequate reference standard assignments to train and test AI is the biggest hurdle before integration into clinical workflow. The next step will be to apply AI to more challenging diagnostic and therapeutic scenarios when there is absence of certitude. Future studies should also seek to address legal regulation and better determine feasibility of implementation in clinical practice.

INTRODUCTION

In 1959, Arthur Samuel defined artificial intelligence (AI) as a field of study that gives a computer the ability to learn without needing to be reprogrammed.[1] In layman's terms, AI algorithms are developed to derive rules and patterns from large amounts of data to calculate the probabilities of various outcomes with new sets of similar data (Figure 1). For instance, Netflix uses AI algorithms to analyze the viewing preferences of millions of people and determine what a viewer is likely to enjoy based on prior viewing behavior. Computers are programmed to continuously update probabilities of a person liking a given television show based on a combination of new all-user data and individual viewing choices.

The initial applications of AI in medicine have focused largely on image-recognition diagnostic tasks such as detecting retinopathy in diabetic people via photographs of the retinal fundus, detecting mammographic lesions, and recognizing skin cancer.[2-4] AI algorithms that address treatment probabilities—such as decision-support tools to assist orthopaedic oncologists in predicting survival and mortality—have also been developed but are not yet widely used in clinical practice.[5,6]

AI might be useful to aid the diagnostic aspects of fracture care. For example, AI applications might improve the diagnosis of true fractures among suspected fractures of the scaphoid or hip, detect key fracture characteristics that might alter prognosis and treatment, or help detect less severe fractures that are often overlooked during a secondary evaluation after complex trauma.[7,8] The key applications of AI will help address the shortcomings of human intelligence that make us susceptible to the magician's sleight of hand and, likewise, to overlook important details in distracting circumstances. In clinical practice, both the routine and complex can be distractions.

We aggregated data from published studies using AI for fracture detection and classification to address the following questions: (1) What is the proportion of correctly detected or classified fractures and the area under the receiving operating characteristic (AUC) curve of AI fracture detection and classification models? (2) What is the performance of AI in this setting compared with the performance of human examiners?

MATERIALS AND METHODS

Article Selection, Quality Assessment, and Data Extraction

We performed a systematic search according to the PRISMA statement[9] using the PubMed, Embase, and Cochrane libraries for studies from the start of each respective database until September 6, 2018. Our review protocol was registered on PROSPERO (CRD42018110167) before data collection. A professional medical librarian helped us

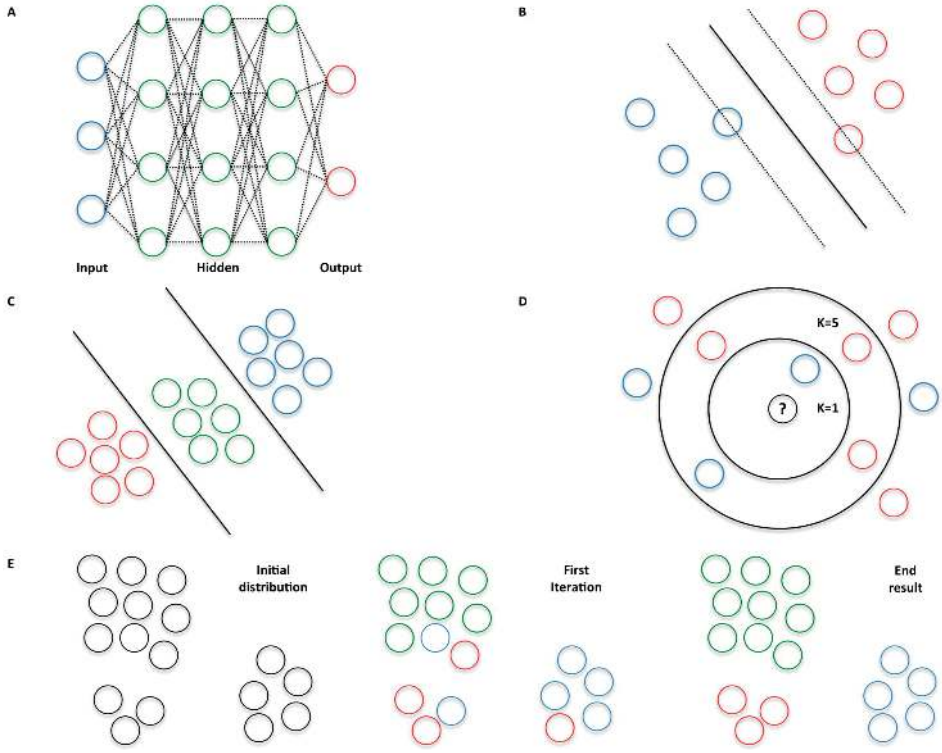


Figure 1: Two common AI techniques exist. Supervised learning applies to iteratively training of an algorithm with a dataset consisting of input features with ground truth labels. For example, wrist radiographs are provided as input features consisting the following labels; fracture versus no fracture. By providing new wrist radiographs without a label, the algorithm learns to make a prediction between both classes on its own. Unsupervised learning applies to data exposure without ground truth labels. During the training phase, the algorithm tries to find labels that best organize the data (i.e. 'clustering'). Generally, unsupervised learning requires more computational power, larger datasets, and its performance is more challenging to evaluate. Therefore, supervised algorithms are most commonly used in medical applications. **(A)** Neural networks are based on interconnected neurons in the human brain. The blue dots represent the input features, whereas the red dots are the output of the algorithm. The green dots mathematically weigh the input features to predict an output. **(B)** Support vector machine defines an optimal separating 'hyperplane' to maximize the distance from the closest points of two classes. **(C)** Linear discriminant analysis is a linear classification technique to distinguish between three or more classes. **(D)** K-nearest neighbours classify an input feature by a majority vote of its K-closest neighbours. For instance, the unknown dot will be assigned blue if K = 1 (inner circle), whereas the unknown dot will be assigned red if K = 5 (outer circle). **(E)** K-means groups objects based on their characteristics by iteratively aggregating clusters to centroids by minimizing the distance to the middle point of the cluster. For example, three clusters are aggregated (i.e. K = 3); green-, red-, and blue dots.

build the search syntax using the following keywords in the title and abstract: (orthopedics OR orthopedic procedures OR traumatology OR fracture* OR skeletal fixation* OR (trauma* AND orthop*)) AND (artificial intelligence OR neural network* or deep learning

OR machine learning OR machine intelligence) AND (predict* OR predictive value of test OR score OR scores OR scoring system OR scoring systems OR observ* OR observer variation OR detect* or evaluat* OR analy* OR assess* OR measure*) (see Appendix; Supplemental Digital Content 1, http://links.lww. com/CORR/A194).

Two reviewers (DWGL, SJJ) independently screened the titles and abstracts, and if a study was considered eligible, they together screened the full-text article using predefined criteria to reach agreement. A third reviewer was not deemed necessary as a high level of consistency during the screening and inclusion process was achieved. Articles met the inclusion criteria if they addressed one or more AI models (a mathematical computing algorithm trained with "big data" to autonomously assign labels to unseen data) for detecting and/or classifying fractures on any radiologic imaging modality. We did not restrict the radiologic imaging modality to detect and/or classify fractures. We excluded studies in which patients were not in an orthopaedic trauma setting, studies evaluating robot-assisted surgery techniques, studies with mixed cohorts without clear subgroup reporting, review articles, letters to the editor, conference abstracts, technique papers, animal and cadaveric studies, and studies not published in English.

The database search yielded 1221 citations, and after removing duplicate articles, we screened 1044 potentially eligible records (Figure 2). Twenty-eight studies were selected for full-text screening, of which eight remained. However, two additional eligible studies were identified through verbal communication in our network and meeting proceedings,

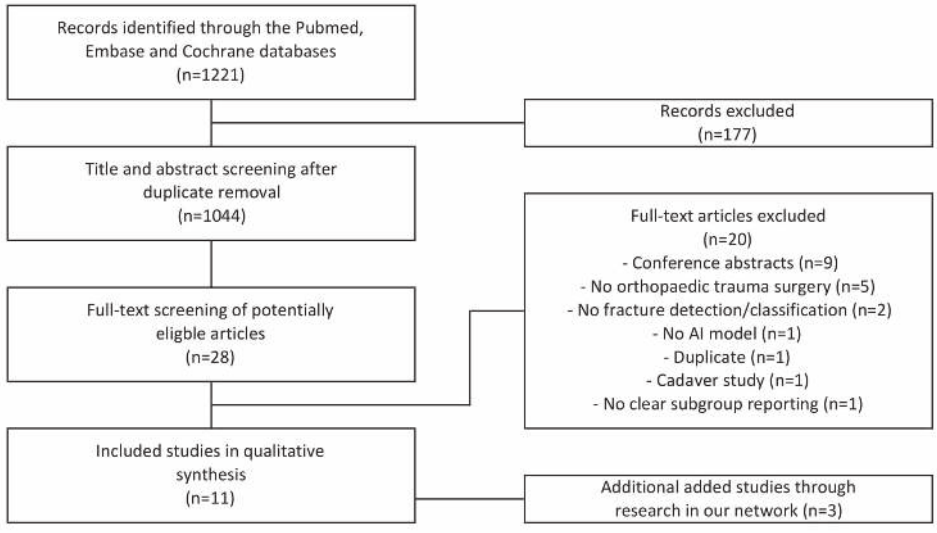


Figure 2: This flowchart depicts the study selection during screening and inclusion of articles for a search period from start of each initial database to September 6, 2018.

but did not appear in our structured systematic searches.[10,11] We did not identify new eligible studies through screening the reference lists of the included studies.

Two reviewers (DWGL, SJJ) independently appraised the quality of all included studies. The Newcastle-Ottawa Scale and methodologic index for nonrandomized studies (MINORS) instruments are commonly used for cohort or case-control studies.[12,13] However, there is no risk of bias assessment tool that is suitable for diagnostic studies. Therefore, we decided to conduct quality assessment using a modified seven-item checklist based on the MINORS criteria, including disclosure, study aim, input feature, determination of ground truth labels, dataset distribution, performance metric, and explanation of the used AI model. Standardized forms were used to extract and record data using an electronic database (Microsoft Excel Version 16.21; Microsoft Inc, Redmond, WA, USA). A consensus meeting between both observers (DWGL, SJJ) was held to overcome disagreements regarding article selection, quality assessment, and data extraction.

Outcome Measures

Our primary study outcome was the proportion of correctly detected or classified fractures and nonfractures to the total number of patients and the area under the receiving operating characteristic (AUC) curve of AI models. A total of 10 studies met inclusion criteria and were used to answer this research question. Our secondary outcome was the performance of AI in this setting compared with the performance of human examiners. Three studies met inclusion criteria and were used to answer this research question.

The following data were obtained from each study: year of publication, input feature (radiologic imaging modality), projection when plain radiography was used as a radiologic imaging modality (for example, AP, oblique, or lateral views),[10,11,14-19] size of the dataset, anatomic location, output classes, AI models that were used, pretrained convolutional neural network (CNN), if applicable, size of the training set, size of the validation set or method, size of the test set, and performance measures (accuracy and AUC curve).

Output classes included fracture detection and/or classification. We considered fracture detection as a binary classifier with two inherent output classes (the presence of any fracture versus absence of a fracture). From what we could discern, these studies evaluated any type of fracture: both displaced fractures, which are easy to detect, and nondisplaced fractures, which can be subtler. Additionally, fracture classification addressed multiple output classes. For example, one study addressed a four-group classification system to distinguish among types of proximal humerus fractures (that is, the Neer classification[20]),[16] whereas another study addressed a subtype of femur fractures (AO-Type 32[21]: a nine-group classification method for diaphyseal femur fractures ranging from simple spiral fractures to complex, irregular, comminuted fractures[15]).

Six studies described the use of a single AI model for detecting and/or classifying the fracture,[10,11,16,17,19,22] and four compared the performance of more than one model.[14,15,18,23]

We analyzed studies describing pretrained CNNs (AI models that were developed using large, separate datasets such as the ImageNet Large Scale Visual Recognition Challenge[24])[10,11,16-19] that were subsequently transferred to new datasets and AI models trained from scratch and implemented for new and unseen data.

Generally, two validation techniques are used to evaluate an AI model after the training phase: a subset of the dataset is retained as a validation set (that is, the size of the validation set) or a validation method is applied. The goal of using a validation set or validation method—especially in situations with small datasets—is to increase model robustness (for example, developing strategies to cope with errors during performance of a specific task). For example, k-fold cross-validation is a validation method that is applied to an automated computer-generated resampling procedure, in which a dataset is divided into smaller sets of different combinations (multiple folds or partitions), which allows it to train throughout many iterations. Although not mutually exclusive, each fold is iteratively used as a test set and the rest is used for training. The size of the test set is a partition of the dataset used for the final evaluation and determines the performance measures of the AI model.

The accuracy and AUC were assessed to provide information on each AI model in the test dataset because these were the most commonly addressed items (eight studies addressed accuracy[10,14-19,23] and five studies addressed the AUC[10,11,16,17,19].) In our study, accuracy applied to the proportion of correctly detected or classified fractures and nonfractures to the total number of patients (such as the proportion of correct predictions over all cases). The AUC corresponds to the probability that a binary classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.[25] An AUC of 1.0 would indicate perfect prediction, whereas 0.5 would indicate a prediction is no better than chance.

Distribution of Fracture Detection and Classification, Anatomical Location, Used AI Models, and Input Features

Nine studies addressed AI models for detecting fractures,[10,11,14,16-19,22,23] whereas one study addressed fracture classification.[15] Chung et al.[16] were the only authors to report on both a fracture detection and fracture classification task (Table 1).

Anatomic fractures were located in the wrist,[11,16-18] hip,[10,19] spine,[22,23] ankle,[18] diaphyseal femur,[15] hand,[18] and proximal humerus.[16]

A pretrained CNN was the most frequently used AI model,[10,11,16-19] followed by neural networks,[14,15,23] k-nearest neighbors,[14,15] support vector machines,[15,22] K-means,[23] and linear discriminant analysis.[15] All AI models were supervised, except for the K-means, which is an unsupervised AI model.

Input features used in the AI models were as follows: eight studies used radiography as an imaging modality,[10,11,14-19] whereas two studies used CT.[22,23] When radiography was the radiologic imaging modality, AP[10,16,18,19] and lateral[11,16-18] projections were most commonly used, followed by posteroanterior,[11] oblique (two different types),[18] and scaphoid (four specific scaphoid views: proximal, distal, ulnar, and radial).[18]

Quality Appraisal

Ten studies were included. Quality appraisal demonstrated that the study aim was clear in seven studies (70%), possibly resulting in outcome bias for the remaining three studies (Figure 3). In seven studies (70%), the inclusion and exclusion criteria for input features (all eligible radiographs and CT scans were included in the dataset) were clearly described, whereas selection bias could not be excluded for the remaining two studies (30%). Seven studies (70%) clearly described how they determined the ground truth (the reference standards in AI), subjecting the remainder to poorly trained AI models. All studies reported a clear distribution of the dataset (training, validation, and testing phases), described how the performance of an AI model was determined (accuracy and AUC), and clearly explained the AI model that was used (see Appendix 2; Supplemental Digital Content 2, http://links.lww.com/CORR/A195).

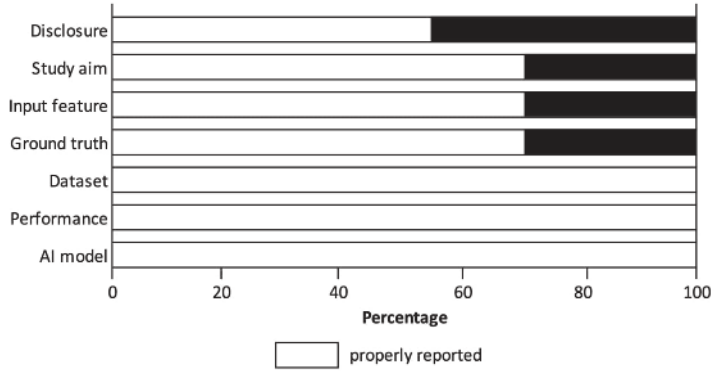


Figure 3: We conducted a quality assessment of included studies using a seven-item checklist based on a modified methodologic index for nonrandomized studies (MINORS) instrument.

Statistical Analysis

Given the heterogeneity of the studies, we reported the range for accuracy and AUC for fracture detection and classification tasks. The sizes of the training, validation, and test sets are reported as percentages of the total number of the dataset.

There was no funding received to perform this work.

Table 1: Studies evaluating the use of AI models in bone fracture detection and/or classification

Author, year	Input features	Imaging direction	Size dataset	Anatomical location	Ground truth label assignment#	Number of output classes	AI models used	Pre-trained CNN	Size training set‡	Size validation set¶ / validation method	Size test set	Performance (accuracy and AUC)
					Fracture detection							
Al-helo et al.[22]	CT	NA	50	Spine	NA	2	K-means	NA	100	NA	NA	98 / NA
Al-helo et al. [22]	CT	NA	50	Spine	NA	2	NN	NA	90	5-LOCV	10	93 / NA
Basha et al. [14]	X-ray	NA	180	NA	NA	2	NN	NA	78	NA	22	88 / NA
Basha et al. [14]	X-ray	NA	180	NA	NA	2	kNN	NA	78	10-FCV	22	86 / NA
Chung et al. [16]	X-ray	AP	1891	Proximal humerus	Combined approach*	2	CNN	ResNet-152	90	10-FCV	10	96 / 1.0
Gale et al. [10]	X-ray	AP	53279	Hip	Radiology reports	2	CNN	DenseNet	85.4	8.3	6.3	97 / 0.99
Kim et al. [17]	X-ray	Lat	1389	Distal radius and ulna	Radiology registrar†	2	CNN	Inception v3	80	10	10	89 / 0.95
Lindsey et al. [11]	X-ray	PA, lat	34990	Wrist	Orthopaedic surgeon	2	CNN	U-Net	80	10	10	NA / 0.97
Olczak et al. [18]	X-ray	AP, lat, obl, scaph	256458	Wrist, hand, ankle	Radiology reports	2	CNN	VGG_16	70	20	10	83 / NA
Olczak et al. [18]	X-ray	AP, lat, obl, scaph	256458	Wrist, hand, ankle	Radiology reports	2	CNN	VGG_19	70	20	10	NA
Olczak et al. [18]	X-ray	AP, lat, obl, scaph	256458	Wrist, hand, ankle	Radiology reports	2	CNN	Network-in-network	70	20	10	NA
Olczak et al. [18]	X-ray	AP, lat, obl, scaph	256458	Wrist, hand, ankle	Radiology reports	2	CNN	VGG CNN S network	70	20	10	NA
Olczak et al. [18]	X-ray	AP, lat, obl, scaph	256458	Wrist, hand, ankle	Radiology reports	2	CNN	BVLC Reference CaffeNet	70	20	10	NA
Urakawa et al. [19]	X-ray	AP	3346	Intertochanteric hip	Single radiologist	2	CNN	VGG_16	80	10	10	96 / 0.98
Yao et al. [21]	CT	NA	40	Thoracic and lumbar spine	Radiology reports	2	SVM	NA	50	10-FCV	50	NA

Table 1: Studies evaluating the use of AI models in bone fracture detection and/or classification (continued)

Author, year	Input features	Imaging direction	Size dataset	Anatomical location	Ground truth label assignment#	Number of output classes	AI models used	Pre-trained CNN	Size training set‡	Size validation set¶ / validation method	Size test set	Performance (accuracy and AUC)
					Fracture classification							
Bayram and Çakiroglu [15]	X-ray	NA	196	Diaphyseal femur	NA	9	LibSVM	NA	100	10-FCV	NA	90 / NA
Bayram and Çakiroglu [15]	X-ray	NA	196	Diaphyseal femur	NA	9	kNN	NA	100	10-FCV	NA	87 / NA
Bayram and Çakiroglu [15]	X-ray	NA	196	Diaphyseal femur	NA	9	NN	NA	100	10-FCV	NA	85 / NA
Bayram and Çakiroglu [15]	X-ray	NA	196	Diaphyseal femur	NA	9	LDA	NA	100	10-FCV	NA	83 / NA
Chung et al. [16]	X-ray	AP	1376	Proximal humerus	Combined approach*	4	CNN	ResNet-152	90	10-FCV	10	77 / 0.94

CNN = convolutional neural network, AUC = area under the receiver operating characteristic curve, ct = computed tomography, K-means = K refers to number of clusters generated by AI model, NN = neural network, 5-LOCV = leave-out cross-validation (10 rounds, five cases were wittheld every round), kNN = k-nearest neighbor, FCV = fold cross-validation, ap = antero-posterior, lat = lateral, front = frontal, obl = oblique (2 different types), scaph = 4 specific scaphoid views (proximal, distal, ulnar, and radial), pa = postero-anterior, 20° = 20° degrees tilted lateral wrist view, SVM = support vector machine, LibSVM = library for support vector machine, LDA = linear discrimant analysis, # the reference standard in AI, ‡ percentage of the total amount of the dataset, ¶ percentage of the total amount included radiographs / ct-scans, † radiology registrar with 3 years of experience, * two shoulder orthopaedic surgeons and one radiologist assigned the ground truth labels with additional use of the corresponding CT-scan if consensus could not be achieved

RESULTS

AI Model Performance

Among the five studies using AUC for fracture detection AI had near perfect prediction (range, 0.95-1.0).[10,11,16,17,19] The accuracy of fracture detection reported in seven studies ranged from 83% to 98%.[10,14,16-19,23]

Seven studies addressed fracture detection on radiographs,[10,11,14,16-19] and two studies addressed fracture detection on CT.[22,23]

In studies addressing fracture classification on radiographs, Chung et al.[16] found an AUC of 0.94 and an accuracy of 77% for classifying proximal humerus fractures into four groups (according to the Neer classification[20]). Bayram and Çakiroglu[15] applied four AI models for classification of diaphyseal femur fractures into nine groups (AO-type 32[21]) and found an accuracy ranging from 83% to 90%.

AI Models Compared with Humans

Three studies compared the performance of AI models with the performance of humans.[16,18,19] Urakawa et al.[19] used an AI model (that is, a pretrained CNN: VGG_16) for detecting hip fractures on an AP radiograph, which had a better AUC than five orthopaedic surgeons did (pretrained CNN: 0.98 [95% CI, 0.97-1.0] versus the five orthopaedic surgeons: 0.97 [95% CI, 0.95-0.97]; $p < 0.001$). Additionally, the difference in accuracy also favored the AI model (pretrained CNN: 96% [95% CI, 93-98] versus the five orthopaedic surgeons: 92% [95% CI, 89-95]; $p < 0.001$).

In a study by Olczak et al.,[18] the accuracy of the best- performing AI model (a pretrained CNN: VGG_16) in detecting wrist, hand, and ankle fractures on several radiographic projections was equivalent to that of two senior orthopaedic surgeons (pretrained CNN: 83 [95% CI, 80-87 versus 82 [95% CI, 78-86] and 82 [95% CI, 78-85] for the two senior orthopaedic surgeons).

For detecting fracture, Chung et al.[16] used a pretrained CNN (Microsoft ResNet-152; Redmond, WA, USA) to detect proximal humerus fractures on an AP radiograph and compared the accuracy of the CNN with that of three human groups: general physicians (n = 28), general orthopaedists (n = 11), and an orthopaedist who specialized in the shoulder (n = 19). The accuracy of the AI model was superior to that of the human groups, although there was no statistical difference between the AI model and the general orthopaedist and shoulder orthopaedist groups (pretrained CNN: 96% [95% CI, 94-97] versus 85% [95% CI, 80-90] for the general physicians, 93% [95% CI, 90-96] for the general orthopaedists, and 93% [95% CI, 87-99] for the orthopaedists who specialized in the shoulder; $p < 0.001$). Additionally, except for one subset (greater tuberosity fractures), the pretrained CNN also demonstrated better accuracy for classifying proximal humerus fractures into four groups (according to the Neer classification[20]).

DISCUSSION

AI can be used to develop predictive models based on large data sets. We analyzed the results of studies using AI for fracture detection and classification to determine the potential utility in fracture care. In a research setting, we found AI models are nearly as good as humans for detecting certain common fractures and—in two studies—outperformed humans for hip and proximal humerus fracture classification.

This study has several limitations. First, the studies addressed the performance of AI models based on only one projection when radiography was used as the input feature; this is in contrast to daily clinical practice, in which a surgeon bases his or her interpretation on multiple projections combined with taking the patient's history and performing a physical examination. AI models can be built to account for features of the interview, examination, and laboratory values (if applicable) along with image analysis. Second, the studies used a variety of approaches for assigning ground truth labels (the reference standard in AI) for each dataset with which the model was trained. For example, ground truth labels might be determined by a fellowship-trained musculoskeletal radiologist or through a thorough screening of reports in the medical record, consensus meeting among physicians with the additional use of more advanced imaging (such as CT images instead of radiographs) to resolve discrepancies, and radiologist reports. All these reference standards are subject to human error. AI models trained with more objective labeling assignments (for example, operative exposure) should result in more accurate and generalizable probabilities. Third, an appropriate risk of bias assessment tool does not exist for diagnostic studies. We therefore modified the methodologic index for nonrandomized studies (MINORS). Fourth, at present there are only a few preliminary studies used in simple diagnostic scenarios that may overestimate of the potential benefit of AI. Additional studies with clinically relevant settings will help evaluate the utility of AI. Fifth, although a broad search strategy encompassing three large databases was used, potentially relevant publications might have been missed. However, we deem this risk to be low, because we did not identify new eligible studies through screening the reference lists of included studies. In addition, we identified nine conference abstracts that have not been published yet, suggesting that AI is a developing research interest.

Our review found that AI was remarkably good at detecting common fractures. It is reasonable to assume that the fracture locations were selected in these studies because they are common and yield large datasets. Most fractures in these areas are displaced and therefore relatively easy to detect by either a human or a computer. More subtle fractures (such as nondisplaced femoral neck or scaphoid fractures) need additional study as AI models might be less accurate. AI algorithms for diagnosing relatively obvious fractures might be useful for clinical scenarios where fractures might be overlooked (for example, multiple trauma) or in primary care or urgent care where a radiologist is

not immediately available,[7,8] potentially replacing radiologists in this setting. AI could also be useful in difficult scenarios, such as suspected scaphoid or hip fractures, if proven to be accurate. A dispassionate examination of the probability of fracture could help surgeons and patients with decision-making. Further research should seek to identify situations in which AI could act in synergy with clinicians in fracture detection tasks, which are generally prone to misinterpretation or uncertainty. However, there are hurdles to overcome before implementation in clinical practice. First, a clinician might be reluctant to use a suggestion by an AI model since there is no human interface, it is not intuitive (complex statistical models), and it cannot be interrogated (the inscrutability of the magic "black box of AI"). The European Union has addressed liability concerns by incorporating a dictum in the General Data Protection Regulations that AI algorithmic decisions about humans must be interpretable and explainable.[26] Second, it remains debatable who would be held responsible if an algorithm errs and causes harm. Thus, appropriate legal regulations should be addressed before implementing AI into the clinical arena outside of research and quality improvement efforts. Lastly, most studies used datasets with ground truth labels that were based on formal reports from radiologists taken from the medical record to train the respective AI algorithms. For many reasons, these datasets have some inherent errors and misinterpretations. We may benefit from better ground truth labels (for example, operative findings or more sophisticated imaging) to develop more accurate AI algorithms.

AI had reasonable accuracy for classifying proximal humerus and diaphyseal femur fractures. Again, there is an issue with the lack of reference standards for the correct or most likely classification in these studies. For example, Chung et al.[16] determined the reference standard for the Neer Classification[20] by consensus of two shoulder surgeons and one radiologist using CT-images on occasion to reach agreement—an arguably inadequate reference standard for a classification that is known to be unreliable, even using CT scans.[27,28] They also introduced selection bias by removing fractures for which consensus could not be reached. Alternatively, AI might use latent class analysis, a statistical technique that calculates the characteristics of diagnostic performance without a reference standard.[29] Bayesian inferences, another field of interest proposed by Kim and MacKinnon[17] could be used to produce more meaningful predictions that accurately reflects the probable outcome, by accounting for the influence of fracture incidence when analysing accuracy.[17]

Two studies found that AI was better than humans at detecting and classifying hip and proximal humerus fractures, and one found equivalent performance for detecting wrist, hand, and ankle fractures.[16,18,19] This suggests that—at least for relatively straightforward diagnostic scenarios—AI can be useful. There are important gaps to consider. These studies based their ground truth on human assessment (for example, radiology reports or a single radiologist's interpretation).[16,18,19] As clinicians are susceptible to error, the

AI models were trained and tested with images that had some level of inaccuracy.[30] As such, AI models might erroneously report good performance, while this would not be detected as a diagnostic error by the model. Additionally, these AI models can diagnose the fracture, but cannot discern which fractures may involve a bone tumor, for example. In contrast, an orthopaedic surgeon or radiologist is more likely to detect additional relevant findings when evaluating radiographs of fractures. Moreover, physicians are able to combine patients' preferences and objective parameters (such as laboratory values) into careful clinical decision-making.

The current thinking about AI application in medicine seems to be that narrow tasks with predefined context are most suitable, such as recognizing the border of an organ to suggest where to stop scanning or detecting suspicious areas in an image.[31] Risk prediction and therapeutics are more challenging for AI. A lack of reliable and accurate standards on which to train and test an algorithm for certain disease entities (such as delirium), makes the probabilities generated by AI less suitable and applicable for the end-user. Furthermore, an algorithm's output is only an association, not a causative relationship.[32] Therefore, physicians should always balance the probable outcome of this output and decide whether it applies to a specific patient. According to Verghese et al.,[33] AI applications and clinicians should always cooperate: AI helps predict and the clinician compassionately explains and decides.

We speculate that AI might outperform humans for many probabilistic tasks that are based on data. However, the largest challenges will be to find ways to collect and analyze large amounts of data efficiently and to overcome legal issues. Despite the current shortcomings, such as inadequate ground truth label assignment, we believe that physicians will benefit by embracing AI rather than ignoring or dismissing it. For fracture care, these models might aid surgeons by drawing their attention to fractures or fracture characteristics that could cause harm if overlooked. Future studies in this area might focus on AI as a tool to assist with complex and uncertain clinical tasks (for example, determining the response of bone tumors to chemotherapy, or detecting nondisplaced or occult fractures) and in decision support.

ACKNOWLEDGMENTS

None.

REFERENCES

1. Samuel AL. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development.* 1959;3(3):210-220.
2. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115-118.
3. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA.* 2016;316(22):2402-2410.
4. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal.* 2017;35:303-312.
5. Karhade AV, Thio Q, Ogink PT, et al. Development of Machine Learning Algorithms for Prediction of 30-Day Mortality After Surgery for Spinal Metastasis. *Neurosurgery.* 2019;85(1):E83-E91.
6. Thio Q, Karhade AV, Ogink PT, et al. Can Machine-learning Techniques Be Used for 5-year Survival Prediction of Patients With Chondrosarcoma? *Clin Orthop Relat Res.* 2018;476(10):2040-2048.
7. Guly HR. Diagnostic errors in an accident and emergency department. *Emerg Med J.* 2001;18(4):263-269.
8. Pfeifer R, Pape HC. Missed injuries in trauma patients: A literature review. *Patient Saf Surg.* 2008;2:20.
9. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* 2009;6(7):e1000097.
10. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting Hip Fractures with Radiologist-Level Performance Using Deep Neural Networks. 2017; Available at: https://arxiv.org/abs/1711.06504. Accessed November 17, 2017.
11. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A.* 2018;115(45):11591-11596.
12. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (minors): development and validation of a new instrument. *ANZ J Surg.* 2003;73(9):712-716.
13. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Avalaible at: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.
14. Basha CMAKZ, Padmaja M, Balaji GN. Computer Aided Fracture Detection System. *J Med Imaging Health Inform.* 2018;8(3):526-531.
15. Bayram F, Cakirotlu M. DIFFRACT: DIaphyseal Femur FRActure Classifier SysTem. *Biocybern Biomed Eng.* 2016;36(1):157-171.
16. Chung KC, Kim HM, Malay S, Shauver MJ, Group W. Predicting Outcomes After Distal Radius Fracture: A 24-Center International Clinical Trial of Older Adults. *J Hand Surg Am.* 2019;44(9):762-771.
17. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol.* 2018;73(5):439-445.
18. Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthop.* 2017;88(6):581-586.
19. 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.* 2018;48:239-244.
20. Neer CS, 2nd. Displaced proximal humeral fractures. I. Classification and evaluation. *J Bone Joint Surg Am.* 1970;52(6):1077-1089.

21. Müller M. *The Comprehensive Classification of Fractures of Long Bones.* New York, NY: Springer-Verlag. 1990.
22. Yao J, Burns JE, Munoz H, Summers RM. Cortical shell unwrapping for vertebral body abnormality detection on computed tomography. *Comput Med Imaging Graph.* 2014;38(7):628-638.
23. Al-Helo S, Alomari RS, Ghosh S, et al. Compression fracture diagnosis in lumbar: a clinical CAD system. *Int J Comput Assist Radiol Surg.* 2013;8(3):461-469.
24. Russakovsky O, Olga R, Jia D, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis.* 2015(115):211–252.
25. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters.* 2006;27:861-874.
26. Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation". Available at: https://arxiv.org/abs/1606.08813. Accessed June 28, 2016.
27. Carofino BC, Leopold SS. Classifications in brief: the Neer classification for proximal humerus fractures. *Clin Orthop Relat Res.* 2013;471(1):39-43.
28. Majed A, Macleod I, Bull AM, et al. Proximal humeral fracture classification systems revisited. *J Shoulder Elbow Surg.* 2011;20(7):1125-1132.
29. LaJoie AS, McCabe SJ, Thomas B, Edgell SE. Determining the sensitivity and specificity of common diagnostic tests for carpal tunnel syndrome using latent class analysis. *Plast Reconstr Surg.* 2005;116(2):502-507.
30. Hallas P, Ellingsen T. Errors in fracture diagnoses in the emergency department--characteristics of patients and diurnal variation. *BMC Emerg Med.* 2006;6:4.
31. Stead WW. Clinical Implications and Challenges of Artificial Intelligence and Deep Learning. *JAMA.* 2018;320(11):1107-1108.
32. Maddox TM, Rumsfeld JS, Payne PRO. Questions for Artificial Intelligence in Health Care. *JAMA.* 2019;321(1):31-32.
33. Verghese A, Shah NH, Harrington RA. What This Computer Needs Is a Physician: Humanism and Artificial Intelligence. *JAMA.* 2018;319(1):19-20.

APPENDIX 1: SEARCH SYNTAXES FOR THE PUBMED, EMBASE, AND COCHRANE DATABASES

Pubmed – September 6th 2018

("Orthopedic Procedures"[Mesh] OR "Orthopedics"[Mesh] OR "Traumatology"[Mesh] OR fracture*[tiab] OR skeletal fixation*[tiab] OR (trauma*[tiab] AND orthop*[tiab])) AND ("Artificial Intelligence"[Mesh] OR artificial intelligence[tiab] OR neural network*[tiab] OR deep learning[tiab] OR machine learning[tiab] OR machine intelligence[tiab]) AND (predict*[tiab] OR predictive value of tests[mh] OR score[tiab] OR scores[tiab] OR scoring system[tiab] OR scoring systems[tiab] OR observ*[tiab] OR observer variation[mh] OR detect*[tiab] OR evaluat*[tiab] OR analy*[tiab] OR assess*[tiab] OR measure*[tiab])

Embase – September 6th 2018

(exp orthopedic surgery/ or exp orthopedics/ or exp traumatology/ or (fracture* or skeletal fixation*).ti,ab,kw. or (trauma* and orthop*).ti,ab,kw.) AND (exp artificial intelligence/ or exp machine learning/ or (artificial intelligence or neural network* or deep learning or machine learning or machine intelligence).ti,ab,kw.) AND (exp "prediction and forecasting"/ or observer variation/ or (predict* or score or scores or scoring system or scoring systems or observ* or detect* or evaluat* or analy* or assess* or measure*).ti,ab,kw.)

Cochrane – September 6th 2018

((Orthopedic Procedures OR Orthopedics OR Traumatology):MeSH OR (fracture* OR skeletal fixation OR (trauma AND orthop*):ti,ab,kw) AND ((Artificial Intelligence):MeSH OR (artificial intelligence OR neural network* OR deep learning OR machine learning OR machine intelligence):ti,ab,kw) AND ((Predictive Value of Tests OR Observer Variation):MeSH OR (predict* OR score OR scores OR scoring system OR scoring systems OR observ* or detect* or evaluat* OR analy* or assess* OR measure*):ti,ab,kw)

Appendix 2: Critical appraisal of included studies

Study type	Author, year	Disclosure	Study aim	Input feature	Ground truth	Dataset distribution	Performance metric	AI model
detection	Al-helo et al.	1	0	0	0	1	1	1
detection	Basha et al.	0	0	0	0	1	1	1
detection	Chung et al.	1	1	1	1	1	1	1
detection	Gale et al.	0	1	1	1	1	1	1
detection	Kim et al.	1	1	1	1	1	1	1
detection	Lindsey et al.	1	1	1	1	1	1	1
detection	Olczak et al.	1	1	1	1	1	1	1
detection	Urakawa et al.	1	1	1	1	1	1	1
detection	Yao et al.	0	1	1	1	1	1	1
classification	Bayram et al.	0	0	0	0	1	1	1

Disclosure
1, Disclosure is reported.
0, Disclosure is not reported.
Study aim
1, Precise and clear study aim.
0, Study aim not specified or unclear.
Input feature
1, Clear eligibility criteria and all eligble 'samples' (i.e. radiographs/CT-scans) have been included.
0, Potential selection bias or eligibility criteria unclear.
Ground truth
1, Clearly stated description how to determine ground truth.
0, Ground truth not specified or unclear.
Dataset distribution
1, Clearly stated distribution of dataset.
0, Unclear dataset distribution (training-, validation, and test phase).
Performance metric
1, Clear definition of performance of AI model.
0, Performance of AI model not specified or unclear.
AI model
1, Clear explanation of used AI model.
0, Unclear how and which AI model is used.

CHAPTER 4

Is Deep Learning On Par with Human Observers for Detection of Radiographically Visible and Occult Fractures of the Scaphoid?

D.W.G. Langerhuizen
A.E.J. Bulstra
S.J. Janssen
D. Ring
G.M.M.J. Kerkhoffs
R.L. Jaarsma
J.N. Doornberg

Clinical Orthopaedics & Related Research 2020 Nov;478(11):2653-2659

Commentary by:
D. McKee in Clinical Orthopaedics & Related Research 2020 Nov;478(11):2660-2662

ABSTRACT

Background

Preliminary experience suggests that deep learning algorithms are nearly as good as humans in detecting common, displaced, and relatively obvious fractures (such as, distal radius or hip fractures). However, it is not known whether this also is true for subtle or relatively nondisplaced fractures that are often difficult to see on radiographs, such as scaphoid fractures.

Questions/purposes

(1) What is the diagnostic accuracy, sensitivity, and specificity of a deep learning algorithm in detecting radiographically visible and occult scaphoid fractures using four radiographic imaging views? (2) Does adding patient demographic (age and sex) information improve the diagnostic performance of the deep learning algorithm? (3) Are orthopaedic surgeons better at diagnostic accuracy, sensitivity, and specificity compared with deep learning? (4) What is the interobserver reliability among five human observers and between human consensus and deep learning algorithm?

Methods

We retrospectively searched the picture archiving and communication system (PACS) to identify 300 patients with a radiographic scaphoid series, until we had 150 fractures (127 visible on radiographs and 23 only visible on MRI) and 150 non-fractures with a corresponding CT or MRI as the reference standard for fracture diagnosis. At our institution, MRIs are usually ordered for patients with scaphoid tenderness and normal radiographs, and a CT with radiographically visible scaphoid fracture. We used a deep learning algorithm (a convolutional neural network [CNN]) for automated fracture detection on radiographs. Deep learning, an advanced subset of artificial intelligence, combines artificial neuronal layers to resemble a neuron cell. CNNs—essentially deep learning algorithms resembling interconnected neurons in the human brain—are most commonly used for image analysis. Area under the receiver operating characteristic curve (AUC) was used to evaluate the algorithm's diagnostic performance. An AUC of 1.0 would indicate perfect prediction, whereas 0.5 would indicate that a prediction is no better than a flip of a coin. The probability of a scaphoid fracture generated by the CNN, sex, and age were included in a multivariable logistic regression to determine whether this would improve the algorithm's diagnostic performance. Diagnostic performance characteristics (accuracy, sensitivity, and specificity) and reliability (kappa statistic) were calculated for the CNN and for the five orthopaedic surgeon observers in our study.

Results

The algorithm had an AUC of 0.77 (95% CI 0.66 to 0.85), 72% accuracy (95% CI 60% to 84%), 84% sensitivity (95% CI 0.74 to 0.94), and 60% specificity (95% CI 0.46 to 0.74). Adding age and sex did not improve diagnostic performance (AUC 0.81 [95% CI 0.73 to 0.89]). Orthopaedic surgeons had better specificity (0.93 [95% CI 0.93 to 0.99]; $p < 0.01$), while accuracy (84% [95% CI 81% to 88%]) and sensitivity (0.76 [95% CI 0.70 to 0.82]; $p = 0.29$) did not differ between the algorithm and human observers. Although the CNN was less specific in diagnosing relatively obvious fractures, it detected five of six occult scaphoid fractures that were missed by all human observers. The interobserver reliability among the five surgeons was substantial (Fleiss' kappa = 0.74 [95% CI 0.66 to 0.83]), but the reliability between the algorithm and human observers was only fair (Cohen's kappa = 0.34 [95% CI 0.17 to 0.50]).

Conclusions

Initial experience with our deep learning algorithm suggests that it has trouble identifying scaphoid fractures that are obvious to human observers. Thirteen false positive suggestions were made by the CNN, which were correctly detected by the five surgeons. Research with larger datasets—preferably also including information from physical examination—or further algorithm refinement is merited.

INTRODUCTION

Deep learning gained great appeal when Google's DeepMind computer defeated the world's number one Go player.[1] Deep learning, an advanced subset of artificial intelligence, combines artificial neuronal layers to resemble a neuron cell. Essentially, these algorithms— highly complex mathematical models—derive rules and patterns from data to estimate the probability of a diagnosis or outcome without human intervention. These algorithms can be applied to imaging tasks such as skin cancer detection on photographs or detection of critical findings in head CT scans.[2,3]

Using different data set sizes, initial experience with fracture detection on radiographs suggests that deep learning algorithms are (nearly) as good as humans at detecting certain common fractures such as distal radius, proximal humerus, and hip fractures.[4] However, many of those fractures are displaced and relatively obvious on radiographs.

It is known that scaphoid fractures can have long-term consequences if not properly diagnosed. A previous study applied five deep learning algorithms to detect wrist, hand (including scaphoid), and ankle fractures; however, they did not report their algorithm's performance for scaphoid fractures specifically.[5] As such, it is not yet clear whether deep learning algorithms will be useful for the detection of relatively subtle and often radiographically invisible nondisplaced femoral neck or scaphoid fractures that are often overlooked by humans, particularly non-specialists.[6]

Therefore, we asked: (1) What is the diagnostic accuracy, sensitivity, and specificity of a deep learning algorithm in detecting radiographically visible and occult scaphoid fractures using four radiographic imaging views? (2) Does adding patient demographic (age and sex) information improve the diagnostic performance of the deep learning algorithm? (3) Are orthopaedic surgeons better at diagnostic accuracy, sensitivity, and specificity compared with deep learning? (4) What is the interobserver reliability among five human observers and between human consensus and deep learning algorithm?

PATIENTS AND METHODS

Data Set and Pre-processing

Our institutional review board approved this retrospective study. Our institution still uses a paper medical record, which makes it difficult to search for patients with specific diagnoses and tests. The picture archiving and communication system (PACS) is electronic and easier to search. We used two strategies to identify at least 300 scaphoid series of radiographs.

The first strategy was based on the fact that clinicians in our institution usually order an MRI in patients with suspected scaphoid fractures and normal radiographs and a CT

with radiographically visible scaphoid fracture. This strategy identified MRI and CT of the scaphoid and then sought corresponding radiographs of scaphoid fractures. We searched the PACS database using the terms "MR scaph", "CT hand", "CT wrist", and "CT extr" and identified 326 patients: 150 that were excluded because the radiographs were incomplete or distorted by cast or splint materials and 176 with adequate radiographic scaphoid series including 13 MRI-confirmed fractures, 59 CT-confirmed fractures, and 104 MRI-confirmed nonfractures.

In the second strategy, we searched PACS for "Xr scaph" and searched them one by one for a corresponding MRI or CT image and an adequate series of radiographs not distorted by plaster. We found 124 additional patients including 10 with MRI-confirmed fractures, 68 with CT-confirmed fractures, 46 MRI-confirmed nonfractures, and 17 CT-confirmed nonfractures. Two observers (DWGL, AEJB) used this strategy to identify patients until we had 150 radiographs of scaphoids with a fracture (127 visible on radiographs and 23 only visible on MRI) and 150 without a fracture, numbers chosen before starting the search and based on typical training strategies. Age and sex demographics were provided by PACS. The mean age at diagnosis was 36 years (SD 16), and 62% (185 of 300) of patients were male. We randomly divided the dataset into a train, a validation, and a test group (180:20:100), each divided 50:50 by presence of a fracture. The radiographically invisible fractures were randomly and evenly distributed between the three groups. To match the predefined image size of the deep learning framework (Figure 1), we manually cropped and resized all Digital Imaging and Communications in Medicine (DICOM) files into a 350 x 300 pixels rectangle capturing the scaphoid (see Appendix 1; Supplemental Digital Content 1, http://links.lww.com/CORR/A353). By automatically rotating, zooming, changing height/width, and horizontal/vertical flipping, all preformatted images were 10-fold augmented with the intent to increase robustness of the algorithm.

Algorithm: Convolutional Neural Network

Convolutional neural networks (CNNs) are complex algorithms resembling interconnected neurons in the human brain. CNNs are a form of deep learning commonly used to analyze images. In deep learning, the computer analyzes both features that are recognizable to humans (for example, the eyes or the nose) and features that are not recognizable to humans (such as edges or transitions). A CNN learns by developing and testing algorithms again and again (in iterations) until it has optimized its ability to identify the feature assigned: in this case, fracture of the scaphoid. When approaching a new image recognition task, it can be helpful to start with a CNN that is already trained to identify features in images. We used an open-source pretrained CNN (Visual Geometry Group, Oxford, United Kingdom[7]) trained on more than 1 million non-medical images with 1000 object categories[8] (see Appendix 2; Supplemental Digital Content 2, http://links. lww.com/CORR/A354).

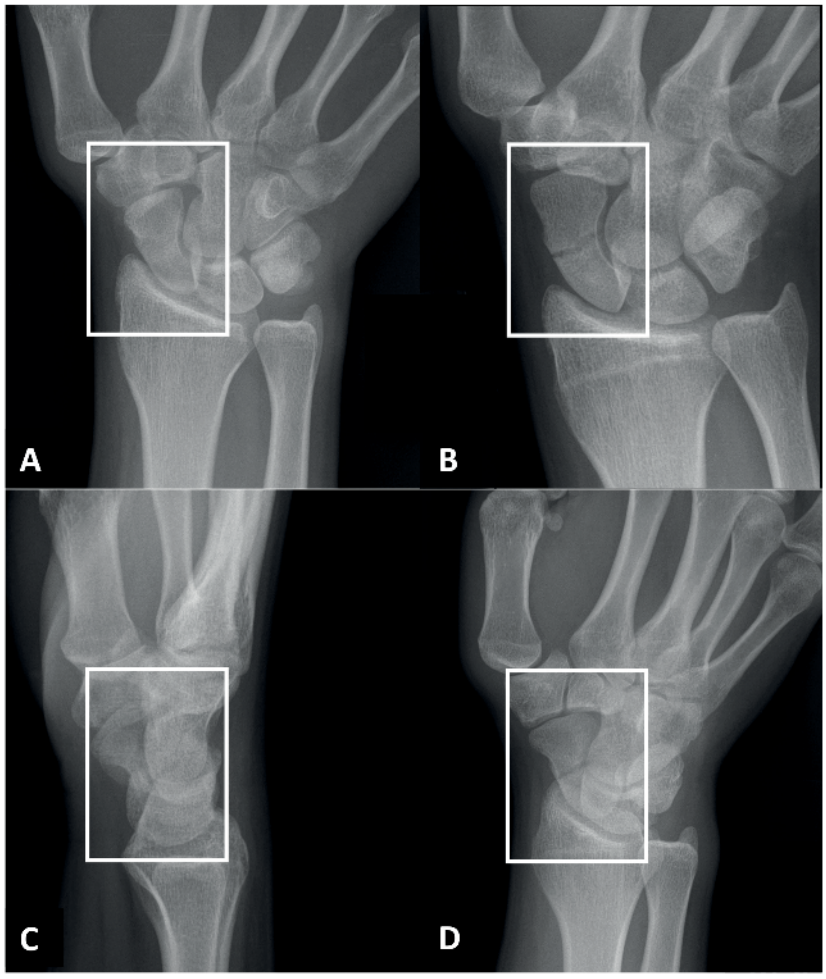


Figure 1: A radiographic scaphoid fracture series for patients with a clinical suspicion for scaphoid fracture at our hospital. The following four projections were fed into the deep learning framework: **(A)** posterior-anterior ulnar deviation; **(B)** uptilt (that is, an elongated view with tube angle adjusted over 30°); **(C)** lateral; and **(D)** 45° oblique projections. The white boxes illustrate the cropped and resized radiographs (350 x 300 pixels) which are fed into the deep learning framework (VGG 16).

A test group of 100 images was randomly selected for use in the tests to determine the algorithm performance. We evaluated the model using the following performance metrics: area under the receiving operating characteristic (AUC) curve, accuracy, sensitivity, and specificity. We set the diagnostic cutoff point at a value that maximized sensitivity, at the cost of a slightly decreased specificity.[6,9,10]

Codes were written in Python Version 3.6.8 (Python Software Foundation, Wilmington, DE, USA) with the packages scikit-learn (0.20.3) and TensorFlow (1.13.1).

Human Observers

We compared the performance metrics of the model with five surgeons (RLJ, JND, MMAJ, NK, JWW). Three orthopaedic trauma surgeons (16, 3, and 2 years after completion of residency training) and two upper limb surgeons (25 and 2 years after completion of residency training) each reviewed the same 100 patients as the model. In our hospital, upper limb surgeons deliver care for the entire upper extremity. The surgeons were not aware of the total number of fracture and nonfracture patients in the test set. All fractures were presented as uncropped Digital Imaging and Communications in Medicine (DICOM) files, which we loaded into Horos (version 3.3.4, Annapolis, MD, USA). Surgeons were asked to identify the presence or absence of a scaphoid fracture on four radiographic views. Again, we calculated the accuracy, sensitivity, and specificity for each surgeon as well as the mean among surgeons for each measure to compare with the CNN.

Statistical Analysis

Continuous variables were presented with mean and SD and categorical variables with frequencies and percentages. Accuracy is defined as the proportion of correctly detected cases among all cases. The AUC reflects the probability that a binary classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.[11] An AUC of 1.0 corresponds to perfect classification, whereas 0.5 indicates a prediction equal to chance. Sensitivity corresponds to the proportion of correctly identified fractures among all actual fractures, while specificity refers to the proportion of correctly identified nonfractures among all nonfractures. We calculated 95% confidence intervals using a Z-score of 1.96. Overlapping 95% CIs indicate no significant difference. A McNemar's test was used to compare sensitivity and specificity between the algorithm and human observers. The probability of a scaphoid fracture generated by the CNN, sex, and age were included in a multivariable logistic regression to determine whether this would improve the algorithm's diagnostic performance.

Kappa, which is a chance-corrected measure, corresponds to the agreement among observers. We used Fleiss' kappa to determine interobserver reliability among surgeons for evaluating the presence or absence of scaphoid fractures. We used Cohen's kappa to calculate reliability between the CNN and majority vote of human observers. According to Landis and Koch[12], a kappa between 0.21 and 0.40 reflects fair agreement, a kappa between 0.41 and 0.60 indicates moderate agreement, a kappa between 0.61 and 0.80 reflects substantial agreement, while a kappa above 0.80 indicates almost perfect agreement.

We performed statistical analyses using Stata 15.0 (StataCorp LP, College Station, TX, USA) and RStudio (Boston, MA, USA) with the packages CalibrationCurves, ggplot2, grid, and precrec.

There were no missing data.

RESULTS

Performance of CNN

For detection of scaphoid fractures among suspected scaphoid fractures, the CNN reported an AUC of 0.77 (95% CI 0.66 to 0.85) (Figure 2). The CNN correctly detected 72 of 100 patients (accuracy 72% [95% CI 60% to 84%]). Eight of 50 confirmed scaphoid fractures were not identified (sensitivity 0.84 [95% CI 0.74 to 0.94]), while 20 of 50 patients without a fracture were incorrectly diagnosed as having a fracture of the scaphoid (specificity 0.60 [95% CI 0.46 to 0.74]).

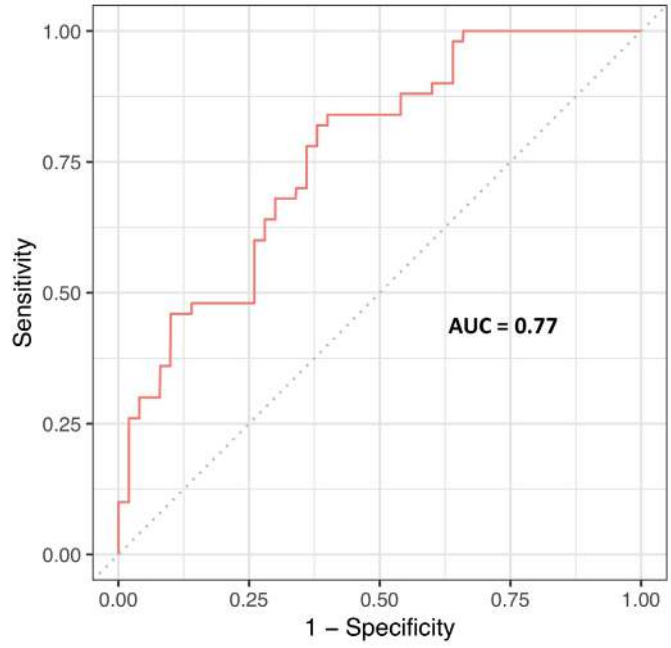


Figure 2: This figure depicts the receiver operating curve for the CNN at the optimal diagnostic cut-off point (0.37).

Performance of CNN Combined with Patient Demographics

Combining age and sex with the generated probabilities of the CNN did not improve the AUC (0.81; 95% CI 0.73 to 0.89). The output of this model was converted into a formula for calculating the probability of a fracture (see Appendix 3; Supplemental Digital Content 3, http://links. lww.com/CORR/A355).

Performance of CNN Compared with Human Observers

Specificity favored the human observers (five orthopaedic surgeons 0.93 [95% CI 0.87 to 0.99] versus CNN 0.60 [95% CI 0.46 to 0.74]; p < 0.01). Accuracy for distinguishing between scaphoid fractures and nonfractures was comparable between human observers and the CNN (five orthopaedic surgeons 84% [95% CI 81% to 88%] versus CNN 72% [95% CI 60 to 84]) (Table 1). Sensitivity was also comparable between the CNN and human observers (five orthopaedic surgeons: 0.76 [95% CI 0.70 to 0.82]) versus CNN: 0.84 [95% CI 0.74 to 0.94]; p = 0.29).

Six scaphoid fractures were missed by all surgeons and therefore considered occult fractures. The CNN detected five of six occult scaphoid fractures. In addition, five human observers detected three fractures that were missed by the CNN. Two fractures, diagnosed by four of five human observers, were also missed by the CNN. In contrast, thirteen false positive suggestion of the CNN, were correctly detected by the surgeons.

Table 1. A comparison of performance metrics between the CNN and the mean of five orthopaedic surgeons

Diagnostic performance characteristic	Orthopaedic surgeons	CNN[a]	p value
Accuracy (95% CI)	84% (81% to 88%)	72% (60% to 84%)	[b]
Sensitivity (95% CI)	0.76 (0.70 to 0.82)	0.84 (0.74 to 0.94)	0.29
Specificity (95% CI)	0.93 (0.87 to 0.99)	0.60 (0.46 to 0.74)	< 0.01

[a]CNN = convolutional neural network at cutoff point 0.37.
[b]We did not calculate a p value, since McNemar's test is sensitive to the proportion of fractures as well as nonfractures.
Bold indicates statistical significance (p < 0.05).

The Interobserver Reliability of Human Observers

Interobserver agreement between five surgeons was higher than between human consensus and the algorithm (0.74 [95% CI 0.66 to 0.83] versus 0.34 [95% CI 0.17 to 0.50]) (Table 2).

Table 2. Contingency table comparing prediction of convolutional neural network to human consensus (agreement ≥ three surgeons)

		Fracture (n = 50)	Non-fracture (n = 50)
Fracture (predicted)	CNN	42	20
	Human consensus	38	1
Non-fracture (predicted)	CNN	8	30
	Human consensus	12	49

CNN = convolutional neural network

DISCUSSION

In medicine, deep learning has primarily been applied to image analysis. In a research setting, use of deep transfer learning showed promising performance for fracture detection and classification for relatively straightforward clinical scenarios.[4] It is not yet clear that deep learning will be useful for radiographic fracture detection in scenarios where fractures are often overlooked by human observers. Using a relatively small data set of 300 patients, our deep learning algorithm demonstrated a moderate better overall performance for detection of radiographically visible and occult fractures (AUC 0.77 [95% CI 0.66 to 0.85]) and human observers had notably better specificity. The algorithm might have performed better if provided with more data.

This study has several limitations. First, we selected our patients from readily available and searchable radiology reports and intentionally introduced a spectrum bias by collecting 150 MRI- or CT-confirmed fractures and 150 confirmed nonfractures. Although this was needed to sufficiently train the algorithm, readers should keep in mind that our data set does not represent the true prevalence of radiographic scaphoid fracture appearance. Second, we were only able to include 300 patients because we could only search a 9-year period starting in January 2010. Three hundred radiographs is a relatively small sample size for deep learning, but more than adequate for logistic regression. A larger data set might improve the diagnostic performance of the CNN. We cannot be certain because, to this point, there is no consensus on a priori sample size in deep learning. It depends on the specific image-analysis task, the quality of the data set, the programming techniques used, and type of deep learning algorithm applied.[13] Third, the ground truth labels (that is, the reference standard diagnosis of scaphoid fracture or not) are based on radiologist interpretations of CT or MRI images, which have limited reliability and untestable accuracy. Given the small number of MRIs with diagnosed fracture and CT with diagnosed nonfractures, we believe any misdiagnoses would have little influence on the model. Fourth, radiographs were manually cropped and resized by one investigator (DWGL), which might introduce bias. However, given that cropping was assisted by an easy-to-use program scripted in Python, we feel it is very likely that another investigator would resize the images similarly. But, one should keep in mind that cropped radiographs may not reflect a clinical scenario, as other potentially relevant findings in a real-size radiograph were not assessable (such as, concomitant fractures or scapholunate dissociation). Furthermore, irrelevant regions in a radiograph were removed and therefore not evaluated by the model. A more in-depth deep learning framework, accounting for the entire wrist radiograph, merits further study. For now, the memory capacity of graphics processing units limits the usable image size. Fifth, among the five human observers, two surgeon raters treated some of the patients in the study, which might have influenced their diagnoses. We feel this would have negligible influ-

ence on our findings. Sixth, although incorporating injury details, signs, and symptoms would have been of interest to incorporate in a logistic regression model as is typical for a clinical prediction rule, they were not commonly reported in a patient's medical record. CNNs only evaluate images, but the probabilities generated can be included in clinical prediction rules.

The AUC of the CNN for detection of scaphoid fractures is not good enough to replace human observers or more sophisticated imaging, but it does suggest the potential to be used as a pre-screen or clinical prediction rule for triage of suspected scaphoid fractures that might benefit additional imaging. Displaced proximal humerus, distal radius, and intertrochanteric hip fractures are relatively easy to detect and not a good test of the potential utility of artificial intelligence.[9,10,14] Subtle and invisible fractures may be more of a challenge. Prior studies using deep learning algorithms to detect radiographically subtle hip and distal radius fractures had better performance than our model.[6,14-16] Larger data sets, use of other pre-trained CNNs, varying degrees of algorithm refinement and hyper-parameter tuning, as well as other anatomical fracture locations may explain why these studies differ with our findings. Also, we might not have had sufficient images to train the upper layers of the pretrained CNN.

Adding sex and age did not improve diagnostic performance. Future research might investigate whether incorporating computer analysis of images improves performance of clinical prediction rules that include demographics, injury details, symptoms, and signs to better triage the use of MRI as well as increase its diagnostic performance by increasing the pretest odds of a fracture.[17,18] The pretest odds could be increased with CNNs, clinical prediction rules, or a combination of both.

Our deep learning algorithm was less specific than human observers but detected five of six occult fractures in the test dataset. On the other hand, caution is warranted because the CNN missed some radiographically visible fractures.

The finding that reliability of fracture diagnosis was substantial (0.74) for the five orthopaedic surgeons and only fair (0.34) between the surgeons and the CNN we interpret as a reflection of the difficulty the deep learning algorithm has with detecting radiographically visible fractures. At the diagnostic cutoff point—chosen to maximize sensitivity—the algorithm's specificity was considerably lower compared with human observers. A different cutoff point may have resulted in more or less the same reliability for detecting scaphoid fractures. It may go without saying that CNNs are known for being highly complex and, to date, not intuitive for the end-user. It is therefore not possible to understand how a CNN reaches its suggestion.

In conclusion, using a relatively small dataset, a deep learning algorithm was inferior to human observers at identifying scaphoid fractures on radiographs. Further study may help evaluate whether a larger dataset and algorithm refinement can increase the performance of deep learning for the diagnosis of scaphoid fractures, some of which are

radiographically invisible. In addition, incorporating predictions from a deep-learning algorithm into clinical prediction rules that also account for demographics, injury details, symptoms, and signs merits further study.

ACKNOWLEDGMENTS

We thank the following orthopaedic surgeons for their participation: M. M. A. Janssen MD, PhD, N. Kruger MD, and J. W. White MBBS, PhD.

REFERENCES

1. British Broadcasting Corporation. Artificial intelligence: Go master Lee Se-dol wins against AlphaGo program. Available at: https://wwwbbccom/news/technology-35797102 Accessed March 13, 2016.
2. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet (London, England).* 2018;392(10162):2388-2396.
3. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115-118.
4. 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.
5. Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthop.* 2017;88(6):581-586.
6. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol.* 2018;73(5):439-445.
7. Zhong S, Li K, Feng R. Deep Convolutional Hamming Ranking Network for Large Scale Image Retrieval. *Available at:* https://ieeexploreieeeorg/document/7052856 *Accessed Augustus 19, 2016.*
8. Russakovsky O, Olga R, Jia D, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis.* 2015(115):211–252.
9. 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.
10. Gan K, Xu D, Lin Y, et al. Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. *Acta Orthop.* 2019;90(4):394-400.
11. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters.* 2006;27:861-874.
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159-174.
13. Ranschaert ER. Artificial Intelligence in Medical Imaging. *eBook Switzerland, AG: Springer.* 2019.
14. 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.* 2018;48:239-244.
15. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting Hip Fractures with Radiologist-Level Performance Using Deep Neural Networks. 2017; Available at: https://arxiv.org/abs/1711.06504. Accessed November 17, 2017.
16. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A.* 2018;45:11591-11596.
17. Duckworth AD, Buijze GA, Moran M, et al. Predictors of fracture following suspected injury to the scaphoid. *J Bone Joint Surg Br.* 2012;94(7):961-968.
18. Rhemrev SJ, Beeres FJ, van Leerdam RH, Hogervorst M, Ring D. Clinical prediction rule for suspected scaphoid fractures: A prospective cohort study. *Injury.* 2010;41(10):1026-1030.

APPENDIX 1. CODE FOR CAPTURING RADIOGRAPHS INTO 350 X 300 PIXELS RECTANGLE

```
#load libraries and packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import pydicom
import os
import sys
import pickle
from scipy import ndimage

savefile = directory

#load file
df = pd.read_pickle(savefile)  #load savefile
df

# create dictionary to lookup images
ext=directory
pt_list=os.listdir(ext)
#enter second nested file
print(ext)
print(pt_list)

d={}
for f in pt_list:
if f==’.DS_Store’:
continue
pt_list2 = ext + ‘/’ + f
dir2 = os.listdir(pt_list2)
extlist = []
for dcmfile in dir2:
if dcmfile==’.DS_Store’:
continue
dcm_ext = pt_list2 + ‘/’ + dcmfile
extlist.append(dcm_ext)
```

```
d[f] = extlist

#find radiograph to crop and resize
pydicom.dcmread(d['xx'][x])

#ptn
p='xx'
#projection
x=#
ext=d[p][x]
print(ext)
#print('study exists of:', len(d[list(d.keys())[p]]), 'images')
dcm=pydicom.dcmread(ext)
date=dcm.StudyDate
time=dcm.StudyTime[:6]
image = dcm.pixel_array
print('date_time_stamp:',date, time)
print(dcm.pixel_array.shape)
plt.imshow(image, cmap=plt.cm.bone)

print(df.iloc[-5:,:])

def rotate_img(img, angl):
rotated_img = ndimage.rotate(img, angle = angl, reshape=False)
return rotated_img

#dcm=pydicom.dcmread(ext).pixel_array
angl = x
image = rotate_img(image, angl)
plt.imshow(image, cmap=plt.cm.bone)
plt.show()

def crop_dicom(img, y_start, x_start, len_y, len_x):
#pix_array=pydicom.dcmread(extension).pixel_array
pix_array = img
print(pix_array.shape)

pix_crop=pix_array[y_start:y_start+len_y,x_start:x_start+len_x]
```

```
return pix_crop

imt='pa'

if imt=='pa':
ly=350
lx=300
elif imt=='lat':
ly= 350
lx = 300
elif imt=='obl':
ly=350
lx=300
elif imt=='up':
ly=350
lx=250

crop_img=crop_dicom(image, y_start= 50, x_start=0, len_y=ly,len_x=lx)
plt.imshow(crop_img, cmap=plt.cm.bone)
shape=crop_img.shape
print('date_time_stamp:',date, time)
print(shape)

#save cropped radiographs
def append_data(img_array, ptno, shape, df, type_):
newrow=[ptno, shape, img_array, type_]
df.loc[len(df)]=newrow
return df
df=append_data(crop_img, p, shape, df, imt)
df.iloc[-5:,:]

#write to disk
df.to_pickle(savefile)
```

APPENDIX 2

Pre-processing of Data

The algorithm was optimized according to the following train, validation, and test split: 180-20-100. All radiographs were manually cropped and resized to match the predefined image size of the deep learning framework (that is, a 200 x 300 pixels rectangle). We downscaled the pixel intensity by averaging each pixel based on minimum and maximum intensity of the radiograph. To increase robustness of the algorithm, we 10-fold augmented the training and validation set by using rotation (-15° and +15°), shifting of height and width (10%), zooming (between 0.8 and 1.1), and horizon flipping. The test set only composed of original radiographs.

Training of Deep Learning Framework

We used keras API (https://keras.io) to run on top of the open-source Imagenet pre-trained Visual Geometry Group (VGG) 16-layer convolutional neural network (CNN) [7]. We ran Intel(R) Xeon(R) W-2175 (clock speed 2.50GHz, 64 GB RAM) with NVIDIA TITAN V (boostclock 1455 MHz, 12 GB HBM2). The outputs of the last CNN-layer were fine-tuned to our scaphoid fracture dataset with a concatenation operation followed by the fully connected top network. End-to-end fine-tuning of the last convolutional layers was performed, while earlier layers—containing more generic features—were kept fixed. We decided not to further fine-tune the convolution layers because it resulted in more overfitting.

To train the algorithms for 30 epochs, we applied a grid search to find the optimal parameters (including fully connected top architecture). The best three top models—evaluated with accuracy—were trained with an early stopping criterion of 0.001 over the last five epochs. The optimal hyper parameters were used to fine tune the last convolutional layers of the four parallel VGG16 CNN architectures.

Table 1. Hyperparameter optimization

Hyperparameter	Values applied in grid search
Fully connected top layers and the nodes used in the top layers	256, 512, 1024, 4096, 256-512, 256-1024, 512-512, 512-1024, 1024-512, 1024-1024, 4096-1024, 512-512-512, 512-1024-512, 512-4096-512, 4096-1024-512
Activation / weight regularization term per layer	1e-8, 1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1
Use of dropout layer	True / False
Drop-out in the drop-out layer	0.3, 0.4, 0.5,0.6,0.7
Learning rate	1e-3, 1e-4, 1e-5
Batch size	32
Optimizer	Stochastic gradient decent without Nesterov momentum of 0.9
Epochs trained	30, 50

To train the algorithms for 30 epochs, we applied a grid search to find the optimal parameters (including fully connected top architecture). The best three top models--evaluated with accuracy--were trained with an early stopping criterion of 0.001 over the last 5 epochs. The optimal hyper parameters were used to fine tune the last convolutional layers of the 4 parallel VGG16 CNN architectures.

Table 2. Optimal settings for algorithms

Model description and optimization	Optimal settings
Algorithm 1	
Model architecture (nodes)	256
Activation/weight regularization term per layer	1e-4/1e-3
Use of drop-out layer	TRUE
Drop-out in drop-out layer	0.5
Learning rate	1.00E-04
Algorithm 2	
Model architecture (nodes)	512-1024
Activation/weight regularization term per layer	1e-5/1e-4
Use of drop-out layer	TRUE
Drop-out in drop-out layer	0.5
Learning rate	1.00E-04
Algorithm 3	
Model architecture (nodes)	512-1024-512
Activation/weight regularization term per layer	1e-6/1e-4
Use of drop-out layer	TRUE
Drop-out in drop-out layer	0.5
Learning rate	1.00E-04

Final specifications of best algorithms with the corresponding optimized hyperparameters

APPENDIX 3: ODDS RATIOS FOR AGE AND SEX AND EQUATION FORMULA OF THE PREDICTION MODEL

Odds ratios

Age: 0.97 (95% CI: 0.94 – 1.01)

Sex: 2.55 (95% CI: 0.76 - 8.55)

Linear predictor

-1.816599 + (probability CNN) * 4.680619 + (age) * -0.0265213 + (sex) * 0.9346456

Equation formula to calculate probability of a scaphoid fracture

EXP(Linear Predictor) / (EXP(Linear Predictor) + 1)

PART III

Clinical Predictors for Surgical Decision Making

CHAPTER 5

Factors Associated with a Recommendation for Operative Treatment for Fracture of the Distal Radius

D.W.G. Langerhuizen
S.J. Janssen
J.T.P. Kortlever
D. Ring
G.M.M.J. Kerkhoffs
R.L. Jaarsma
J.N. Doornberg

Journal of Wrist Surgery 2021 Aug;10(4):316-321

ABSTRACT

Background

Evidence suggests that there is substantial and unexplained surgeon-to-surgeon variation in recommendation of operative treatment for fractures of the distal radius. We surveyed a global collaborative to understand bias and variation among surgeons to identify patient factors that influence recommendation for operative treatment of a fracture of the distal radius.

Question/Purposes

(1) What factors are associated with recommendation for operative treatment of a fracture of the distal radius? (2) Which factors are rated as the most influential on recommendation of operative treatment?

Methods

One hundred thirty-one upper extremity and fracture surgeons evaluated 20 fictitious patient scenarios with randomly assigned factors (e.g. personal, clinical, and radiologic factors) for patients with a fracture of the distal radius. They addressed the following questions: (1) Do you recommend operative treatment for this patient (yes/no)? We determined the influence of each factor on this recommendation using random forest algorithms. Also, participants rated the influence of each factor—excluding age and sex— on a scale from 0 (not at all important) to 10 (extremely important).

Results

Random forest algorithms determined that age and angulation were having the most influence on recommendation for operative treatment of a fracture of the distal radius. Angulation on the lateral radiograph and presence or absence of lunate subluxation were rated as having the greatest influence and smoking status and stress levels the lowest influence on advice to patients.

Conclusions

The observation that—other than age—personal factors have limited influence on surgeon recommendations for surgery may reflect how surgeon cognitive biases, personal preferences, different perspectives, and incentives may contribute to variations in care. Future research can determine whether decision aids—those that use patient-specific probabilities based on predictive analytics in particular—might help match patient treatment choices to what matters most to them, in part by helping to neutralize the influence of common misconceptions as well as surgeon bias and incentives.

INTRODUCTION

In the Netherlands, about 20-30% of patients with a fracture of the distal radius are treated operatively compared to 70-80% in Australia.[1,2] Given that fracture patterns are similar in both countries, it is likely that surgeons are having undue bias on treatment decisions.

Decision aids are intended to neutralize the influence of the surgeon and limit misdiagnosis of patient preferences by ensuring their test and treatment choices are based on what matters most to them (their values) rather than misconceptions or surgeon bias. Estimates of the probabilities of various adverse events and the level of symptoms and limitations based on their specific injury pattern and personal characteristics might improve the appeal and utility of the decision aid. Artificial-intelligence (AI) algorithms–probability calculators developed by programs that can iteratively learn from additions of data–may provide more detailed and accurate estimates for various outcomes.

To date, data-driven models can help determine who will develop diseases such as diabetes type 2 and chronic kidney disease, and they can estimate years of survival in patients with bone tumours.[3-6] More sophisticated statistical models using large amounts of data may result in better probability modelling, which can facilitate decision-making. As a first step in developing AI predictive algorithms for fractures of the distal radius, it might help to understand factors influencing surgeon bias and variation in recommendations for surgical management. Scenario-based survey studies help determine the sources of potentially unhelpful practice variation. Variation that persists in spite of clinical guidelines and appropriateness criteria.

Therefore, we surveyed a large international group of surgeons 1) to identify factors in fictitious case scenarios that influence recommendation for operative treatment of a fracture of the distal radius; and 2) to rate these same factors on a scale from 0 to 10. These factors can subsequently be used to develop distal radius fracture specific decision aids intended to help patients come to decisions consistent with their values, independent of surgeon bias.

METHODS

Participants

All 630 members of the 'Science of Variation Group' (SOVG) were invited to complete our survey. The SOVG is a global web-based collaboration based on camaraderie, and without financial incentive, intended to evaluate variation in interpretation, classification, and treatment of musculoskeletal illness.[7,8] Among 178 members that felt this survey pertained to their area of expertise, 131 (74%) completed the survey and were kept for

analysis (Table 1). Participants specialized in fracture surgery or upper-extremity. We used SurveyMonkey (Palo Alto, CA, USA) to create the online survey. The invitation was sent on March 29, 2019 with 2 reminders at 2 and 4 weeks.

Table 1: Surgeon and practice characteristics

	Total = 131	
	N	%
Sex		
Male	121	92
Female	10	8
Location of practice		
Canada/United States	67	51
Europe	40	31
Other*	24	18
Subspecialty		
Hand and wrist	63	48
Orthopaedic trauma	53	40
Shoulder and elbow	15	11
Years in practice		
0-5	30	23
6-10	31	24
11-20	56	43
21-30	14	11
Cases per month		
0-5	55	42
6-10	52	40
>10	24	18
Supervising trainees		
Yes	16	12
No	115	88

N = number of participating surgeons.
*Other locations of practice were Asia in 2 (2%), Australia in 5 (4%), and South-America in 17 (13%)

Factors Influencing Surgeon Recommendation for Operative Treatment

We presented each participating surgeon a unique set of 20 randomly created fictitious scenarios of patients with a fracture of the distal radius. For each patient scenario, participating surgeons were asked: Do you recommend operative treatment for this patient (yes/no)? The following 20 factors were randomly assigned in each scenario: 1) age (either between 43-57 years or 63-77 years); 2) sex (male *versus* female); 3) smoking status (yes *versus* no); 4) mechanism of injury (high-energy *versus* low-energy); 5) health and activity level of patient (healthy and active *versus* infirm and inactive); 6) patient wishes to avoid surgery (yes *versus* no); 7) coping strategies (very effective *versus* moderately effective); 8) psychological distress (moderate symptoms of depression or anxiety *versus* slight symptoms of depression or anxiety); 9) stress (moderate financial, housing, health or relationship stress *versus* slight stress); 10) step-off (<2mm *versus* ≥2mm); 11) gap (<2mm *versus* ≥2mm); 12) radial shortening (<3mm *versus* ≥3mm); 13) radial angulation (≥15° dorsal angulation *versus* ≥20° volar angulation *versus* neither volar nor dorsal angulation); 14) radial inclination (≥15° *versus* <15°); 15) lunate subluxation (presence or absence); 16) dorsal metaphyseal comminution (presence or absence); 17) volar metaphyseal comminution (presence or absence); 18) volar rim fracture (presence or absence); 19) die punch fragment (presence or absence); and 20) volar lunate facet fragment (presence or absence).

Surgeon Rating of Factor Influence

Each participating surgeon rated the patient factors for influence on advice to patients on a scale from 0 (not at all important) to 10 (extremely important). Age and sex were accidentally omitted.

Statistical analysis

We used random forest algorithms to determine the influence of each patient factor on recommendation for surgery in the patient scenarios.[9] Random forest is a machine learning technique commonly used to measure the relative influence (referred to as importance) of each variable on variation in the outcome. The factor with the largest influence on decisions is discerned via analysis of multiple, mathematically created decision trees based on conditional statements (i.e. if a condition is met, then an action is performed). The influence of all other factors is then compared to this factor by dividing the score of each variable over the factor with the largest influence. The most important variable has a score of 1.0.

Among the rated list of 18 patient factors, factors influencing surgical decision-making were presented with mean and standard deviation (SD).

There were no missing data. All statistical analyses were performed using Stata 15.0 (StataCorp LP, College Station, TX).

RESULTS

Factors Influencing Surgeon Recommendation for Operative Treatment

Random forest algorithms showed that age was rated most influential and assigned a variable influence of 1 (Figure 1). Angulation was the second most influential factor (ranked 0.78 due to its relative influence). All other factors, varying from 0.67 (patient factor: avoid surgery) to 0.63 (patient factor: gap), were deemed of less influence.

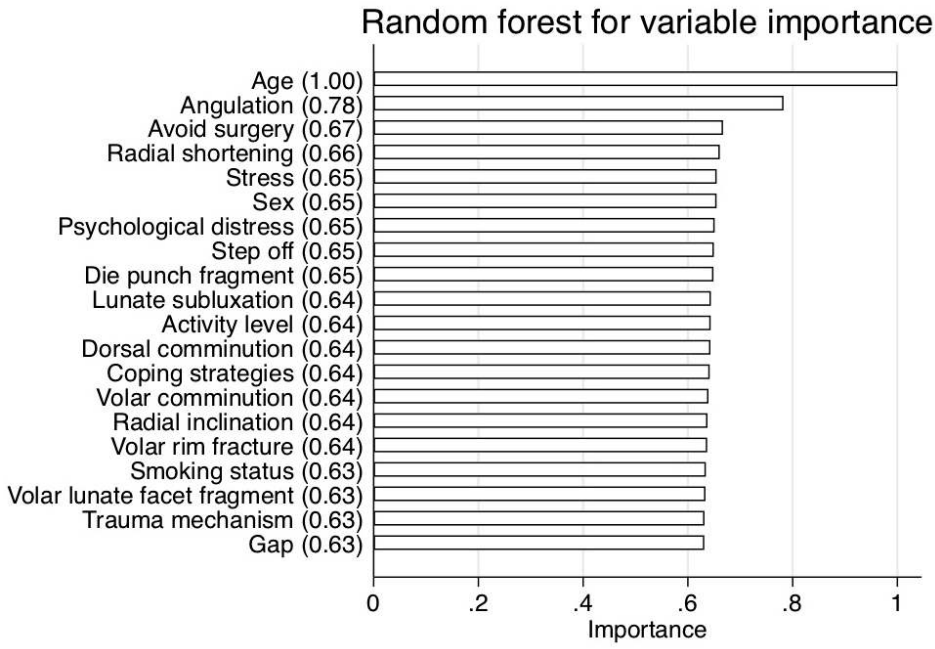


Figure 1: Factor influence depicted for all patient factors included in 20 'fictitious' case scenarios. Between parentheses factor influence for each variable.

Surgeon Rating of Factor Influence

Participants rated angulation (≥15 dorsal vs ≥20 volar vs. neither volar nor dorsal angulation) as having greatest influence on recommending operative treatment with a mean of 7.8 (SD 1.6), followed by presence or absence of lunate subluxation with a mean of 7.7 (SD 1.9) (Table 2). Smoking status (yes/no) was rated as having lowest influence (mean 3.0; SD 2.5), followed by stress (moderate financial, housing or relationship stress vs. slight; mean 4.0; SD 2.3).

Table 2: List of factors influencing outcome as rated by participants on a scale from 0 to 10

	Mean (±standard deviation)
Angulation (≥15 dorsal vs. ≥20 volar vs. neither volar nor dorsal angulation)	7.8 ± 1.6
Lunate subluxation (presence or absence)	7.7 ± 1.9
Die punch (presence or absence)	7.3 ± 1.9
Volar lunate facet (presence or absence)	7.2 ± 2.0
Step-off (<2mm vs. ≥2mm)	7.1 ± 2.0
Volar metaphyseal comminution (presence or absence)	6.9 ± 2.3
Volar rim (presence or absence)	6.9 ± 2.1
Radial shortening (<3mm vs. ≥3mm)	6.8 ± 2.0
Health and activity level of patient (healthy and active vs. infirm and inactive)	6.6 ± 2.3
Dorsal metaphyseal comminution (presence or absence)	6.4 ± 2.3
Patient wishes to avoid surgery (yes/no)	5.9 ± 2.4
Radial inclination (≥15 vs. <15)	5.9 ± 2.1
Gap (<2mm vs. ≥2mm)	5.5 ± 2.3
Mechanism of injury (high vs. low-energy)	5.4 ± 2.6
Effective coping strategies (very effective vs. moderately effective)	4.7 ± 2.3
Psychological distress (moderate symptoms of depression or anxiety vs. slight)	4.2 ± 2.4
Stress (moderate finanical, housing, or relationship stress vs. slight)	4.0 ± 2.3
Smoking status (yes/no)	3.0 ± 2.5

DISCUSSION

Surgeons seem to have undue bias on a decision for operative treatment of a fracture of the distal radius.[1,2] Decision aids are intended, in part, to reduce treatment inefficiencies and practice variation.[10] To inform efforts to guide people to a decision based on what matters most to them and not based on misconceptions or biases, we studied factors that have the strongest influence on surgeon recommendations. Based on 'fictitious' case scenarios, it seems that age and angulation have the strongest influence on surgeon offer of operative treatment. Radiographic parameters and fracture characteristics were rated most important for guiding patients to operative or non-operative treatment.

This study has several limitations. First, the reader needs to consider whether the participating surgeons are representative of the average surgeon. Most participants work in academic practice, which might limit generalizability. On the other hand, practice setting accounts for minimal variation in our studies. Furthermore, surgeons are mainly from North America and Europe (107 out of 131 participating surgeons [82%]). Second, prior studies with similar 'fictitious' case scenario survey designs only incorporated 5 to 6 factors.[11,12] The number of patient factors in our case scenarios may have interfered with surgical decision-making. As a result, most factors–except age and an-

gulation–seem to have comparable influence on treatment recommendations. Indeed, Halford et al.[13] found that humans are only capable of processing four variables into a single cognitive representation at a time. Third, although each fictitious case scenario assigned each patient factor (including age and sex), we accidentally forgot to include age and sex as patient factors in the list where surgeons ranked importance. Given that age had substantial influence in the patient scenarios, it would have likely been ranked highly. Fourth, we only used categorical variables in our scenarios to facilitate statistical analysis. For example, angulation was categorized. This may have led to some confusion. For instance, we anticipate some surgeons may not have clearly understood "neither volar nor dorsal angulation." Our chosen threshold (e.g. volar angulation ≥20°) were not extreme enough to create clear indications for surgery for many surgeons. These types of potential weaknesses apply to most studies of artificial patient scenarios. Fifth, radial shortening was one of the factors surveyed in our study, but ulnar variance is the preferred measure nowadays. Sixth, including other factors such as presence or absence of polytrauma or carpal tunnel syndrome may also have been of interest, but were not included as we decided to specifically focus on 20 factors that are considered most important for distal radius fractures. Seventh, the randomization process may have created rare combinations such as dorsally displaced fractures with volar lunate facet fractures, but this should be uncommon enough to have a limited influence. In addition, one should keep in mind that psychosocial factors such as psychosocial distress or effective coping strategies may have been assessed differently during a face-to-face assessment compared to fictive written case scenarios. As such, face-to-face evaluation may even introduce more surgeon bias, potentially resulting in greater variation than already found in this study. This potentially may also apply to radiographs instead of concrete written descriptions.

We found age and angulation to be the most influential factors for recommending operative treatment, while all other patient factors partitioned the data in a top-down decision tree structure in subsets with similar gained information. One can conclude that the impact from all other factors is similar and a hierarchy of impact could not be discerned among these factors. Kyriakedes et al.[12] also found that age–together with fracture displacement–is associated with surgeon recommendation for operative treatment of distal radius fractures. Data driven computer algorithms have the potential to provide estimated probabilities of treatment outcomes directly to patients, which might limit the influence of surgeon bias and preferences.[14,15] And computers can account for more factors simultaneously than humans.

The finding that surgeons rate radiographic parameters and fracture characteristics as having greater influence on treatment recommendations than patient preferences, health, activity level, mind-set, and circumstances suggests that surgeons feel they are doing their best when they stick to the proper pathoanatomical indications. On the

other than this finding does suggest that surgeons may underestimate the influence of personal factors on ultimate outcomes when guiding patients as they choose a treatment strategy.[16-20]

The discrepancy between what surgeons describe as important factors and what the statistical analysis demonstrates are influencing recommendations are evidence of surgeon-to-surgeon variation in the relatively influence of various factors, and perhaps the degree to which stated importance matches actual influence. Such discrepancies between declared and measured influence may reflect surgeon cognitive biases, past experiences, incentives, and personal preferences. For instance, surgeon personality traits, such as self-awareness or uncertainty, are substantially related to treatment decisions in musculoskeletal fracture care.[21,22] In addition, surgeons are less likely to choose surgery for themselves than they are to recommend surgery for a patient.[23]

Personal factors seem to have limited influence on surgeon recommendation for operative treatment. Surgeons may be estimating probabilities of various outcomes largely based on radiographic parameters, making a value judgement about those estimates, and then making a recommendation to patients. Future studies should focus on analysis of data by using sophisticated predictive analytics commonly referred to as artificial intelligence. This may provide more accurate estimates of various outcomes than surgeon expertise, wisdom, and gestalt. Furthermore, it may go without saying that what matters most to patients (patient values) ought to take precedence over what matters to surgeon (surgeon values, incentives, and biases).

REFERENCES

1. Ansari U, Adie S, Harris IA, Naylor JM. Practice variation in common fracture presentations: a survey of orthopaedic surgeons. *Injury.* 2011;42(4):403-407.
2. Walenkamp MM, Mulders MA, Goslings JC, Westert GP, Schep NW. Analysis of variation in the surgical treatment of patients with distal radial fractures in the Netherlands. *J Hand Surg Eur Vol.* 2016.
3. Karhade AV, Thio Q, Ogink PT, et al. Development of Machine Learning Algorithms for Prediction of 30-Day Mortality After Surgery for Spinal Metastasis. *Neurosurgery.* 2018.
4. Thio Q, Karhade AV, Ogink PT, et al. Can Machine-learning Techniques Be Used for 5-year Survival Prediction of Patients With Chondrosarcoma? *Clin Orthop Relat Res.* 2018;476(10):2040-2048.
5. Perotte A, Ranganath R, Hirsch JS, Blei D, Elhadad N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. *J Am Med Inform Assoc.* 2015;22(4):872-880.
6. Bennett WL, Maruthur NM, Singh S, et al. Comparative effectiveness and safety of medications for type 2 diabetes: an update including new drugs and 2-drug combinations. *Ann Intern Med.* 2011;154(9):602-613.
7. Janssen SJ, Hermanussen HH, Guitton TG, van den Bekerom MP, van Deurzen DF, Ring D. Greater Tuberosity Fractures: Does Fracture Assessment and Treatment Recommendation Vary Based on Imaging Modality? *Clin Orthop Relat Res.* 2016;474(5):1257-1265.
8. Bruinsma WE, Guitton TG, Warner JJ, Ring D. 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(17):1600-1604.
9. Zou RY, Schonlau M. The Random Forest Algorithm for Statistical Learning with Applications in Stata. *The Stata Journal.* 2018.
10. Members of the Writing R, Voting Panels of the AUCotToDRF, Watters WC, Sanders JO, Murray J, Patel N. The American Academy of Orthopaedic Surgeons Appropriate Use Criteria on the treatment of distal radius fractures. *J Bone Joint Surg Am.* 2014;96(2):160-161.
11. Hageman MG, Becker SJ, Bot AG, Guitton T, Ring D, Science of Variation G. Variation in recommendation for surgical treatment for compressive neuropathy. *J Hand Surg Am.* 2013;38(5):856-862.
12. Kyriakedes JC, Crijns TJ, Teunis T, Ring D, Bafus BT, Science of Variation G. International Survey: Factors Associated with Operative Treatment of Distal Radius Fractures and Implications for the American Academy of Orthopaedic Surgeons Appropriate Use Criteria. *J Orthop Trauma.* 2019.
13. Halford GS, Baker R, McCredden JE, Bain JD. How many variables can humans process? *Psychol Sci.* 2005;16(1):70-76.
14. Maddox TM, Rumsfeld JS, Payne PRO. Questions for Artificial Intelligence in Health Care. *JAMA.* 2019;321(1):31-32.
15. Verghese A, Shah NH, Harrington RA. What This Computer Needs Is a Physician: Humanism and Artificial Intelligence. *JAMA.* 2018;319(1):19-20.
16. Bailey R, Kaskutas V, Fox I, Baum CM, Mackinnon SE. Effect of upper extremity nerve damage on activity participation, pain, depression, and quality of life. *J Hand Surg Am.* 2009;34(9):1682-1688.
17. Crichlow RJ, Andres PL, Morrison SM, Haley SM, Vrahas MS. Depression in orthopaedic trauma patients. Prevalence and severity. *J Bone Joint Surg Am.* 2006;88(9):1927-1933.
18. Jaquet JB, Kalmijn S, Kuypers PD, Hofman A, Passchier J, Hovius SE. Early psychological stress after forearm nerve injuries: a predictor for long-term functional outcome and return to productivity. *Ann Plast Surg.* 2002;49(1):82-90.

19. Nunez F, Vranceanu AM, Ring D. Determinants of pain in patients with carpal tunnel syndrome. *Clin Orthop Relat Res.* 2010;468(12):3328-3332.
20. Vranceanu AM, Jupiter JB, Mudgal CS, Ring D. Predictors of pain intensity and disability after minor hand surgery. *J Hand Surg Am.* 2010;35(6):956-960.
21. Teunis T, Janssen S, Guitton TG, Ring D, Parisien R. Do Orthopaedic Surgeons Acknowledge Uncertainty? *Clin Orthop Relat Res.* 2016;474(6):1360-1369.
22. Teunis T, Janssen SJ, Guitton TG, Vranceanu AM, Goos B, Ring D. Surgeon personality is associated with recommendation for operative treatment. *Hand (N Y).* 2015;10(4):779-784.
23. Janssen SJ, Teunis T, Guitton TG, Ring D, Science of Variation G. Do Surgeons Treat Their Patients Like They Would Treat Themselves? *Clin Orthop Relat Res.* 2015;473(11):3564-3572.

PART IV

3D Printing for Preoperative Planning

CHAPTER 6

Do 3-D Printed Handheld Models Improve Surgeon Reliability for Recognition of Intraarticular Distal Radius Fracture Characteristics

D.W.G. Langerhuizen
J.N. Doornberg
M.M.A. Janssen
G.M.M.J. Kerkhoffs
R.L. Jaarsma
S.J. Janssen

Clinical Orthopaedics & Related Research 2020 Dec;478(12):2901-2908

Commentary by:
C. L. Forthman in Clinical Orthopaedics & Related Research 2020 Dec;478(12):2909-2911.

ABSTRACT

Background

For fracture care, radiographs and two-dimensional (2-D) and three-dimensional (3-D) CT are primarily used for preoperative planning and postoperative evaluation. Intraarticular distal radius fractures are technically challenging to treat, and meticulous preoperative planning is paramount to improve the patient's outcome. Three-dimensionally printed handheld models might improve the surgeon's interpretation of specific fracture characteristics and patterns preoperatively and could therefore be clinically valuable; however, the additional value of 3-D printed handheld models for fractures of the distal radius, a high-volume and commonly complex fracture due to its intraarticular configuration, has yet to be determined.

Questions/purposes

(1) Does the reliability of assessing specific fracture characteristics that guide surgical decision-making for distal radius fractures improve with 3-D printed handheld models? (2) Does surgeon agreement on the overall fracture classification improve with 3-D printed handheld models? (3) Does the surgeon's confidence improve when assessing the overall fracture configuration with an additional 3-D model?

Methods

We consecutively included 20 intraarticular distal radius fractures treated at a Level 1 trauma center between May 2018 and November 2018. Ten surgeons evaluated the presence or absence of specific fracture characteristics (volar rim fracture, die punch, volar lunate facet, dorsal comminution, step-off > 2 mm, and gap > 2 mm), fracture classification according to the AO/Orthopaedic Trauma Association (OTA) classification scheme, and their confidence in assessing the overall fracture according to the classification scheme, rated on a scale from 0 to 10 (0 = not at all confident to 10 = very confident). Of 10 participants regularly treating distal radius fractures, seven were orthopaedic trauma surgeons and three upper limb surgeons with experience levels ranging from 1 to 25 years after completion of residency training. Fractures were assessed twice, with 1 month between each assessment. Initially, fractures were assessed using radiographs and 2-D and 3-D CT images (conventional assessment); the second time, the evaluation was based on radiographs and 2-D and 3-D CT images with an additional 3-D handheld model (3-D printed handheld model assessment). On both occasions, fracture characteristics were evaluated upon a surgeon's own interpretation, without specific instruction before assessment. We provided a sheet demonstrating the AO/OTA classification scheme before evaluation on each session. Multi-rater Fleiss's kappa was used to determine intersurgeon reliability for assessing fracture characteristics and

classification. Confidence regarding assessment of the overall fracture classification was assessed using a paired t-test.

Results

We found that 3-D printed models of intraarticular distal radius fractures led to no change in kappa values for the reliability of all characteristics: volar rim (conventional kappa 0.19 [95% CI 0.06 to 0.32], kappa for 3-D handheld model 0.23 [95% CI 0.11 to 0.36], difference of kappas 0.04 [95% CI -0.14 to 0.22]; p = 0.66), die punch (conventional kappa 0.38 [95% CI 0.15 to 0.61], kappa for 3-D handheld model 0.50 [95% CI 0.23 to 0.78], difference of kappas 0.12 [95% CI -0.23 to 0.47]; p = 0.52), volar lunate facet (conventional kappa 0.31 [95% CI 0.14 to 0.49], kappa for 3-D handheld model 0.48 [95% CI 0.23 to 0.72], difference of kappas 0.17 [95% CI -0.12 to 0.46]; p = 0.26), dorsal comminution (conventional kappa 0.36 [95% CI 0.13 to 0.58], kappa for 3-D handheld model 0.31 [95% CI 0.11 to 0.51], difference of kappas -0.05 [95% CI -0.34 to 0.24]; p = 0.74), step-off > 2 mm (conventional kappa 0.55 [95% CI 0.29 to 0.82], kappa for 3-D handheld model 0.58 [95% CI 0.31 to 0.85], difference of kappas 0.03 [95% CI -0.34 to 0.40]; p = 0.87), gap > 2 mm (conventional kappa 0.59 [95% CI 0.39 to 0.79], kappa for 3-D handheld model 0.69 [95% CI 0.50 to 0.89], difference of kappas 0.10 [95% CI -0.17 to 0.37]; p = 0.48). Although there appeared to be categorical improvement in kappa values for some fracture characteristics, overlapping CIs indicated no change. Fracture classification did not improve (conventional diagnostics: kappa 0.27 [95% CI 0.14 to 0.39], conventional diagnostics with an additional 3-D handheld model: kappa 0.25 [95% CI 0.15 to 0.35], difference of kappas: -0.02 [95% CI -0.18 to 0.14]; p = 0.81). There was no improvement in self-assessed confidence in terms of assessment of overall fracture configuration when a 3-D model was added to the evaluation process (conventional diagnostics 7.8 [SD 0.79 {95% CI 7.2 to 8.3}], 3-D handheld model 8.5 [SD 0.71 {95% CI 8.0 to 9.0}], difference of score: 0.7 [95% CI -1.69 to 0.16], p = 0.09).

Conclusions

Intersurgeon reliability for evaluating the characteristics of and classifying intraarticular distal radius fractures did not improve with an additional 3-D model. Further studies should evaluate the added value of 3-D printed handheld models for teaching surgical residents and medical trainees to define the future role of 3-D printing in caring for fractures of the distal radius.

INTRODUCTION

Orthopaedic trauma surgeons have conventionally used radiographs and two-dimensional (2-D) and three-dimensional (3-D) CT images for preoperative planning.[1] In general, 2-D and 3-D CT images have improved reliability for the classification of fractures compared with radiographs,[1-4] but classification remains an inconsistent exercise, often limited by substantial between-surgeon variation. Several studies reported poor reproducibility of fracture classifications and characterization, as well as large variation in subsequent treatment decision-making.[5-7] For distal radius fractures, a previous study showed that the AO classification using radiographs and 2-D and 3-D CT images among a large international group of surgeons resulted in substantial agreement for Type A fractures, but only fair agreement for Type B fractures and moderate agreement for Type C fractures.[8]

One could argue that for surgical decision-making, fracture characterization—the description of key elements of a fracture that might influence the decision of whether or how to perform surgery—is more important than fracture classification. For instance, 3-D CT images improve reliability and accuracy for recognizing specific characteristics of distal radial fractures.[1] In light of that, it is reasonable to wonder whether 3-D printed models of complex intraarticular fractures might be even more helpful; these models have become relatively inexpensive and relatively available. Prior studies have found that 3-D printed handheld models improved the reliability of classifying acetabular fractures and recognizing specific distal humerus and coronoid fracture characteristics.[9-11] For distal radius fractures, the additional clinical value of 3-D handheld models for patient-clinician communication have been reported, but they did not address characterization and classification.[12] To the best of our knowledge, there are no studies reporting on the additional preoperative clinical value of 3-D printed handheld models to recognize fracture characteristics and agree on the classification of intraarticular distal radius fractures. As so, the potential of 3-D handheld models for preoperative management, especially for a high-volume fracture that is often challenging due to its articular involvement, has yet to be determined.

Therefore, we asked: (1) Does the reliability of assessing specific fracture characteristics that guide surgical decision-making for distal radius fractures improve with 3-D printed handheld models? (2) Does surgeon agreement on the overall fracture classification improve with 3-D printed handheld models? (3) Does the surgeon's confidence improve when assessing the overall fracture configuration with an additional 3-D model?

PATIENTS AND METHODS

Study Design and Participants

This cross-sectional survey was approved by our institutional review board. We consecutively included 20 patients 18 years and older with an intraarticular distal radius fracture (AO Type 23 B1 to C3) who presented to a Level 1 trauma center between May 2018 and November 2018. We searched the picture archiving and communication system, using the term "CT wrist", to identify patients who had a wrist CT scan. Only patients with availability of corresponding radiographs and 2-D and 3-D CT images were selected. Ten surgeons, of whom seven were fellowship-trained orthopaedic trauma surgeons and three were upper-limb surgeons working at the same Level 1 trauma center from which the patients were selected, were asked to participate as observers. Among the surgeons, experience ranged between 1 and 25 years after completion of residency training. All observers determined six fracture characteristics and classified the fracture according to the AO/Orthopaedic Trauma Association (OTA) Fracture and Dislocation Classification.[13] Participants then indicated their overall confidence regarding the fracture's configuration. Fracture characteristics were evaluated upon a surgeon's own interpretation, without instruction before assessment. An information sheet demonstrating the AO/OTA classification scheme before assessment was provided. A surgeons' confidence was determined on an ordinal Likert scale ranging from 0 to 10.

Data were collected using standardized outcome sheets. Observations were made with 1 month between each assessment. The initial evaluation was based on radiographs and 2-D and 3-D CT images; the second evaluation was based on radiographs and 2-D and 3-D CT images with an additional 3-D printed handheld model.

Survey Design

The survey included 20 patients, and for each patient, we asked the surgeons the following questions: (1) Are there any of the following fracture characteristics? (Volar rim fracture, die punch, involvement of the volar lunate facet, dorsal comminution, step-off > 2 mm, and gap > 2 mm). (2) According to the AO/OTA's classification system,[14] how would you classify the fracture? (AO Type 23 B1 to C3). And, (3) On a scale from 0 to 10, how confident are you about the overall fracture according to the classification scheme (0 = not at all confident to 10 = very confident)?

We considered a die punch fracture as a central impacted intraarticular fragment. A volar rim fragment is distal from the watershed line where the radiolunate ligament originates. As such, it is located radially from the volar ulnar corner fragment.

Image Viewer and 3-D Printing

For all patients, the Digital Imaging and Communications in Medicine (DICOM) files from all radiographs and 2-D and 3-D CT images were loaded into Horos (version 3.3.4, Annapolis, MD, USA), an open-source medical imaging viewer (Figure 1).

An independent researcher who was not involved in the patients' care created all 3-D handheld models at our facility. Axial CT images with a slice thickness < 1.0 mm were obtained and saved as DICOM files and imported into 3-D Slicer (version 4.11.0, Boston, MA, USA). We used a threshold of 200 Hounsfield units to identify the distal radius and its fracture fragments because this resulted in minimal distortion from the surrounding soft tissue. All 3-D models were exported as surface tessellation language files into Ultimaker Cura software (version 3.6, Ultimaker BV, Geldermalsen, the Netherlands) for final preparation and subsequent conversion to G-code. The following pre-processing parameters were used: layer height, 0.15 mm; infill density, 100%; print speed, 50 mm/s; and extruder temperature 210° C. A 3-D printer (Ultimaker 2 +, Ultimaker BV) was used to create the 3-D handheld models on a 1:1 scale, with polylactic acid (PLA) as construction material. Total costs of 3-D handheld printed models was estimated at USD 10 per case (printing time about 4.5 hours [USD 2/hour overhead costs] and USD 0.50 PLA material cost per print).

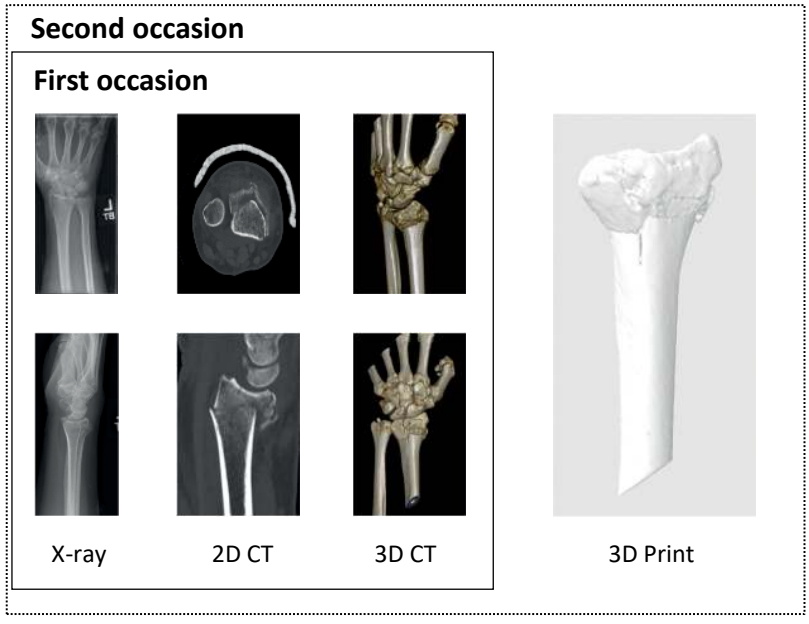


Figure 1: On the first occasion, radiographs, 2D- and 3D CT-scans were assessed on a 2D computer screen (black solid line). As part of standard protocol in our hospital, these images are obtained to optimize preoperative planning for fractures of the distal radius. For the second occasion, in addition to standard care, 3D hand-held models were printed with polylactic acid (PLA) as construction material (black dotted line).

Statistical Analysis

Multi-rater Fleiss's kappa was used to determine the intersurgeon agreement. The kappa value is a chance-corrected quantitative measure representing the degree to which observers agree with each other. To interpret kappa values, Landis and Koch[15] proposed the following: a kappa between 0.01 and 0.20 reflects slight agreement, between 0.21 and 0.40 reflects fair agreement, between 0.41 to 0.60 reflects moderate agreement, between 0.61 and 0.80 reflects substantial agreement, and greater than 0.81 reflects almost perfect agreement.

Bootstrapping (number of resamples: 1000) was used to calculate the standard error, z-statistic, 95% CI, and p values for the kappa values to compare groups. Differences in confidence in assessing fracture configuration with and without 3-D handheld model were compared using the paired t-test. A two-tailed p value less than 0.05 was considered significant. All statistical analyses were performed using Stata 15 (StataCorp LLC, College Station, TX, USA). There were no missing values for any of the collected variables.

RESULTS

3-D printed models (3-D hand held model assessment) of intraarticular distal radius fractures led to no improvement in kappa values for the reliability for all characteristics compared with radiographs with 2-D and 3-D CT images (conventional assessment): volar rim (conventional kappa 0.19 [95% CI 0.06 to 0.32], kappa for 3-D handheld model 0.23 [95% CI 0.11 to 0.36], difference of kappas 0.04 [95% CI 0.14 to 0.22]; p = 0.66), die punch (conventional kappa 0.38 [95% CI 0.15 to 0.61], kappa for 3-D handheld model 0.50 [95% CI 0.23 to 0.78], difference of kappas 0.12 [95% CI -0.23 to 0.47]; p = 0.52), volar lunate facet (conventional kappa 0.31 [95% CI 0.14 to 0.49], kappa for 3-D handheld model 0.48 [95% CI 0.23 to 0.72], difference of kappas 0.17 [95% CI -0.12 to 0.46]; p = 0.26), dorsal comminution (conventional kappa 0.36 [95% CI 0.13 to 0.58], kappa for 3-D handheld model 0.31 [95% CI 0.11 to 0.51], difference of kappas -0.05 [95% CI -0.34 to 0.24]; p = 0.74), step-off > 2 mm (conventional kappa 0.55 [95% CI 0.29 to 0.82], kappa for 3-D handheld model 0.58 [95% CI 0.31 to 0.85], difference of kappas 0.03 [95% CI -0.34 to 0.40]; p = 0.87), gap > 2 mm (conventional kappa 0.59 [95% CI 0.39 to 0.79], kappa for 3-D handheld model 0.69 [95% CI 0.50 to 0.89], difference of kappas 0.10 [95% CI -0.17 to 0.37]; p = 0.48) (Table 1).

The surgeons' agreement on the overall fracture classification demonstrated no improvement: kappa for conventional diagnostics, 0.27 (95% CI 0.14 to 0.39); kappa for conventional diagnostics with an additional 3-D handheld model, 0.25 (95% CI 0.15 to 0.35); difference of kappas -0.02 (95% CI -0.18 to 0.14; p = 0.81). There was no improvement in self-assessed confidence in terms of assessment of overall fracture configura-

Table 1. Intraobserver reliability for determining fracture characteristics

Characteristic	Examination 1: conventional imaging			Examination 2: conventional imaging with 3-D handheld models			Difference examination 1 and 2		
	Kappa	Category	95%CI	Kappa	Category	95%CI	Difference of kappa	95% CI	p value
Volar rim	0.19	Slight	0.06 to 0.32	0.23	Fair	0.11 to 0.36	0.04	-0.14 to 0.22	0.66
Die punch	0.38	Fair	0.15 to 0.61	0.5	Moderate	0.23 to 0.78	0.12	-0.23 to 0.47	0.52
Volar lunate facet	0.31	Fair	0.14 to 0.49	0.48	Moderate	0.23 to 0.72	0.17	-0.12 to 0.46	0.26
Dorsal comminution	0.36	Fair	0.13 to 0.58	0.31	Fair	0.11 to 0.51	-0.05	-0.34 to 0.24	0.74
Step-off > 2 mm	0.55	Moderate	0.29 to 0.82	0.58	Moderate	0.31 to 0.85	0.03	-0.34 to 0.40	0.87
Gap > 2 mm	0.59	Moderate	0.39 to 0.79	0.69	Substantial	0.50 to 0.89	0.1	-0.17 to 0.37	0.48

tion when a 3-D model was added to the evaluation process (conventional diagnostics 7.8 [SD 0.79] versus conventional diagnostics with an additional 3-D handheld model 8.5 [SD 0.71], mean difference: 0.70 [95% CI -1.69 to 0.16]; p = 0.09).

DISCUSSION

For distal radius fractures, radiographs in combination with 2-D and 3-D CT images are commonly used for preoperative planning.[1] Using the AO/OTA classification scheme, interobserver agreement among surgeons for distal radius fractures on radiographs and CT was substantial for Type A, fair for Type B, and moderate for Type C fractures.[8] Previous experience with 3-D printed handheld models demonstrated improved reliability among surgeons for classification of, for example, acetabular fractures.[10] The additional value of 3-D printed handheld models for preoperative assessment of intraarticular distal radius fractures is unclear. We found that 3-D printed models did not improve fracture characterization or confidence in the surgeon's assessment.

This study has several limitations. First, we only selected patients with a distal radius fracture who underwent CT, as this was necessary for subsequent 3-D printing. In general, patients who undergo CT are more likely to have relatively complex fracture patterns, resulting in a spectrum bias considering all distal radius fractures. However, we do not believe that including a different subset of patients with simpler fracture patterns will affect our results. We therefore consider this a minor limitation. Second, surgeons performed two evaluations, and might have remembered the first at the time of the second. However, we felt that one month between each assessment was sufficient to mitigate the risk of recall bias and therefore see this as a minor limitation.[1,16] Third, only 10 surgeons participated in our study. A higher number of observers might have resulted in us finding statistical differences between the first and second evaluations. However, the absolute difference in kappa values is small and so even if there were a statistical difference, we do not feel that it would be clinically relevant.

We found that an additional 3-D printed handheld model did not improve interobserver reliability for any of the included fracture characteristics. 3-D CT images improved assessment of fracture characteristics (such as, articular depression, fragment displacement, comminution) compared with 2-D CT images and radiographs in several studies.[1] However, there was no additional value of a 3-D printed handheld model versus 3-D CT images when assessing specific fracture characteristics. This finding was confirmed by three other studies that demonstrate no or only slight improvement of 3-D printed handheld models for assessment of specific fracture characteristics in coronoid, distal humerus, and radial head fractures over 3-D CT images (Table 2).[9,11,17]

Table 2. Studies evaluating the additional use of 3-D printed handheld models for fracture characteristics

Author, year	Anatomical location	Classification scheme	Observers	Number of observers	Kappa radiopgrahs and 2-D CT	Kappa radiographs, 2-D CT, and 3-D CT	Kappa 3-D printed handheld model	Kappa intra-operative view
Brouwer et al., 2012[9]	Distal humerus	Coronal fracture line	Surgeon, first assistant	2	0.43	0.24	0.46	0.53
	Distal humerus	> Three articular fragments	Surgeon, first assistant	2	0.49	0.6	0.67	0.66
	Distal humerus	Metaphyseal comminution	Surgeon, first assistant	2	0.63	0.71	0.71	0.66
	Distal humerus	Separated articular fragments	Surgeon, first assistant	2	0.47	0.41	0.41	0.32
	Distal humerus	Impaction	Surgeon, first assistant	2	0.14	0.26	0.21	0.29
Guitton et al., 2013[11]	Coronoid	Fracture of the anteromedial facet	Surgeon, first assistant	2	0.24	0.53	0.4	0.4
	Coronoid	Fracture of the tip of the coronoid	Surgeon, first assistant	2	0.08	0.35	0.66	0.44
	Coronoid	Comminuted fracture	Surgeon, first assistant	2	0.36	0.32	0.34	0.4
	Coronoid	Presence of impacted articular fragments	Surgeon, first assistant	2	0.16	0.32	0.52	0.26
	Coronoid	Subluxated/dislocated concentrically located elbow	Surgeon, first assistant	2	0.41	0.59	0.58	0.48
Guitton et al., 2014[17]	Radial head	Fracture line	Surgeons, first assistants	34	0.69	0.54	0.59	0.54
	Radial head	Comminution	Surgeons, first assistants	34	0.31	0.48	0.57	0.4
	Radial head	Articular surface	Surgeons, first assistants	34	0.12	0.27	0.22	0.28
	Radial head	Gap > 2 mm	Surgeons, first assistants	34	0.37	0.59	0.57	0.37
	Radial head	Impaction	Surgeons, first assistants	34	0.17	0.24	0.12	0.23
	Radial head	> Three fragments	Surgeons, first assistants	34	0.29	0.5	0.64	0.57
	Radial head	Small fragments	Surgeons, first assistants	34	0.26	0.34	0.33	0.42

Table 3. Studies evaluating the additional use of 3-D printed handheld models for fracture classification

Author, year	Anatomical fracture location	Classification scheme	Observers	Number of observers	Kappa radiographs and 2-D CT	Kappa radiographs, 2-D, and 3-D CT	Kappa 3-D printed handheld model	Kappa intra-operative view
Brouwer et al., 2012[9]	Distal humerus	AO classifcation	Surgeon, first assistant	2	0.74	0.78	0.74	0.64
Brouwers et al., 2018[10]	Acetabulum	Judet Letournel classification	Senior surgeon	7	0.33	0.42	0.59	
	Acetabulum	Judet Letournel classification	Junior surgeon	5	0.18	0.42	0.59	
	Acetabulum	Judet Letournel classification	Senior surgical resident	5	0.17	0.43	0.66	
	Acetabulum	Judet Letournel classification	Junior surgical resident	5	0.19	0.37	0.51	
	Acetabulum	Judet Letournel classification	Intern	5	0.16	0.38	0.61	
Guitton et al., 2013[11]	Coronoid	Mayo classification	Surgeon, first assistant	2	0.17	0.4	0.4	0.41
Guitton et al., 2014[17]	Radial head	Broberg Morrey	Surgeons, first assistants	34	0.23	0.26	0.37	0.38
Misselyn et al., 2018[18]	Calcaneus	Sanders classification	Foot and ankle surgeons	6	0.45		0.67	
	Calcaneus	Sanders classification	Trainee-surgeons, radiologists	18	0.26		0.47	

Assessment of the AO classification type did not improve with the additional use of 3-D handheld models in our study. In line with a specific fracture characteristic assessment, there seemed to be an added value of 3-D CT images over 2-D CT images and radiographs.[2-4] However, the added value of 3-D printed handheld models in addition to 3-D CT images for fracture classification is debatable, as prior evidence contradicts our finding of no difference (Table 3). For instance, 3-D handheld models versus 3-D CT images did improve fracture classification of acetabular fractures (difference of kappas 0.17).[18] Possible explanations for this contradicting finding are: (1) evaluation of a different anatomical area with its own complexities and classification scheme (Judet-Letournel[19]) and (2) a less frequently occurring fracture resulting in smaller caseload and, therefore, less experience. In a study on intraarticular calcaneal fractures, 3-D handheld models improved classification reliability (difference of kappas 0.22) using the Sanders classification.[20] However, this study compared 3-D handheld models only to 2-D CT coronal images, leaving other 2-D CT reconstructions and 3-D CT images out of comparison.[18]

In line with previous research, confidence regarding the overall fracture configuration did not substantially change with an additional 3-D handheld model.[12] It is reasonable to assume that confidence expressed by our experienced fellowship-trained surgeons was already high at baseline, regardless of the imaging technique used, leaving relatively little room for improvement with an additional 3-D printed handheld model.

There is a possibility that 3-D printed handheld models may have greater potential for less-experienced surgeons because these models help visualize fracture patterns in a more realistic 3-D view than conventional imaging. Prior studies found that reliability of fracture classification improved for less-experienced surgeons in training (junior residents and medical trainees) compared with fellowship trained orthopaedic surgeons when assessing acetabular and calcaneal fracture classification.[10,18] One can imagine sitting with a resident to discuss the surgical plan for approaching, reducing, and fixing certain fracture fragments based on a 3-D handheld model. Interestingly, operative time for intraarticular distal radius fractures reduced when 3-D models were provided preoperatively, and, although we did not specifically evaluate, several participants of our study believed that the second assessment (with a 3-D printed handheld model) was faster.[21]

At our institution, PACS automatically generates 3-D CT by reconstructing axial 2-D CT slices. Surgeons use these 3-D CT images together with radiographs and 2-D CT as standard care, without additional costs. The implementation of 3-D handheld models into clinical practice is relatively easy and straightforward nowadays as the data is available via standard care (DICOM files from the 2-D CT scans) and the software for pre-processing is open-access and intuitive to use. In addition to routinely obtained preoperative imaging for intraarticular distal radius fractures, the costs associated with 3-D printed hand

models are about USD 10 each. Despite this, our study demonstrated no clear advantage of 3-D handheld models in conjunction with radiographs, 2-D and 3-D CT images and we therefore deem implementation of 3-D models in clinical practice unnecessary.

In conclusion, this study showed that 3-D handheld models do not improve reliability for assessing and classifying intraarticular distal radius fractures. Further studies should seek to evaluate the added value of 3-D handheld models for teaching surgical residents and medical trainees to define the future role of 3-D printing in caring for distal radius fractures.

ACKNOWLEDGMENTS

We thank the following participants: J.H. Pot MD, B. Jadav MD, J. Nestorson, MD, PhD, M.M.A. Janssen MD, PhD, R. L. Jaarsma MD, PhD, FRACS, B. J. Duijnisveld MD, PhD, J. W. White MBBS, PhD, O. L. Leonardsson MD, PhD, R. S. Strasser MD, G. I. Bain MD, PhD. We thank K. Denk MSc, for administrative and logistical support.

REFERENCES

1. Harness NG, Ring D, Zurakowski D, Harris GJ, Jupiter JB. The influence of three-dimensional computed tomography reconstructions on the characterization and treatment of distal radial fractures. *J Bone Joint Surg Am.* 2006;88(6):1315-1323.
2. Brunner A, Honigmann P, Treumann T, Babst R. The impact of stereo-visualisation of three-dimensional CT datasets on the inter- and intraobserver reliability of the AO/OTA and Neer classifications in the assessment of fractures of the proximal humerus. *J Bone Joint Surg Br.* 2009;91(6):766-771.
3. 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(8):1795-1801.
4. Foroohar A, Tosti R, Richmond JM, Gaughan JP, Ilyas AM. Classification and treatment of proximal humerus fractures: inter-observer reliability and agreement across imaging modalities and experience. *J Orthop Surg Res.* 2011;6:38.
5. Doornberg JN, Guitton TG, Ring D. Diagnosis of elbow fracture patterns on radiographs: interobserver reliability and diagnostic accuracy. *Clin Orthop Relat Res.* 2013;471(4):1373-1378.
6. Ghoshal A, Enninghorst N, Sisak K, Balogh ZJ. An interobserver reliability comparison between the Orthopaedic Trauma Association's open fracture classification and the Gustilo and Anderson classification. *Bone Joint J.* 2018;100-b(2):242-246.
7. Neuhaus V, Bot AG, Guitton TG, et al. Scapula fractures: interobserver reliability of classification and treatment. *J Orthop Trauma.* 2014;28(3):124-129.
8. Jayakumar P, Teunis T, Gimenez BB, Verstreken F, Di Mascio L, Jupiter JB. AO Distal Radius Fracture Classification: Global Perspective on Observer Agreement. *J Wrist Surg.* 2017;6(1):46-53.
9. Brouwer KM, Lindenhovius AL, Dyer GS, Zurakowski D, Mudgal CS, Ring D. Diagnostic accuracy of 2- and 3-dimensional imaging and modeling of distal humerus fractures. *J Shoulder Elbow Surg.* 2012;21(6):772-776.
10. Brouwers L, Pull ter Gunne AF, de Jongh MAC, et al. The Value of 3D Printed Models in Understanding Acetabular Fractures. *3D Printing and Additive Manufacturing.* 2018;5(1):37-46.
11. Guitton TG, Kinaci A, Ring D. Diagnostic accuracy of 2- and 3-dimensional computed tomography and solid modeling of coronoid fractures. *J Shoulder Elbow Surg.* 2013;22(6):782-786.
12. Bizzotto N, Tami I, Tami A, et al. 3D Printed models of distal radius fractures. *Injury.* 2016;47(4):976-978.
13. Marsh JL, Slongo TF, Agel J, et al. Fracture and dislocation classification compendium - 2007: Orthopaedic Trauma Association classification, database and outcomes committee. *J Orthop Trauma.* 2007;21(10 Suppl):S1-133.
14. Müller M. *The Comprehensive Classification of Fractures of Long Bones.* New York, NY: Springer-Verlag. 1990.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159-174.
16. Clarke PM, Fiebig DG, Gerdtham UG. Optimal recall length in survey design. *J Health Econ.* 2008;27(5):1275-1284.
17. Guitton TG, Brouwer K, Lindenhovius AL, et al. Diagnostic accuracy of two-dimensional and three-dimensional imaging and modeling of radial head fractures. *J Hand Microsurg.* 2014;6(1):13-17.

18. Misselyn D, Nijs S, Fieuws S, Shaheen E, Schepers T. Improved Interobserver Reliability of the Sanders Classification in Calcaneal Fractures Using Segmented Three-Dimensional Prints. *J Foot Ankle Surg.* 2018;57(3):440-444.
19. Judet R, Judet J, Letournel E. Fractures of the Acetabulum: Classification and Surgical Approaches for Open Reduction. Preliminary Report. *J Bone Joint Surg Am.* 1964;46:1615-1646.
20. Sanders R, Fortin P, DiPasquale T, Walling A. Operative treatment in 120 displaced intraarticular calcaneal fractures. Results using a prognostic computed tomography scan classification. *Clin Orthop Relat Res.* 1993(290):87-95.
21. Chen C, Cai L, Zheng W, Wang J, Guo X, Chen H. The efficacy of using 3D printing models in the treatment of fractures: a randomised clinical trial. *BMC Musculoskelet Disord.* 2019;20(1):65.

PART V

3D Fluoroscopy for Intraoperative Assessment

CHAPTER 7

Diagnosis of Dorsal Screw Penetration After Volar Plating of a Distal Radial Fracture: Intraoperative Dorsal Tangential Views Versus 3D Fluoroscopy

D.W.G. Langerhuizen
M. Bergsma
C.A. Selles
R.L. Jaarsma
J.C. Goslings
N.W.L. Schep
J.N. Doornberg

The Bone & Joint Journal 2020 Jul;102-B(7):874-880.

ABSTRACT

Aims

The aim of this study was to investigate whether intraoperative 3D fluoroscopic imaging outperforms dorsal tangential views in the detection of dorsal cortex screw penetration after volar plating of an intra-articular distal radial fracture, as identified on postoperative CT imaging.

Methods

A total of 165 prospectively enrolled patients who underwent volar plating for an intra-articular distal radial fracture were retrospectively evaluated to study three intraoperative imaging protocols: 1) standard 2D fluoroscopic imaging with anteroposterior (AP) and elevated lateral images (n = 55); 2) 2D fluoroscopic imaging with AP, lateral, and dorsal tangential views images (n = 50); and 3) 3D fluoroscopy (n = 60). Multiplanar reconstructions of postoperative CT scans served as the reference standard.

Results

In order to detect dorsal screw penetration, the sensitivity of dorsal tangential views was 39% with a negative predictive value (NPV) of 91% and an accuracy of 91%; compared with a sensitivity of 25% for 3D fluoroscopy with a NPV of 93% and an accuracy of 93%. On the postoperative CT scans, we found penetrating screws in: 1) 40% of patients in the 2D fluoroscopy group; 2) in 32% of those in the 2D fluoroscopy group with AP, lateral, and dorsal tangential views; and 3) in 25% of patients in the 3D fluoroscopy group. In all three groups, the second compartment was prone to penetration, while the postoperative incidence decreased when more advanced imaging was used. There were no penetrating screws in the third compartment (extensor pollicis longus groove) in the 3D fluoroscopy groups, and one in the dorsal tangential views group.

Conclusion

Advanced intraoperative imaging helps to identify screws which have penetrated the dorsal compartments of the wrist. However, based on diagnostic performance characteristics, one cannot conclude that 3D fluoroscopy outperforms dorsal tangential views when used for this purpose. Dorsal tangential views are sufficiently accurate to detect dorsal screw penetration, and arguably more efficacious than 3D fluoroscopy.

INTRODUCTION

Open reduction and internal fixation (ORIF) with a volar plate is increasingly being used for the treatment of a distal radial fracture.[1-7] One potential technical error is protrusion into the dorsal compartments with penetrating screws which are obscured on lateral fluoroscopic imaging by Lister's tubercle, with the risk of extensor tendon irritation and rupture.[8] Intraoperative dorsal tangential views, in which the forearm is placed in 75° inclination to the operating table with the wrist in flexion, have been shown to be a promising technique for avoiding dorsal cortical screw penetration (Figure 1).[9-11] Several cadaveric and preclinical studies have shown the accurate identification of screw penetration using this technique.[12,13] Subsequently, Ganesh et al[14] were the first to report the incidence of dorsal cortical penetration by screws using postoperative CT scans as the reference standard, addressing the accuracy of dorsal tangential views. In their pilot study, the incidence of penetrating screws on postoperative CT scans was 17% (five of 30 patients).

In orthopaedic trauma, previous authors have described increased rates of revision intraoperative fixation using advanced 3D fluoroscopic imaging, reducing rates of further revision surgery compared with conventional 2D fluoroscopy.[15-17] Mehling et al[18] described revision of misplaced screws after volar plating for 51 patients with a distal radial fracture. In one-third of patients, screws were found to be too long, too radial, or intra-articularly placed on intraoperative dorsal tangential views and these screws were not detected with conventional 2D fluoroscopy.

For the purpose of detecting dorsal penetration, we wished to know whether 3D fluoroscopy adds value compared with dorsal tangential views. Therefore, the aim of this retrospective study of three prospective cohorts of 165 patients treated with volar plating for a distal radial fracture was to determine whether intraoperative 3D fluoroscopy would be preferred to dorsal tangential views in the identification of screw penetration with postoperative CT imaging as the reference standard. Specifically, we compared 1) the diagnostic performance of dorsal tangential views versus 3D fluoroscopy; 2) the incidence of postoperative screw penetration in three groups of intraoperative imaging using a) conventional 2D fluoroscopy, b) 2D fluoroscopy with dorsal tangential views, and c) 3D fluoroscopy; 3) specific compartments at risk using the different imaging modalities; and 4) the difference in number of penetrating screws in different patterns of fracture (AO-type 23).[19]

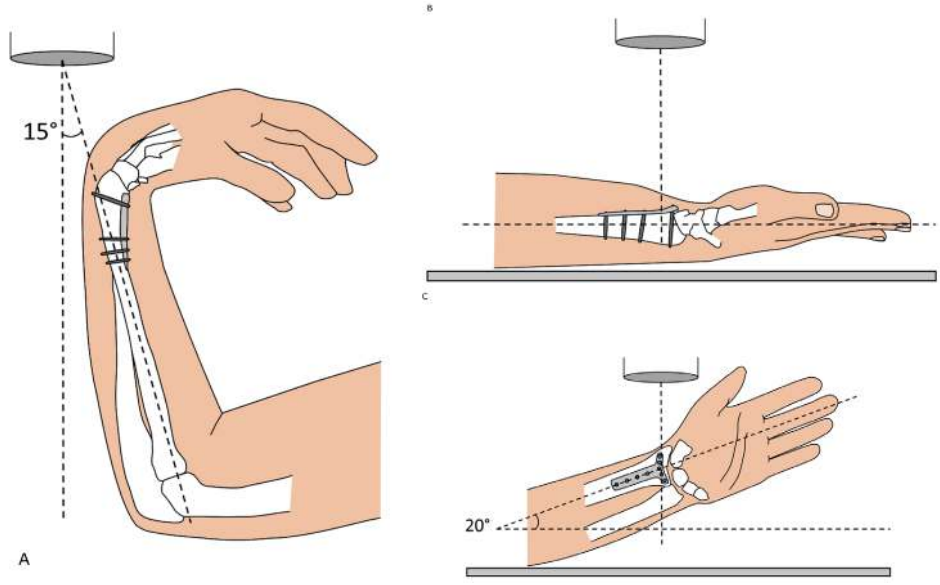


Figure 1: Diagrams of a) the dorsal tangential view, in which the arm is intraoperatively placed in 75° inclination to the operating table with the wrist in flexion, enables inspection of the dorsal cortex of the distal radius; b) the anteroposterior view, with the arm on the table and the volar aspect upward perpendicular to the radiograph beam; c) the elevated lateral view, elevating the wrist 20° to enhance visualization of the radio-carpal joint.

METHODS

In accordance with the Declaration of Helsinki, we retrospectively reviewed prospectively collected postoperative CT scans of two prospective trials of adult patients with an intra-articular distal radial fracture with the approval of our Institutional Review Boards.[20,21] This study was designed as a multicentre prospective matched cohort study, with retrospective image analysis of postoperative CT scans as the reference standard.

Study design – prospective matched cohort study

In the initial study, patients with fractures of the calcaneum, ankle, tibial pilon, and distal radius undergoing ORIF were included in a prospective multicentre randomized clinical trial (the EF3X- trial) investigating the effectiveness of the intraoperative use of advanced 3D versus 2D fluoroscopy.[20] The main aim was to evaluate the quality of reduction and fixation of the fracture, with postoperative CT scans serving as the reference standard. We prospectively included patients that underwent volar plating of the distal radius. The analysis of patients with other extremity fractures, as well as the initial research question (i.e. quality of reduction), will be the subject of future publications.

In the second study, patients with an intra-articular distal radial fracture were prospectively included in a single Level-1 centre cohort study to evaluate the diagnostic performance characteristics of dorsal tangential views to detect dorsal screw penetration after volar plating, with postoperative CT as the reference standard. The current study extends this study to allow comparison of dorsal screw penetration in patients that did not have intraoperative dorsal tangential views, versus intraoperative dorsal tangential views, versus intraoperative 3D fluoroscopy, to evaluate the diagnostic performance of dorsal tangential views.[21]

In this study, three prospective cohorts of patients with a distal radial fracture undergoing ORIF from both prospective trials were combined to allow evaluation of the different intraoperative imaging strategies (Table 1): standard 2D fluoroscopy with anteroposterior (AP) and elevated lateral images (n = 55); 2D fluoroscopy with AP, elevated lateral, and dorsal tangential views (n = 50); and 3D fluoroscopy (n = 60).

Table 1. Patient characteristics of the three groups.

Characteristic	Cohort 1: 2DF	Cohort 2: 2DF + DTV	Cohort 3: 3DF
Patients, n	55	50	60
Mean age, yrs (range)	56 (24 to 76)	57 (18 to 87)	56 (22 to 79)
Sex, n (%)			
Male	18 (33)	16 (32)	24 (40)
Female	37 (67)	34 (68)	36 (60)
Side of fracture, n (%)			
Left	31 (56)	20 (40)	39 (65)
Right	24 (44)	30 (60)	21 (35)
AO/OTA-type 23, n (%)			
A	8 (15)	N/A	4 (7)
B	10 (18)	16 (32)	9 (15)
C	37 (67)	34 (68)	47 (78)

*DF, 2D fluoroscopy; 3DF, 3D fluoroscopy; DTV, dorsal tangential view; N/A, not applicable; OTA, Orthopaedic Trauma Association.

Patient cohort 1 (2D fluoroscopy) and cohort 3 (3D fluoroscopy)

From the first study, two groups were included: the 2D and 3D fluoroscopy groups. Out of a total 206 patients with a distal radial fracture, 103 patients were allocated to 2D and 103 patients to 3D fluoroscopy. For the purpose of this study, we excluded 91 patients: 21 without a volar plate, 29 with additional dorsal and/or lateral plates that obscured dorsal screw penetration on postoperative CT, 29 with a postoperative CT scan of insufficient quality to serve as the reference standard, and 12 without a postoperative reference CT scan.

We included the 115 patients who underwent volar plating for an intra-articular distal radial fracture: 55 were randomized to intraoperative 2D fluoroscopy, and 60 to 3D fluoroscopy. All patients were treated by or under the supervision of a senior orthopaedic or trauma consultant at one of the participating hospitals, between October 2009 and July 2014. A volar approach through the bed of flexor carpi radialis was used to expose the radius, as a modified Henry approach.[22] Volar locking plates were used (locking compression plates (LCPs) 2.4 mm[23] and variable angle VA-LCPs 2.4 mm[24] (DePuy Synthes, Oberdorf, Switzerland)). In the 2D fluoroscopy group, AP and elevated lateral views were used intraoperatively at the surgeons' discretion. Dorsal tangential views were not part of hospital protocols and not used in the respective surgeons' practice. Therefore, the 2D fluoroscopy group served as the reference.

Patient cohort 2 (2D fluoroscopy with dorsal tangential views)

From the second initial study, we included 50 prospectively enrolled patients who were treated with a VA-LCP (Synthes, North Ryde, Australia) for an intra-articular distal radial fracture at our Level-1 Trauma Centre, between May 2017 and August 2018, for retrospective use of this prospective data.[21]

Overall CT reference standard: assessment of dorsal cortex screw penetration

We defined dorsal penetration as screws penetrating by ≥ 0.5 mm.[25] All patients had a postoperative CT scan of the wrist within one week with a slice thickness of < 1 mm (Somatom Definition AS+, Siemens, Erlangen, Germany). The scans were obtained in an axial plane and saved as Digital Imaging and Communications in Medicine (DICOM) files. We created triplanar reconstruction in OsiriX lite version 9 (open-source software; Pixmeo, Geneva, Switzerland)[26] with an adjusted axial plane parallel to the screw. We only evaluated the most distal row of the locking plate as we consider that the anatomical space for the extensor tendons was the most limited in this region. Two observers (DWGL, MB), who were not involved in the patients' care, independently evaluated each screw for 1) penetration of the dorsal cortex, 2) the total length of the screw, 3) the amount of dorsal penetration in millimetres (mm), and 4) the anatomical location of the violated compartment.

Measurements were obtained by following these steps (Figure 2): 1) The axial, sagittal, and coronal planes were adjusted parallel to each respective distal angular stable penetrating screw. 2) In the axial plane, we measured the total length of the dorsal penetrating screw from its head to its tip. 3) We constructed a line at each side of the penetrating screw. 4) We determined the distance of penetration from the tip of the screw to the dorsal cortex by measuring the distance from the tip of the screw to this line. A video on the following URL demonstrates the measurement technique: www.traumaplatform.org/currentprojects.

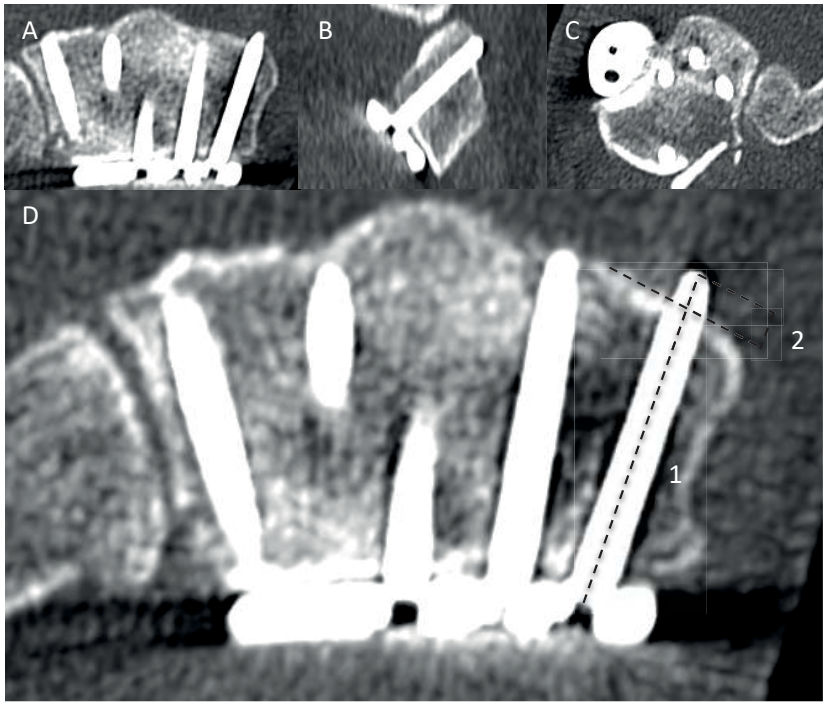


Figure 2: Radiographs of triplanar reconstructions with the a) axial, b) sagittal, and c) coronal planes adjusted parallel to the distal angular stable penetrating screw. Measurements were performed in the adjusted axial plane parallel to the penetrating screw in the most radial position (i.e. second compartment). d) Number 1 is the line on which the total length of the penetrating screw was measured. Line number 2 represents the penetrating distance of the screw.

Statistical analysis

Two independent observers not involved in patient care (DWGL, MB) conducted initial measurements in a set of 20 randomly selected cases in order to assess the interobserver reliability of the new CT measurement technique. We used Kappa, a quantitative measure accounting for agreement by chance among observers, to assess the interobserver agreement for penetration of the screw and the linear-weighted Kappa for the total length of a screw. Confidence intervals (CIs) were calculated by using the standard error. According to Landis and Koch,[27] the Kappa for dorsal screw penetration and the linear-weighted Kappa for the total length of a screw was almost perfect: 0.84 (95% CI 0.74 to 0.94) and 0.82 (95% CI 0.70 to 0.95). A Kappa above 0.81 indicates almost perfect agreement. Interobserver agreement of screw penetration was calculated using an intraclass correlation coefficient (ICC) with a two-way random-effects model with absolute agreement to assess how much each measurement differed from that of the other observer. The ICC for the distance of dorsal penetration was excellent: 0.96 (95% CI 0.95 to 0.97).[28]

Patient characteristics were summarized with frequencies and percentages for categorical variables, and mean and range for continuous variables. Diagnostic performance characteristics were calculated according to standard formulae. Sensitivity applied to the proportion of actual correctly identified positives. Accuracy corresponded to the proportion of correctly predicted penetrating and non-penetrating screws over all the measured screws. The negative predictive value (NPV) was defined as the probability of not having a penetrating screw when it was not detected intraoperatively. Using analysis of one proportion, CIs were calculated to compare sensitivity, accuracy, and the NPV for the different imaging strategies. Overlap between lower and upper boundaries of the respective 95% CIs indicate no significant difference.

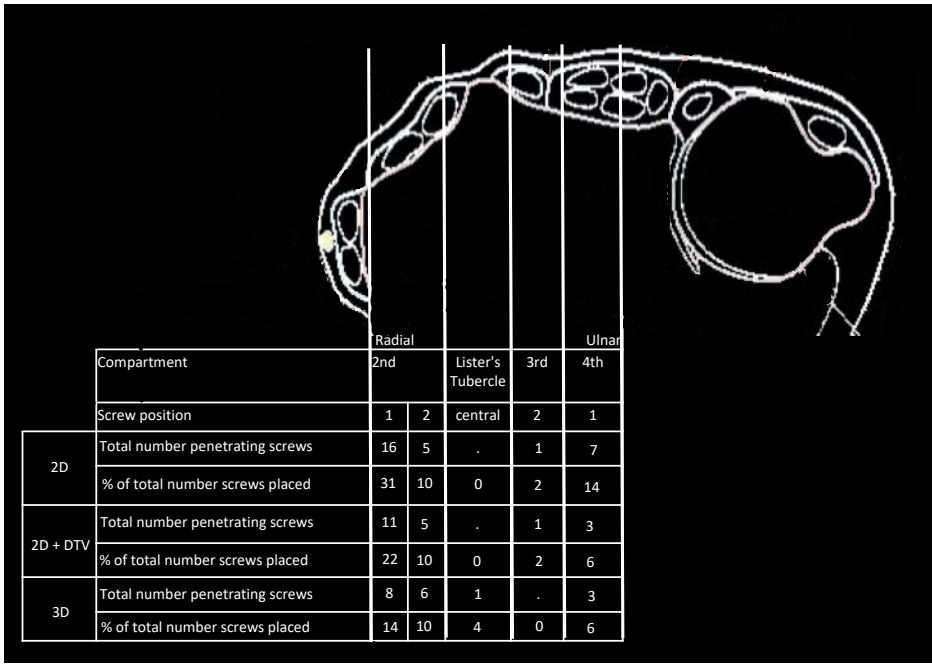
We used a chi-squared test for linear trend to compare fracture patterns (i.e. AO-type 23) among patients either with or without at least one dorsally penetrating screw. We determined it clinically relevant for surgeons to be able to reduce the incidence of dorsally penetrating screws from one in three patients as reported in the current literature (32%) with use of dorsal tangential views, to less than one in ten patients with use of 3D fluoroscopy (< 10%).[29] In order to prove this clinically relevant decrease statistically, power calculations showed that 47 patients were needed in each group to achieve 80% power ($\alpha = 0.05$, $\beta = 0.20$). All analyses were performed using Stata 15 (StataCorp, College Station, Texas, USA). Statistical significance was set at $p < 0.05$.

RESULTS

Patient cohort 1 (2D fluoroscopy): baseline incidence of postoperative screw penetration

Without the routine use of intraoperative dorsal tangential views or 3D fluoroscopy, 40% of patients (22/55) had at least one dorsal screw penetrating (≥ 0.5 mm) on postoperative CT imaging, and in 13% (7/55) two screws were penetrating. In total, 29/225 screws (13%) were penetrating with a mean distance of penetration of 1.1 mm (0.6 to 4.9) and a median length of 20 mm (interquartile range (IQR) 18 to 22).

The position of the screws at risk for dorsal penetration were 16/29 screws (55%) in the most radial position (second compartment), 5/29 (17%) in the second most radial position (second compartment), 7/29 (17%) in the most ulnar position (fourth compartment), 1/29 (3%) in the second most ulnar position (third compartment), and none in the central position (Lister's Tubercle, in plates with five holes) (Figure 3).



	Compartment	2nd		Lister's Tubercle	3rd	4th
	Screw position	1	2	central	2	1
2D	Total number penetrating screws	16	5	.	1	7
	% of total number screws placed	31	10	0	2	14
2D + DTV	Total number penetrating screws	11	5	.	1	3
	% of total number screws placed	22	10	0	2	6
3D	Total number penetrating screws	8	6	1	.	3
	% of total number screws placed	14	10	4	0	6

Figure 3: Diagram showing that screws in the second compartment were at highest risk of being too long, followed by the fourth, third, and central compartments.

Patient cohort 2 (2D fluoroscopy with dorsal tangential views)

The sensitivity of intraoperative dorsal tangential views to detect penetrating screws was 39% (95% CI 22 to 55) with a NPV of 91 (95% CI 88 to 96) and an accuracy of 91% (95% CI 89 to 95; Table 2).

With the routine use of dorsal tangential views, 32% of patients (16/50) had at least one screw penetrating on CT imaging, whereas in 4% (2/50) two screws were penetrating, and in 4% (2/50) three screws were penetrating. In total, 20/218 screws (9%) were penetrating with a mean distance of penetration of 1.5 mm (0.5 to 4.5) and a median length of 21 mm (IQR 18 to 24). Of the penetrating screws; 11/20 screws (55%) were in the most radial position (second compartment), 5/20 (25%) were in the second most radial position (second compartment), 3/20 (15%) were in the most ulnar position (fourth compartment), 1/20 (5%) was in the second most ulnar position (i.e. third compartment), whereas none were in the central position (in plates with five holes). Additionally, 2/218 screws (1%) were placed intra-articularly (one in the radio-carpal joint and one in the distal radial ulnar joint).

Table 2. Diagnostic performance characteristics per imaging strategy. Cohort 1, the conventional 2D fluoroscopy cohort (i.e. the baseline cohort) was not included in the table.

Diagnostic performance characteristics	Cohort 2: 2DF + DTV	Cohort 3: 3DF
Patients, n	50	60
Sensitivity* (95% CI)	39 (22 to 55)	25 (8 to 42)
Negative predictive value* (95% CI)	91 (88 to 96)	93 (90 to 96)
Accuracy* (95% CI)	91 (89 to 95)	93 (90 to 96)

*2DF, 2D fluoroscopic; 3DF, 3D fluoroscopic; CI, confidence interval; DTV, dorsal tangential views.

Patient cohort 3 (3D fluoroscopy)

The sensitivity of intraoperative 3D fluoroscopy was 25% (95% CI 8 to 42) with a NPV of 93% (95% CI 90 to 96) and an accuracy of 93% (95% CI 90 to 96).

With intraoperative 3D fluoroscopy, 25% of patients (15/60) had at least one screw penetrating on CT imaging, whereas in 5% (3/60) two screws were penetrating. In total, 18/248 (7%) were penetrating with a mean distance of penetration of 1.6 mm (0.7 to 4.0) and a median length of 20 mm (IQR 18 to 20).

Advanced imaging with fluoroscopy showed penetration in the following positions: 8/18 screws (44%) in the most radial position (second compartment), 6/18 screws (33%) in the second most radial position (second compartment), 3/18 screws (17%) in the most ulnar position (fourth compartment), and 1/18 screw (5%) in the central position. No screws were penetrating in the second most ulnar position (third compartment).

Influence of fracture patterns on postoperative screw penetration

Fracture patterns, assessed with the AO-type 23 A to C, were equally distributed (p = 0.467, chi-squared test) among patients with or without at least one penetrating screw (Table 3).

Table 3. Patients with or without at least one penetrating screw.

Penetrating screw	AO-type 23A	AO-type 23B	AO-type 23C	p-value*
Patients, n	12	35	118	
No penetrating screw, n (%)	9 (8)	23 (22)	75 (70)	0.467
≥ 1 penetrating screw, n (n%)	3 (5)	12 (21)	43 (74)	0.467

*Chi-squared test

DISCUSSION

Open reduction and internal fixation with a volar approach is commonly used in the treatment of distal radial fractures.[1-7] Based on diagnostic performance characteristics and the incidence of screw penetration in the dorsal compartments, one can conclude

that the use of 3D fluoroscopy is not better than dorsal tangential views. Moreover, dorsal tangential views are arguably more efficacious for the purpose of detecting penetration than advanced 3D fluoroscopy when taking into account the use of resources for intraoperative 3D imaging.

This study has strengths and weaknesses. A strength is the use of a new reliable CT-measurement technique to evaluate and measure dorsal cortex screw penetration as the reference standard by which to compare the diagnostic performance characteristics of 3D fluoroscopy and dorsal tangential views. Weaknesses include the fact that the study was designed as an imaging study, thus lacking clinical data about the incidence of extensor tendon related complications such as tenosynovitis and tendon rupture. Secondly, it was a case control study of prospective cohorts rather than a prospective RCT comparing the use of dorsal tangential views with 3D fluoroscopy.

Sensitivity tended to increase for dorsal tangential views compared with 3D fluoroscopy (39% vs 25%), while the accuracy was high and similar between both groups (92% vs 93%). We suggest that the difference in sensitivity between the groups may be due to surgeons being familiar with dorsal tangential views as part of our hospital protocol, while the intraoperative use of advanced 3D fluoroscopy was not routine for most surgeons. When compared with the existing literature, previous studies involving dorsal tangential views reported sensitivities ranging from 58% to 70%, while, to the best of our knowledge, no authors have described the sensitivity for 3D fluoroscopy in the identification of dorsal screw penetration.[9,30,31]

The incidence of screw penetration was 25% for 3D fluoroscopy versus 32% in the dorsal tangential views group, and compared with a baseline incidence of 40% in the conventional 2D fluoroscopy group. This is comparable to one other study. A prospective trial evaluated the incidence of unrecognized dorsal screw penetration on postoperative CT scans after the intraoperative use of 2D fluoroscopy without dorsal tangential views or 3D fluoroscopy.[32] The authors found penetrating screws in 37% of 30 patients. In contrast to studies using advanced imaging, Ganesh et al[14] found prominent screw tips in 17% of 26 patients with ≥ 1 mm penetration on postoperative CT imaging with the intraoperative use of dorsal tangential views. Although our results suggest that screw penetration can be reduced, and the incidence was lowest in the most advanced imaging 3D fluoroscopy group, still one in four patients had prominent dorsal screw tips in an extensor compartment. One could argue that using 3D fluoroscopy is time-consuming and expensive compared with dorsal tangential views, and requires a trained team to obtain the 3D images intraoperatively.

The second compartment was at the highest risk for screw penetration; however, while using more advanced imaging the number of penetrating screws decreased. Additionally, no protruding screws remained in the third compartment (extensor pollicis longus (EPL) groove) with 3D fluoroscopy, whereas there was one penetrating screw in

the dorsal tangential views group (2%). This is in line with previous studies, in which the second compartment was the most commonly violated, followed by the third and fourth compartments.[25,32] Despite the high number of penetrating screws in the radial compartment, the slope on the dorsal aspect of the radius provides some room for error, as reports on tendon ruptures in this compartment are scarce.[33] In contrast, EPL in the third compartment might be more easily injured due to the small space and the narrow tendon sheath.[34] The incidence of extensor tendon related complications due to penetrating screws varies between 0% and 30%, perhaps leaving most penetrating screws being asymptomatic.[25,32,35,36] However, dorsal penetration can be avoided by a combination of meticulous technique, including subtracting 2 mm from the measured depth, and the use of correct imaging strategies such as dorsal tangential views. Given the additional costs and higher one-year complication rates associated with operative management compared with conservative management in older adults, one should aim to avoid the morbidity of extensor tendinitis, and surgical intervention for late rupture.[2,37]

Finally, although there is evidence that comminution of the dorsal cortex may hinder accurate intraoperative screw assessment, our findings reflect the fact that more severe fractures did not significantly impede the identification of dorsal cortical penetration ($p = 0.467$, chi-squared test).[26]

In conclusion, this study supports the use of dorsal tangential views to minimize dorsal penetrating screws after volar plating. One could argue that 3D fluoroscopy is not required to be part of a surgeon's armamentarium to avoid screw penetration, as it did not improve the diagnostic performance, while implementing this technique in the daily routine may be labour-intensive and expensive.

REFERENCES

1. Costa ML, Achten J, Rangan A, Lamb SE, Parsons NR. Percutaneous fixation with Kirschner wires versus volar locking-plate fixation in adults with dorsally displaced fracture of distal radius: five-year follow-up of a randomized controlled trial. *Bone Joint J.* 2019;101-B(8):978-983.
2. DeGeorge BR, Jr., Van Houten HK, Mwangi R, Sangaralingham LR, Larson AN, Kakar S. Outcomes and Complications in the Management of Distal Radial Fractures in the Elderly. *J Bone Joint Surg Am.* 2020;102(1):37-44.
3. Kakar S. Clinical Faceoff: Controversies in the Management of Distal Radius Fractures. *Clin Orthop Relat Res.* 2015;473(10):3098-3104.
4. Koval K, Haidukewych GJ, Service B, Zirgibel BJ. Controversies in the management of distal radius fractures. *J Am Acad Orthop Surg.* 2014;22(9):566-575.
5. Koval KJ, Harrast JJ, Anglen JO, Weinstein JN. Fractures of the distal part of the radius. The evolution of practice over time. Where's the evidence? *J Bone Joint Surg Am.* 2008;90(9):1855-1861.
6. Members of the Writing R, Voting Panels of the AUCotToDRF, Watters WC, Sanders JO, Murray J, Patel N. The American Academy of Orthopaedic Surgeons Appropriate Use Criteria on the treatment of distal radius fractures. *J Bone Joint Surg Am.* 2014;96(2):160-161.
7. Murray J, Gross L. Treatment of distal radius fractures. *J Am Acad Orthop Surg.* 2013;21(8):502-505.
8. Al-Rashid M, Theivendran K, Craigen MA. Delayed ruptures of the extensor tendon secondary to the use of volar locking compression plates for distal radial fractures. *J Bone Joint Surg Br.* 2006;88(12):1610-1612.
9. Brunner A, Siebert C, Stieger C, Kastius A, Link BC, Babst R. The dorsal tangential X-ray view to determine dorsal screw penetration during volar plating of distal radius fractures. *J Hand Surg Am.* 2015;40(1):27-33.
10. Haug LC, Glodny B, Deml C, Lutz M, Attal R. A new radiological method to detect dorsally penetrating screws when using volar locking plates in distal radial fractures. The dorsal horizon view. *Bone Joint J.* 2013;95-b(8):1101-1105.
11. Hill BW, Shakir I, Cannada LK. Dorsal Screw Penetration With the Use of Volar Plating of Distal Radius Fractures: How Can You Best Detect? *J Orthop Trauma.* 2015;29(10):e408-413.
12. Joseph SJ, Harvey JN. The dorsal horizon view: detecting screw protrusion at the distal radius. *J Hand Surg Am.* 2011;36(10):1691-1693.
13. Ozer K, Wolf JM, Watkins B, Hak DJ. Comparison of 4 fluoroscopic views for dorsal cortex screw penetration after volar plating of the distal radius. *J Hand Surg Am.* 2012;37(5):963-967.
14. Ganesh D, Service B, Zirgibel B, Koval K. The Detection of Prominent Hardware in Volar Locked Plating of Distal Radius Fractures: Intraoperative Fluoroscopy Versus Computed Tomography. *J Orthop Trauma.* 2016;30(11):618-621.
15. Hufner T, Stubig T, Citak M, Gosling T, Krettek C, Kendoff D. Utility of intraoperative three-dimensional imaging at the hip and knee joints with and without navigation. *J Bone Joint Surg Am.* 2009;91 Suppl 1:33-42.
16. Richter M, Geerling J, Zech S, Goesling T, Krettek C. Intraoperative three-dimensional imaging with a motorized mobile C-arm (SIREMOBIL ISO-C-3D) in foot and ankle trauma care: a preliminary report. *J Orthop Trauma.* 2005;19(4):259-266.
17. Wich M, Spranger N, Ekkernkamp A. [Intraoperative imaging with the ISO C(3D)]. *Der Chirurg; Zeitschrift fur alle Gebiete der operativen Medizen.* 2004;75(10):982-987.

18. Mehling I, Rittstieg P, Mehling AP, Kuchle R, Muller LP, Rommens PM. Intraoperative C-arm CT imaging in angular stable plate osteosynthesis of distal radius fractures. *J Hand Surg Eur Vol.* 2013;38(7):751-757.
19. Müller M. *The Comprehensive Classification of Fractures of Long Bones.* New York, NY: Springer-Verlag. 1990.
20. Beerekamp MS, Ubbink DT, Maas M, 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. Bergsma M, Bulstra AE, Morris D, Janssen M, Jaarsma R, Doornberg J. Accuracy of Dorsal Tangential Views to Avoid Screw Penetration with Volar Plating of Distal Radius Fractures. *J Orthop Trauma.* 2020.
22. Henry MH, Griggs SM, Levaro F, Clifton J, Masson MV. Volar approach to dorsal displaced fractures of the distal radius. *Techniques in hand & upper extremity surgery.* 2001;5(1):31-41.
23. 2.4 mm LCP Distal Radius System. A comprehensive plating system to address a variety of fracture patterns. 2009;http://synthes.vo.llnwd.net/o16/Mobile/Synthes North America/Product Support Materials/Technique Guides/SUSA/SUTG2.4DRPltJ4569F.pdf.
24. Variable Angle LCP Two-Column Volar Distal Radius Plate 2.4. For fragment-specific fracture fixation with variable angle locking technology. Surgical Technique. DePuySynthes. 2015. http://synthes.vo.llnwd.net/o16/LLNWMB8/INT Mobile/Synthes International/Product Support Material/legacy_Synthes_PDF/ DSEM-TRM-0815-0464_LR.pdf (date last accessed 28 April 2020).
25. Sugun TS, Karabay N, Gurbuz Y, Ozaksar K, Toros T, Kayalar M. Screw prominences related to palmar locking plating of distal radius. *J Hand Surg Eur Vol.* 2011;36(4):320-324.
26. Rosset A, Spadola L, Ratib O. OsiriX: an open-source software for navigating in multidimensional DICOM images. *Journal of digital imaging.* 2004;17(3):205-216.
27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159-174.
28. Cicchetti D. Guidelines, criteria and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess.* 1994;6:284-290.
29. Bergsma M, Doornberg JN, Duit R, et al. Volar plating in distal radius fractures: A prospective clinical study on efficacy of dorsal tangential views to avoid screw penetration. *Injury.* 2018;49(10):1810-1815.
30. Kiyak G. In vivo confirmation of the reliability of the dorsal tangential view of the wrist. *Hand Surg Rehabil.* 2018;37(1):56-59.
31. Oc Y, Kilinc BE, Gulcu A, Varol A, Ertugrul R, Kara A. Ultrasonography or direct radiography? A comparison of two techniques to detect dorsal screw penetration after volar plate fixation. *J Orthop Surg Res.* 2018;13(1):70.
32. Diong TW, Haflah NHM, Kassim AYM, Habshi S, Shukur MH. Use of Computed Tomography in Determining the Occurrence of Dorsal and Intra-articular Screw Penetration in Volar Locking Plate Osteosynthesis of Distal Radius Fracture. *J Hand Surg Asian Pac Vol.* 2018;23(1):26-32.
33. Azzi AJ, Aldekhayel S, Boehm KS, Zadeh T. Tendon Rupture and Tenosynovitis following Internal Fixation of Distal Radius Fractures: A Systematic Review. *Plast Reconstr Surg.* 2017;139(3):717e-724e.
34. Bianchi S, van Aaken J, Glauser T, Martinoli C, Beaulieu JY, Della Santa D. Screw impingement on the extensor tendons in distal radius fractures treated by volar plating: sonographic appearance. *AJR American journal of roentgenology.* 2008;191(5):W199-203.

35. Arora R, Lutz M, Hennerbichler A, Krappinger D, Espen D, Gabl M. Complications following internal fixation of unstable distal radius fracture with a palmar locking-plate. *J Orthop Trauma.* 2007;21(5):316-322.
36. McKay SD, MacDermid JC, Roth JH, Richards RS. Assessment of complications of distal radius fractures and development of a complication checklist. *J Hand Surg Am.* 2001;26(5):916-922.
37. Pang EQ, Truntzer J, Baker L, Harris AHS, Gardner MJ, Kamal RN. Cost minimization analysis of the treatment of distal radial fractures in the elderly. *Bone Joint J.* 2018;100-B(2):205-211.

PART VI

Summary and Discussion

CHAPTER 8

Discussion

In this PhD thesis, efforts have been made to improve diagnosis and characterization of wrist trauma. The aim was to focus on diagnostic applications of artificial-intelligence algorithms as well as three-dimensional imaging strategies. The general discussion will interpret the findings and propose potential future research.

PART I: RISK STRATIFICATION IN THE EMERGENCY DEPARTMENT

Chapter 2

In line with the Ottawa Ankle Rules, simplicity of a model is believed to enhance use in clinical practice.[1,2] By only incorporating four variables, we deployed a freely available machine learning prediction tool that allows to calculate the probability of a fracture of the distal radius following wrist trauma with 98% sensitivity and 24% specificity.

It might be of interest to study whether our machine learning-based model reduces radiographic referral in a randomized controlled fashion, as we speculate this may lower healthcare related costs and simultaneously reduce length of stay in the ED. In a prospective cohort study, the initially developed Amsterdam Wrist Rules –based on the same dataset– decreased radiographic referral with 15%: however, eight variables were incorporated into this clinical decision rule, of which four are considered burdensome for the patient (e.g. radioulnar ballottement test).[3] Using the same dataset, but now machine learning, a comparable accuracy with only four variables was achieved.

Our online accessible application can guide patients and clinicians to improve shared-decision making. The current model only includes distal radius fractures, but future studies--combining large prospective wrist fracture databases--will be valuable for development of more extensive models encompassing all wrist injury related fractures. However, carpal fractures are outside the realm of our model. Therefore, end-users should apply our tool in conjunction with other models to evaluate significant wrist fractures.

PART II: DEEP LEARNING FOR FRACTURE DETECTION

Chapter 3

Recent breakthroughs and widely used online services such as Spotify and Netflix play a key role in artificial intelligence's excitement. Some believe AI technology is still in its early days and forecasted to accelerate the global gross domestic income with 14% in 2030.[4] While AI is entering mainstream in many medical specialties, various sectors already adopted these innovative strategies to boost their growth.

Chapter 3 demonstrated that AI and human examiners perform on par evaluating radiographs for certain fracture detection and classification tasks, and sometimes AI outperforms humans. However, the majority of these studies evaluated relatively easy discernible fractures, for example proximal humerus and distal radius.

Utilization and implementation of fracture detection and classification models in clinical practice is currently faced by challenges. For instance, labelling the outcome of interest (i.e. reference standard) for training an AI model is often a labour intensive manual task requiring considerable time. Not only a wide spectrum of clinical experience among physicians (i.e. fellowship trained musculoskeletal radiologists *versus* radiology residents) but also poor reliability of most fracture classification systems lead to substantial variation in quality of labelled data sets with which algorithms are trained.[5-8] Latent class analysis, Bayesian statistics, and consensus meetings among experts are potential ways to improve the reference standard.[9,10] AI derived models are known for being highly complex and often inscrutable. It is therefore difficult to explain how certain algorithms reach their conclusion. As a result, clinicians may be reluctant to rely on suggestions generated by a model that cannot be fully comprehended. For the end-user, areas indicating presence of a fracture might be useful to improve transparency, interpretability and trust in the model, given the closer relationship between input data and subsequent prediction model.[11] Furthermore, large amounts of data are necessary to sufficiently train AI models: however, in medicine, vast quantities of data are often difficult to obtain. Although electronic medical records (EMRs) are collecting health data digitally, routine care data is often stored as "crude oil" restricting its reuse for research purposes.[12] Finally, it is yet to be elucidated how most AI models--trained on a particular data set--are able to carry their previously gained experience to new circumstances (i.e. external validation).

Chapter 4

Clinical applications of AI have high potential to change the way we practice fracture care. In a research environment, viable AI models have been developed in orthopaedic surgery.[13-16] However, careful consideration is required to determine scenarios for which AI applications are beneficial. Evidence suggest that scaphoid fractures are easily missed on radiographs in the acute setting.[17] If not treated correctly, these fractures can lead to severe long-term consequences. Chapter 4 utilized a deep learning model to identify radiographically visible and occult scaphoid fractures. Using a relatively small dataset encompassing 300 patients, our model showed a sensitivity and accuracy on par with five orthopaedic surgeons, but with lower specificity.

Given that our model is not able to look beyond the outcome of interest (i.e. presence or absence of a scaphoid fracture), it will miss for example a scapholunate dissociation

or bone tumour. In contrast, orthopaedic trauma surgeons or radiologists are more likely to diagnose these additional relevant findings.

At this stage, attempts to create deep learning algorithms for image-based analysis are hindered by the difficulty of fitting real-sized radiographs.[18] We speculate that more computer memory capacity might eventually allow incorporation of real-sized radiographs into an algorithm. It is of relevance to evaluate whether model development with full scale image projections--reflecting clinical practice--yield better performance or could detect other relevant findings in the entire radiograph.

We found that incorporating age and sex demographics alone did not improve the performance characteristics of our model. More extensive information on physical examination, symptoms, and injury details added to predictions from a deep learning algorithm might prove valuable in developing clinical prediction rules that could accurately predict presence of a scaphoid fracture following wrist trauma. Future research might also assess whether more data will drive model performance.

PART III: CLINICAL PREDICTORS FOR SURGICAL DECISION MAKING

Chapter 5

There is a dearth of evidence in many clinical scenarios. Despite clinical guidelines and appropriateness criteria to better facilitate distal radius fracture management, treatment variations dominate clinical practice.[19-22] Chapter 5 was set out as a scenario-based survey study to better understand practice variation among surgeons. Based on fictitious distal radius fracture case scenarios, statistical analysis revealed that age and angulation were most influential in recommending operative treatment.

For fractures of the distal radius, the Appropriate Use Criteria--developed by the American Academy of Orthopaedic Surgeons--is a decision aid to help surgeons choose an evidence-based treatment.[23] These criteria indicate the best available option considering the following five factors: AO/OTA fracture type, mechanism of injury, patient activity level, patient health (American Society of Anesthesiologists' [ASA] status), and other injuries. While prior studies, in line with our findings, demonstrated that age and fracture displacement are important factors that drive treatment recommendation among surgeons, these factors are not yet adopted in the Appropriate Use Criteria.[24,25] Also, low agreement between the "appropriate" treatments recommended by the Appropriate Use Criteria and a surgeon's actual given treatment were found, especially for more severe distal radius fracture types.[25] As such, we can conclude that, in spite of appropriateness criteria, variation in clinical practice persists.

We speculate that factors derived from our study may provide insights that can be used when developing distal radius fracture specific decision aids. These decision aids

intend to help surgeon and patient come to a treatment choice that best matches a patient's requirements and expectations. AI predictive models--sophisticated statistical calculators commonly developed with large amounts of data--have the potential to estimate tailored treatment probabilities based on a patient's specific risk profile and fracture characteristics without taking preferences, misconceptions, and surgeon bias into account. These predictive models will limit treatment inconsistencies among surgeons.[7,26,27] In contrast to humans, who can only cognitively process four factors at the same time, these computer models are able to simultaneously process far more factors than just four.[28]

PART IV: 3D PRINTING FOR PREOPERATIVE PLANNING

Chapter 6

Optimal understanding of intraarticular distal radius fractures is paramount to facilitate preoperative planning. For distal radius fractures, 3D CT images in addition to radiographs and 2D CT images improve assessment of specific fracture characteristics.[29] In line with prior research on other anatomical fracture locations, our study demonstrated that combining a 3D printed handheld model for evaluating specific distal radius fracture characteristics with fracture classification does not improve preoperative reliability among surgeons.[30-32] 3D handheld models for teaching surgical residents and medical trainees merits further study as well as determining whether using sterilized 3D hand held models intra-operatively might prove valuable. Today, 3D printing has become cheap, user-friendly, and implementation into clinical practice is straightforward. This enhances the widespread adoption of the 3D printing techniques into medical specialties. Outside its potential value for teaching, we do not recommend its use in clinical practice for caring of intraarticular distal radius fractures.

PART V: 3D FLUOROSCOPY FOR INTRAOPERATIVE ASSESSMENT

Chapter 7

Intra-operative anteroposterior and lateral 2D fluoroscopy views have been traditionally used to evaluate fracture reduction and implant positioning for patients with a distal radius fracture. However, the complex 3D shape of the dorsal cortex often obscures evaluation of correct screw positioning. Dorsal tangential views (DTV) have been shown promising for reducing post-operative iatrogenic dorsal cortex penetration after volar plating for distal radius fractures.[33-35]

Literature is conflicting as to whether intra-operative use of 3D fluoroscopy might be a valuable adjunct. Also, 3D fluoroscopy detected misplaced screws in one third of patients that underwent volar plating for distal radius fractures: however, these screws were missed on conventional 2D fluoroscopy.[36] On the other hand, quality of distal radius fracture reduction and fixation demonstrated no difference between intra-operative use of 2D- and 3D fluoroscopy: however, these findings might potentially be slightly underpowered and therefore final conclusions might not be arrived at.[37]

In our study, we compared the intra-operative use of conventional 2DF (anteroposterior and elevated lateral projections), conventional 2DF with additional DTVs, and 3D fluoroscopy by determining the post-operative incidence of dorsal cortex screw penetration after volar plating for a distal radius fracture. We found that 3DF did not improve the diagnostic performance over DTVs. Based on these findings, we deem the intra-operative role of 3D fluoroscopy limited for detection of dorsal cortex penetrating screws for fractures of the distal radius. We concur that using DTVs, as compared to 3D fluoroscopy, are potentially more efficient in the daily clinical routine, easier to obtain intra-operatively, and also less expensive.

FUTURE PERSPECTIVES

There are many fragmented individual endeavours to improve trauma care. Substantial inconsistencies are seen among surgeons for treatment recommendation for fractures of the distal radius.[20,21] Open-source data--preferably from institutions across different continents--should be collected to improve fracture detection as well as enhance treatment management. At the forefront of these initiatives is the single-centre “Medical Information Mart for Intensive Care” (MIMIC)-database, an openly available anonymised data set including patient demographics, clinical data, and medications of about 60.000 patients admitted to critical care units of Beth Israel Deaconess Medical Center.[38] Using those databases, AI models may potentially reduce humans’ subjective interpretation as they only rely on input variables. On the other hand, bias in data is a problem not easily tackled.

A previous study incorporated about 4,000 patients to derive a formula to predict loss of threshold fracture alignment for distal radius fractures: however, a subsequent validation study demonstrated the calculator lacks generalisability when used in a different patient population.[39,40] Research efforts--utilizing deep learning algorithms with large data sets from different institutions--might prove valuable for deriving models that can accurately predict or even outperform human examiners, for example to help forecast distal radius fracture instability.

Another core interest for distal radius fracture management is to accurately estimate the patient-reported outcome for both conservative and surgical treatment. This might help surgeon and patient to initially decide on the best available treatment option for an individual patient. Artificial-intelligence algorithms, along with dynamic learning features, have the potential to improve over time as more patient data is provided.

Following the landmark example of the MIMIC collaborative in the field of intensive care medicine, our group coined the "Machine Learning Consortium" in the field of orthopaedic trauma surgery to enhance an open access mentality and drive synergy. The ML Consortium facilitates sharing of patient data and radiographs between institutions and across continents to combine strengths of each respective institute. For example, legislation and commercial contracts prohibit our collaborators at the Massachusetts General Hospital (MGH) in Boston to integrate their ML algorithms in the EMR. In contrast, such coupling of ML algorithms is legally allowed and technically facilitated at the University Medical Centre Groningen. Many of such endeavours are now underway within our ML consortium including development and internal validation of the initial algorithm in one centre, external validation in a subsequent collaborative centre, and "silent" prospective running as the ultimate test.

In summary, narrow-patterns tasks such as image-based analyses are at the forefront of the digital medicine AI era. Medicolegal regulations need to be confronted before widespread adoption into clinical practice is feasible. Although the European Union's recently embedded General Data Protection Regulation ensures data protection and privacy across Europe, data-sharing regulation is a matter of variability among different countries, trust, and cost. In addition, inscrutability of many AI algorithms may not be able to explain why the model errs. Finally, patient behaviour has large impact on for example treatment outcome, but it remains elusive whether suggestions made with the help of AI will lead to effective action or behavioural changes.

REFERENCES

1. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. *Ann Emerg Med.* 1992;21(4):384-390.
2. Stiell IG, Wells GA. Methodologic standards for the development of clinical decision rules in emergency medicine. *Ann Emerg Med.* 1999;33(4):437-447.
3. Mulders MAM, Walenkamp MMJ, Sosef NL, et al. The Amsterdam Wrist Rules to reduce the need for radiography after a suspected distal radius fracture: an implementation study. *Eur J Trauma Emerg Surg.* 2019.
4. Available at: https://www.pwc.com.au/health/ai/pwc-adopting-ai-in-healthcare-why-change-19feb2019.pdf, Accessed February 19, 2019.
5. Doornberg JN, Guitton TG, Ring D. Diagnosis of elbow fracture patterns on radiographs: interobserver reliability and diagnostic accuracy. *Clin Orthop Relat Res.* 2013;471(4):1373-1378.
6. Ghoshal A, Enninghorst N, Sisak K, Balogh ZJ. An interobserver reliability comparison between the Orthopaedic Trauma Association's open fracture classification and the Gustilo and Anderson classification. *Bone Joint J.* 2018;100-b(2):242-246.
7. Jayakumar P, Teunis T, Gimenez BB, Verstreken F, Di Mascio L, Jupiter JB. AO Distal Radius Fracture Classification: Global Perspective on Observer Agreement. *J Wrist Surg.* 2017;6(1):46-53.
8. Neuhaus V, Bot AG, Guitton TG, et al. Scapula fractures: interobserver reliability of classification and treatment. *J Orthop Trauma.* 2014;28(3):124-129.
9. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol.* 2018;73(5):439-445.
10. LaJoie AS, McCabe SJ, Thomas B, Edgell SE. Determining the sensitivity and specificity of common diagnostic tests for carpal tunnel syndrome using latent class analysis. *Plast Reconstr Surg.* 2005;116(2):502-507.
11. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A.* 2018;45:11591-11596.
12. Emanuel EJ, Wachter RM. Artificial Intelligence in Health Care: Will the Value Match the Hype? *JAMA.* 2019;321(23):2281-2282.
13. Hendrickx L, Sobol G, Langerhuizen DWG, et al. Incidence, Predictors and Fracture Mapping of (Occult) Posterior Malleolar Fractures Associated with Tibial Shaft Fractures. *J Orthop Trauma.* 2019 Dec;33(12):452-458.
14. Karhade AV, Thio Q, Ogink PT, et al. Development of Machine Learning Algorithms for Prediction of 30-Day Mortality After Surgery for Spinal Metastasis. *Neurosurgery.* 2019;85(1):E83-E91.
15. Ogink PT, Karhade AV, Thio Q, et al. Predicting discharge placement after elective surgery for lumbar spinal stenosis using machine learning methods. *Eur Spine J.* 2019;28(6):1433-1440.
16. Thio Q, Karhade AV, Ogink PT, et al. Can Machine-learning Techniques Be Used for 5-year Survival Prediction of Patients With Chondrosarcoma? *Clin Orthop Relat Res.* 2018;476(10):2040-2048.
17. Suh N, Grewal R. Controversies and best practices for acute scaphoid fracture management. *J Hand Surg Eur Vol.* 2018;43(1):4-12.
18. Ranschaert ER. Artificial Intelligence in Medical Imaging. *eBook Switzerland, AG: Springer.* 2019.
19. Heelkunde NVv. Richtlijn distale radius fracturen: diagnostiek en behandeling. Available at: https://heelkunde.nl/sites/heelkunde.nl/files/richtlijnen-definitief/Richtlijn_Distale_radius_fracturen_definitieve_versie_0511.pdf. Accessed 2010.

20. Ansari U, Adie S, Harris IA, Naylor JM. Practice variation in common fracture presentations: a survey of orthopaedic surgeons. *Injury.* 2011;42(4):403-407.
21. Walenkamp MM, Mulders MA, Goslings JC, Westert GP, Schep NW. Analysis of variation in the surgical treatment of patients with distal radial fractures in the Netherlands. *J Hand Surg Eur Vol.* 2016.
22. Murray J, Gross L. Treatment of distal radius fractures. *J Am Acad Orthop Surg.* 2013;21(8):502-505.
23. Appropriate use criteria web-based application: distal radius fractures treatment. Available at: www.aaos.org/aucapp. Accessed August 1, 2018.
24. Kyriakedes JC, Crijns TJ, Teunis T, Ring D, Bafus BT, Science of Variation G. International Survey: Factors Associated with Operative Treatment of Distal Radius Fractures and Implications for the American Academy of Orthopaedic Surgeons Appropriate Use Criteria. *J Orthop Trauma.* 2019.
25. Kyriakedes JC, Tsai EY, Weinberg DS, et al. Distal Radius Fractures: AAOS Appropriate Use Criteria Versus Actual Management at a Level I Trauma Center. *Hand (N Y).* 2018;13(2):209-214.
26. Maddox TM, Rumsfeld JS, Payne PRO. Questions for Artificial Intelligence in Health Care. *JAMA.* 2019;321(1):31-32.
27. Verghese A, Shah NH, Harrington RA. What This Computer Needs Is a Physician: Humanism and Artificial Intelligence. *JAMA.* 2018;319(1):19-20.
28. Halford GS, Baker R, McCredden JE, Bain JD. How many variables can humans process? *Psychol Sci.* 2005;16(1):70-76.
29. Harness NG, Ring D, Zurakowski D, Harris GJ, Jupiter JB. The influence of three-dimensional computed tomography reconstructions on the characterization and treatment of distal radial fractures. *J Bone Joint Surg Am.* 2006;88(6):1315-1323.
30. Brouwer KM, Lindenhovius AL, Dyer GS, Zurakowski D, Mudgal CS, Ring D. Diagnostic accuracy of 2- and 3-dimensional imaging and modeling of distal humerus fractures. *J Shoulder Elbow Surg.* 2012;21(6):772-776.
31. Guitton TG, Brouwer K, Lindenhovius AL, et al. Diagnostic accuracy of two-dimensional and three-dimensional imaging and modeling of radial head fractures. *J Hand Microsurg.* 2014;6(1):13-17.
32. Guitton TG, Kinaci A, Ring D. Diagnostic accuracy of 2- and 3-dimensional computed tomography and solid modeling of coronoid fractures. *J Shoulder Elbow Surg.* 2013;22(6):782-786.
33. Bergsma M, Doornberg JN, Duit R, et al. Volar plating in distal radius fractures: A prospective clinical study on efficacy of dorsal tangential views to avoid screw penetration. *Injury.* 2018.
34. Ganesh D, Service B, Zirgibel B, Koval K. The Detection of Prominent Hardware in Volar Locked Plating of Distal Radius Fractures: Intraoperative Fluoroscopy Versus Computed Tomography. *J Orthop Trauma.* 2016;30(11):618-621.
35. Haug LC, Glodny B, Deml C, Lutz M, Attal R. A new radiological method to detect dorsally penetrating screws when using volar locking plates in distal radial fractures. The dorsal horizon view. *Bone Joint J.* 2013;95-b(8):1101-1105.
36. Mehling I, Rittstieg P, Mehling AP, Kuchle R, Muller LP, Rommens PM. Intraoperative C-arm CT imaging in angular stable plate osteosynthesis of distal radius fractures. *J Hand Surg Eur Vol.* 2013;38(7):751-757.
37. Selles CA, Beerekamp MSH, Leenhouts PA, et al. The Value of Intraoperative 3-Dimensional Fluoroscopy in the Treatment of Distal Radius Fractures: A Randomized Clinical Trial. *J Hand Surg Am.* 2020;45(3):189-195.
38. MIMIC-III afaccdJA, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available at: http://www.nature.com/articles/sdata201635.

39. Mackenney PJ, McQueen MM, Elton R. Prediction of instability in distal radial fractures. *J Bone Joint Surg Am.* 2006;88(9):1944-1951.
40. Walenkamp MMJ, Mulders MAM, van Hilst J, Goslings JC, Schep NWL. Prediction of Distal Radius Fracture Redisplacement: A Validation Study. *J Orthop Trauma.* 2018;32(3):e92-e96.

CHAPTER 9

Summary in English

PART I: RISK STRATIFICATION IN THE EMERGENCY DEPARTMENT

Although only one third of patients presenting to the ED with wrist pain after sustaining an injury have a fracture of the distal radius, the majority is referred for radiographic evaluation. Therefore, **Chapter 2** intended to develop and externally validate a machine learning algorithm to predict the probability of a fracture of the distal radius for patients presenting to the ED after sustaining wrist trauma. For this, we included 854 patients that were prospectively enrolled at EDs of five hospitals; 488 patients in the derivation cohort and 366 in the validation cohort. Among nineteen variables, we used a random forest algorithm to determine the most influential predictors for incorporation into the algorithm (i.e. age, swelling of the wrist, visible deformation, and distal radius tender to palpation). Four machine-learning derived algorithms were developed on the derivation cohort: boosted decision tree, support vector machine, neural network, and Bayes point machine. Each algorithm's performance was assessed according to the following metrics: (1) c-discrimination (i.e. AUC); (2) calibration; and (3) Brier-score. All models showed nearly similar performance: c-statistics ranged between 0.86 and 0.88, while the Brier score was 0.16 for all models. Calibration slopes ranged between 0.72 and 0.84 and calibration intercepts ranged between -0.05 and -0.21. Bayes point machine was the best-fit algorithm. At a threshold of 0.05, the sensitivity and specificity were 0.98 and 0.24 respectively. The Bayes point machine algorithm was incorporated into an open access web-based application (accessible: http://traumaplatform.shinyapps.io/distalradius_ed).

PART II: DEEP LEARNING FOR FRACTURE DETECTION

This part starts with **Chapter 3**, an overview of aggregated literature (systematic review) addressing the accuracy and AUC of AI fracture detection and classification models. Also, we evaluated the performance of AI in a research setting compared with the performance of human examiners. For fracture detection, we found that the AUC reflected near perfect prediction (range, 0.95-1.0), and the accuracy ranged from 83% to 98%. For fracture classification, the AUC was 0.94, and the accuracy ranged from 77% to 90%. AI outperformed human examiners for detecting and classifying hip and proximal humerus fractures, and showed equivalent performance for detecting wrist, hand and ankle fractures. Fracture detection and classification using AI shows promising performance. AI may enhance processing and communicating probabilistic tasks in medicine, including orthopaedic surgery. At present, inadequate reference standard assignments to train and test AI is the biggest hurdle before integration into clinical workflow. The next step

will be to apply AI to more challenging diagnostic and therapeutic scenarios when there is absence of certitude.

Preliminary experience suggests that deep learning algorithms are nearly as good as humans in detection of common, displaced, and relatively obvious fractures (e.g. distal radius or hip fractures). **Chapter 4** tested the utility of a deep learning algorithm for scaphoid fractures, often a subtle or non-displaced fracture that is difficult to diagnose on radiographs. Specifically, we studied: (1) the diagnostic performance characteristics of a deep transfer learning algorithm in detecting scaphoid fractures using four radiographic imaging views; (2) whether the algorithm together with patient demographics would improve performance characteristics; (3) the algorithm's diagnostic performance as compared to five orthopaedic surgeons; and (4) the reliability of five human observers as well as the reliability between the algorithm and human consensus. Consecutive patients evaluated for a possible scaphoid fracture with radiographs and CT or MRI as reference standard were included until we had 150 fractures and 150 non-fractures as defined by radiologist diagnosis. We utilized a deep learning algorithm (a convolutional neural network [CNN]) for automated fracture detection on radiographs. The algorithm had an AUC of 0.77, 72% accuracy, 84% sensitivity, and 60% specificity. Adding age and sex--by using a multivariable logistic regression--had no significant influence on diagnostic performance. Specificity favoured the orthopaedic surgeons, while sensitivity and accuracy did not differ between the algorithm and human observers. The reliability among five surgeons was substantial (Fleiss' Kappa = 0.74). The reliability between the algorithm and human observers was only fair (Cohen's Kappa = 0.34).

PART III: CLINICAL PREDICTORS FOR SURGICAL DECISION MAKING

Evidence suggests that there is substantial and unexplained surgeon-to-surgeon variation in recommendation of operative treatment for fractures of the distal radius. In **Chapter 5**, we surveyed a global collaborative to understand bias and variation among surgeons to identify patient factors that influence recommendation for operative treatment of a fracture of the distal radius. One hundred thirty-one upper extremity and fracture surgeons evaluated 20 fictitious patient scenarios with randomly assigned factors (e.g. personal, clinical, and radiologic factors) for patients with a fracture of the distal radius. They addressed the following question: Do you recommend operative treatment for this patient (yes/no)? We determined the influence of each factor on this recommendation using random forest algorithms. Also, participants rated the influence of each factor--excluding age and sex--on a scale from 0 (not at all important) to 10 (extremely important). Random forest algorithms determined that age and angulation were having the most influence on recommendation for operative treatment of a fracture of the

distal radius. Angulation on the lateral radiograph and presence or absence of lunate subluxation were rated as having the greatest influence and smoking status and stress levels the lowest. The observation that--other than age--personal factors have limited influence on surgeon recommendations for surgery may reflect how cognitive biases, personal preferences, different perspectives, and incentives may contribute to variations in care.

PART IV: 3D PRINTING FOR PREOPERATIVE PLANNING

Three-dimensional printed hand-held models might improve the surgeons' interpretation of specific fracture characteristics pre-operatively and may therefore facilitate management. In **Chapter 6**, we determined whether the reliability of six specific distal radius fracture characteristics improve with additional pre-operative use of 3D printed hand-held models. Also, reliability of fracture classification (AO-type 23) and the surgeons' confidence (scale from 0 to 10) when assessing overall fracture configuration were assessed. On two occasions, ten surgeons evaluated 20 intraarticular distal radius fractures for presence or absence of the following fracture characteristics: volar rim fracture, die punch, volar lunate facet, dorsal comminution, step-off >2mm, and gap >2mm. Surgeons only used radiographs, 2D- and 3D CT-scans during the first occasion (i.e. conventional diagnostics). A month later, they used conventional diagnostics with an additional 3D printed hand-held model. We found that 3D printed hand-held models of intraarticular distal radius fractures led to no change in kappa values for the reliability of all characteristics. Fracture classification did not improve (conventional diagnostics: kappa, 0.27 [95% CI, 0.14 – 0.39] *versus* conventional diagnostics with an additional 3D printed hand-held model: kappa, 0.25 [95% CI, 0.15 – 0.35]). Confidence regarding overall fracture configuration showed no statistical difference (conventional diagnostics: 7.8 [95% CI, 7.2–8.3] *versus* conventional diagnostics with an additional 3D hand-held model: 8.5 [95% CI, 8.0–9.0]; p=0.09).

PART V: 3D FLUOROSCOPY FOR INTRAOPERATIVE ASSESSMENT

Chapter 7 investigated whether 3D fluoroscopy imaging outperforms DTVs to detect dorsal cortex screw penetration after volar plating for an intraarticular distal radius fracture. One-hundred sixty-five patients who underwent volar plating for an intraarticular distal radius fracture were evaluated to study three intra-operative imaging protocols: 2D fluoroscopy imaging with antero-posterior (AP) and elevated lateral images (n=55); 2D fluoroscopy imaging with AP, lateral, and DTV images (n=50); and 3D fluoroscopy

(n=60). Multi-planar reconstructions of post-operative computed tomography (CT) scans served as the reference standard. To detect dorsal cortex screw penetration, sensitivity of DTVs was 39% with a negative predictive value of 91% and an accuracy of 91%. For 3D fluoroscopy imaging, sensitivity was 25% with a negative predictive value of 93%, and an accuracy of 93%. On the CT reference standard post-operatively, we found penetrating screws in 40% of patients in the 2D fluoroscopy reference cohort; in 32% of patients in the 2D fluoroscopy cohort with AP, lateral and DTV images; and in 25% of patients in the 3D fluoroscopy cohort. The 2nd compartment was prone for penetration in all three imaging groups, while post-operative incidence decreased when more advanced imaging was used. No penetrating screws remained in situ with 3D fluoroscopy in the third compartment (extensor pollicis longus [EPL] groove), and one in the DTV-group. We concluded that advanced intra-operative imaging aids in identifying protruding screws in the dorsal wrist compartments. However, one cannot conclude that 3D fluoroscopy outperforms DTVs for this purpose.

CHAPTER 10

Summary in Dutch

DEEL 1: RISICO STRATIFICATIE OP DE SPOEDEISENDE HULP

Een derde van de patiënten die zich presenteren op de spoedeisende hulp na een letsel van de pols heeft een fractuur van de distale radius. De meerderheid van deze patiënten wordt echter verwezen voor aanvullende röntgendiagnostiek. In **hoofdstuk 2** hebben wij machine learning algoritmen ontwikkeld en extern gevalideerd om te voorspellen of een patiënt een fractuur van de distale radius heeft. Deze algoritmen zijn ontwikkeld met data van 854 patiënten die prospectief zijn geïncludeerd op de spoedeisende hulp van vijf ziekenhuizen. De algoritmen zijn initieel getraind met 488 patiënten en vervolgens gevalideerd met 366 patiënten. Met behulp van random forest algoritmes zijn uiteindelijk vier van de negentien variabelen opgenomen in het algoritme: leeftijd, zwelling van de pols, zichtbare vervorming en pijn bij palpatie van de distale radius. Een boosted decision tree, support vector machine, neural network, en Bayes point machine zijn getraind. C-discriminaties lieten een variatie zien van 0.86 tot 0.88, terwijl de Brier-score vergelijkbaar was voor ieder algoritme (0.16). De kalibratiehellingen varieerden tussen 0.72 en 0.84 en kalibratiesnijpunten tussen -0.05 en -0.21. Bayes point machine was het best presterende algoritme, met een sensitiviteit van 98% en specificiteit van 24%. Deze is vervolgens ontwikkeld als een vrij toegankelijke onlineapplicatie: (http://trauma-platform.shinyapps.io/distalradius_ed).

DEEL II: DEEP LEARNING VOOR FRACTUUR DETECTIE

Het tweede deel start met **hoofdstuk 3**, een overzicht van de literatuur waarin de AUC en accuracy van kunstmatige intelligente algoritmen voor fractuur diagnose en fractuur classificatie uiteengezet wordt. Tevens zijn de prestaties van deze algoritmen vergeleken met prestaties van artsen. Voor fractuur diagnose vonden we een AUC tussen 0.95 en 1.0, terwijl de accuracy varieerde van 83% tot 98%. De AUC voor fractuurclassificatie was 0.94 en de accuracy varieerde van 77% tot 90%. In vergelijking met artsen waren de kunstmatige intelligente algoritmen beter in het diagnosticeren en classificeren van heup- en proximale humerusfracturen. De algoritmen en artsen presteerden vergelijkbaar voor het diagnosticeren van pols-, hand- en enkelfracturen. Tot op heden bestaan veel ontoereikende referentiestandaarden voor het trainen en testen van deze algoritmen. Dit is een van de grootste uitdagingen voordat implementatie van deze toepassingen in de klinische praktijk haalbaar wordt geacht.

Onderzoek toonde aan dat deep learning algoritmen dezelfde prestaties hebben als artsen voor diagnose van veel voorkomende, verplaatste en relatief makkelijk aan te tonen fracturen (bijvoorbeeld distale radius en heupfracturen). **Hoofdstuk 4** onderzoekt of een deep learning algoritme scaphoïd fracturen kan diagnosticeren op röntgenfoto's.

De prestaties van het algoritme hebben wij vergeleken met vijf orthopedisch chirurgen. In totaal zijn 300 patiënten onderzocht die zich presenteerden op de spoedeisende hulp met een verdenking op een scaphoid fractuur. Hiervan hadden 150 patiënten een fractuur en 150 patiënten geen fractuur. De referentiestandaard voor het algoritme, aan- of afwezigheid van een fractuur, werd bevestigd op CT of MRI door een radioloog. Het algoritme had een AUC van 0.77, een accuracy van 72%, een sensitiviteit van 84% en een specificiteit van 60%. De prestaties van het algoritme verbeterden niet door toevoeging van leeftijd en geslacht. Specificiteit was beter voor orthopedisch chirurgen. Het algoritme en orthopedisch chirurgen hadden een vergelijkbare sensitiviteit en accuracy. De betrouwbaarheid onder vijf chirurgen was aanzienlijk (Fleiss' Kappa = 0.74). De betrouwbaarheid tussen het algoritme en chirurgen was redelijk (Cohen's Kappa = 0.34).

PART III: KLINISCHE VOORSPELLERS VOOR CHIRURGISCHE BESLUITVORMING

Er is substantiële en onverklaarbare variatie onder chirurgen voor het aanbevelen van een operatieve behandeling voor distale radius fracturen. In **hoofdstuk 5** proberen we meer inzicht te krijgen in deze variatie door een wereldwijd panel van chirurgen te vragen welke patiëntfactoren de aanbeveling voor operatie van een distale radius fractuur beïnvloeden. Honderdeenendertig chirurgen hebben twintig fictieve patiëntscenario's geëvalueerd. Een computerprogramma heeft scenario's gecreëerd waarin verschillende persoonlijke-, klinische- en patiëntfactoren willekeurig werden toegewezen. Bij iedere casus werd de volgende vraag gesteld: zou u wel of geen operatieve behandeling aanbevelen voor deze patiënt? Vervolgens werd de invloed van elke patiëntfactor op de behandelingskeuze berekend met behulp van een random forest algoritme. Daarnaast beoordeelden de deelnemers de invloed van elke factor op een schaal van 0 tot 10. Het algoritme toonde aan dat leeftijd en angulatie factoren zijn met de meeste invloed op het aanbevelen van een operatieve behandeling. Angulatie op de laterale röntgenfoto en aan- of afwezigheid van lunatum subluxatie waren als meest invloedrijke factoren beoordeeld, terwijl rookstatus en stress als minst belangrijk werden beschouwd. Behalve leeftijd en geslacht lijken persoonlijke factoren een beperkte invloed te hebben op de aanbeveling voor operatieve behandeling.

PART IV: 3D GEPRINTE MODELLEN VOOR PREOPERATIEVE VOORBEREIDING

In **hoofdstuk 6** onderzochten we voor zes fractuurkenmerken van de distale radius of de betrouwbaarheid onder 10 chirurgen toeneemt door toevoeging van 3D geprinte draagbare modellen. Daarnaast werd de betrouwbaarheid voor fractuurclassificatie (AO-type 23) beoordeeld. Iedere chirurg beoordeelde 20 intra-articulaire distale radius fracturen op aan- of afwezigheid van de volgende zes fractuurkenmerken: volaire rim aantasting, die punch, volaire lunatum facet, dorsale verkleining, step-off >2mm en gap >2mm. Chirurgen gebruikten tijdens de eerste evaluatie röntgenfoto's, 2D- en 3D CT-scans (d.w.z. conventionele diagnostiek). Tijdens de tweede evaluatie, een maand later, werd naast conventionele diagnostiek ook een extra 3D geprint draagbaar model gebruikt. De toevoeging van 3D geprinte draagbare modellen heeft niet geleid tot een verbetering van de zes onderzochte fractuurkarakteristieken. Daarnaast leidde het gebruik van 3D geprinte draagbare modellen niet tot verbeterde fractuurclassificatie (conventionele diagnostiek: kappa, 0.27 [95% BI, 0.14 - 0.39] *versus* conventionele diagnostiek met een extra 3D-geprint draagbaar model: kappa, 0.25 [95% BI, 0.15 - 0.35]).

PART V: 3D DOORLICHTING VOOR INTRAOPERATIEVE BEOORDELING

In **hoofdstuk 7** hebben wij onderzocht of 3D fluoroscopie een betere intra-operatieve beeldvormingsmodaliteit is dan DTV om dorsaal uitstekende schroeven aan te tonen na volaire plaatfixatie van een distale radius fractuur. Drie verschillende intra-operatieve beeldvormingsstrategieën zijn vergeleken: 2D fluoroscopie met AP en laterale afbeeldingen (n=55), 2D fluoroscopie met AP-, laterale- en DTV-afbeeldingen (n=50) en 3D fluoroscopie (n=60). Reconstructiebeelden van postoperatieve CT-scans werden gebruikt om aan- of afwezigheid van uitstekende schroeven in de dorsale cortex te bepalen. De sensitiviteit van DTV was 39%, de negatief voorspellende waarde was 91% en de accuracy 91%. Voor 3D fluoroscopie was de sensitiviteit 25%, de negatief voorspellende waarde 93% en de accuracy ook 93%. Bij 40% van de patiënten werd in het 2D fluoroscopie referentiecohort (AP en laterale afbeeldingen) een uitstekende schroef gezien. In het DTV-cohort was dit 32%, terwijl dit 25% was voor patiënten in het 3D fluoroscopie cohort. Voor alle drie de beeldvormingsstrategieën werden in het 2de dorsale compartiment de meest uitstekende schroeven gezien. Naarmate meer geavanceerde beeldvorming werd gebruikt nam de hoeveelheid uitstekende schroeven af. Voor 3D fluoroscopie zagen we geen uitstekende schroeven in het derde compartiment, terwijl dit één schroef betrof in het DTV-cohort. Geavanceerde intra-operatieve beeldvormingstechnieken kunnen

helpen bij het identificeren van uitstekende schroeven in de dorsale compartimenten van de distale radius. Voor het aantonen van uitstekende schroeven kan men niet concluderen dat intra-operatieve 3D fluoroscopie beter presteerde dan DTV.

APPENDICES

Abbreviations
Portfolio
Report of Scholarship
Acknowledgements
About the author

ABBREVIATIONS

2D	=	two-dimensional
3D	=	three-dimensional
95% CI	=	95% confidence interval
AI	=	artificial intelligence
AUC	=	area under the receiver operating characteristic curve
CT	=	computed tomography
DL	=	deep learning
DTV	=	dorsal tangential view
ED	=	emergency department
IQR	=	interquartile range
ML	=	machine learning
MRI	=	magnetic resonance imaging
ORIF	=	open reduction internal fixation
SD	=	standard deviation
SOVG	=	science of variation group

REPORT OF SCHOLARSHIP

PUBLICATIONS, PEER-REVIEWED

D.W.G. Langerhuizen, S.J. Janssen, Q.M. van der Vliet, K.A. Rasking, M.L. Ferrone, F.J. Hornicek, J.H. Schwab, S.A. Lozano-Calderon. Metastasectomy, intralesional resection, or stabilization only in the treatment of bone metastases from renal cell carcinoma. *Journal of Surgical Oncology 2016.*

N.R. Paulino Pereira, **D.W.G. Langerhuizen**, S.J. Janssen, F.J. Hornicek, M.L. Ferrone, M.B. Harris, J.H. Schwab. Are perioperative allogeneic blood transfusion associated with 90-days infection after operative treatment for bone metastases? *Journal of Surgical Oncology 2016.*

S.J. Janssen, **D.W.G. Langerhuizen**, J.H. Schwab, J.A.M. Bramer. Outcome After Endoprosthetic Reconstruction of Proximal Femoral Tumors: A Systematic Review. *Journal of Surgical Oncology 2018.*

D.W.G. Langerhuizen, S.J. Janssen, W.H. Mallee, M.P.J. van den Bekerom, D. Ring, G.M.M. J. Kerkhoffs, R.L. Jaarsma, 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.*

L.A.M. Hendrickx, G.L. Sobol, **D.W.G. Langerhuizen**, A.E.J. Bulstra, J. Hreha, S. Sprague, M.S. Sirkin, D. Ring, G.M.M.J. Kerkhoffs, R.L. Jaarsma, J.N. Doornberg, Machine Learning Consortium. A Machine Learning Algorithm to Predict the Probability of (Occult) Posterior Malleolar Fractures Associated with Tibial Shaft Fractures to Guide “Malleolus First” Fixation. *Journal of Orthopaedic Trauma 2019.*

D.W.G. Langerhuizen, A.E.J. Bulstra, S.J. Janssen, D. Ring, G.M.M.J. Kerkhoffs, R.L. Jaarsma, J.N. Doornberg. Is Deep Learning On Par with Human Observers for Detection of Radiographically Visible and Occult Fractures of the Scaphoid? *Clinical Orthopaedics & Related Research 2020.*

D.W.G. Langerhuizen, M. Bergsma, C.A. Selles, R.L. Jaarsma, J.C. Goslings, N.W.L. Schep, J.N. Doornberg. Diagnosis of dorsal screw penetration after volar plating of a distal radius fracture. *The Bone & Joint Journal 2020.*

D.W.G. Langerhuizen, J.N. Doornberg, M.M.A. Janssen, G.M.M.J. Kerkhoffs, R.L. Jaarsma, S.J. Janssen. Do 3D Printed Handheld Models Improve Surgeon Reliability for

Recognition of Intraarticular Distal Radius Fracture Characteristics? *Clinical Orthopaedics & Related Research 2020.*

D.W.G. Langerhuizen, L.E. Brown, J.N. Doornberg. D. Ring, G.M.M.J. Kerhoffs, S.J. Janssen. Analysis of Online Reviews of Orthopaedic Surgeons and Orthopaedic Practices Using Natural Language Processing. *Journal of American Academy of Orthopaedic Surgeons 2021.*

A.N. Steffens, **D.W.G. Langerhuizen**, J.N. Doornberg, D. Ring, S.J. Janssen. Emotional tones in scientific writing: comparison of commercially funded studies and non-commercially funded orthopedic studies. *Acta Orthopaedica 2021.*

S.J. Janssen, **D.W.G. Langerhuizen**, G.M.M.J. Kerkhoffs, D. Ring. Payments by Industry to Residency Program Directors in the United States: A Cross-Sectional Study. *Academic Medicine 2021.*

D.W.G. Langerhuizen, S.J. Janssen, J.T.P. Kortlever, D. Ring, G.M.M.J. Kerkhoffs, R.L. Jaarsma, J.N. Doornberg. Factors Associated with a Recommendation for Operative Treatment for Fracture of the Distal Radius. *Journal of Wrist Surgery 2021.*

D.W.G. Langerhuizen, L.P.E. Verweij, J.C. van der Wouden, G.M.M.J. Kerkhoffs, S.J. Janssen. Antihypertensive drugs demonstrate varying levels of hip fracture risk: A systematic review and meta-analysis. *Injury 2021.*

A.B. Walinga, T. Stornebrink, **D.W.G. Langerhuizen**, P.A. Struijs, S.J. Janssen, G.M.M.J. Kerkhoffs. What are the Best Diagnostic Tests for Diagnosing Bacterial Arthritis of a Native Joint? A Systematic Review of 27 Studies. *The Bone & Joint Journal 2021.*

PODIUM PRESENTATIONS (PRESENTER)

N.R. Paulino Pereira, **D.W.G. Langerhuizen**, S.J. Janssen, F.J. Hornicek, M.L. Ferrone, M.B. Harris, J.H. Schwab. Are Perioperative Allogeneic Blood Transfusions Associated With 90-days Infection After Operative Treatment For Bone Metastases? October 2016, Musculoskeletal Tumor Society, Detroit, Michigan, USA.

S.J. Janssen, **D.W.G. Langerhuizen**, J.H. Schwab, J.A.M. Bramer. Outcome After Endoprosthetic Reconstruction of Proximal Femoral Tumors: A Systematic Review. May 2018, European Musculoskeletal Oncology Society, Amsterdam, The Netherlands.

D.W.G. Langerhuizen, S.J. Janssen, W.H. Mallee, M.P.J. van den Bekerom, D. Ring, G.M.M.J. Kerkhoffs, J.N. Doornberg, R.L. Jaarsma. Artificial Intelligence in Bone Fracture Detection and Classification: A Systematic Review. October 2018, AOA SA/NT, Adelaide, Australia.

B.J.A. Schoolmeesters, **D.W.G. Langerhuizen**, G.M.M.J. Kerkhoffs, D. Eygendaal, M.P.J. van den Bekerom, R.L. Jaarsma, J.N. Doornberg, B. Jadav. 3D Printed Handheld Models in Orthopaedic Trauma Surgery: Does it Improve our Patients' Outcomes? A Systematic Review. June 2019, AOTS/ASM, Cairns, Australia.

L.A.M. Hendrickx, G. Sobol, **D.W.G. Langerhuizen**, A.E.J. Bulstra, J. Hreha, S. Sprags, M. Sirkin, D. Ring, G.M.M.J. Kerkhoffs, R.L. Jaarsma, J.N. Doornberg. A Machine Learning Algorithm to Predict the Probability of (Occult) Posterior Malleolar Fractures Associated with Tibial Shaft Fractures to Guide "Malleolus First" Fixation. June 2019, AOTS/ASM, Cairns, Australia.

D.W.G. Langerhuizen, M.B. Bergsma, C.A. Selles, R.L. Jaarsma, N.W. Schep, J.N. Doornberg. Diagnosis of Dorsal Screw Penetration after Volar Plating of the Distal Radius: Intra-Operative Dorsal Tangential Views versus Three-Dimensional Fluoroscopy. Augustus 2019, AOA SA/NT, Adelaide, Australia.

B.J.A. Schoolmeesters, **D.W.G. Langerhuizen**, G.M.M.J. Kerkhoffs, D. Eygendaal, M.P.J. van den Bekerom, R.L. Jaarsma, J.N. Doornberg, B. Jadav. 3D Printed Hand Held Models in Orthopaedic Trauma Surgery: A Systematic Review. Augustus 2019, AOA SA/NT, Adelaide, Australia.

L.A.M. Hendrickx, G. Sobol, **D.W.G. Langerhuizen**, A.E.J. Bulstra, J. Hreha, S. Sprags, M. Sirkin, D. Ring, G.M.M.J. Kerkhoffs, R.L. Jaarsma, J.N. Doornberg. A Machine Learning Algorithm to Predict the Probability of (Occult) Posterior Malleolar Fractures Associated with Tibial Shaft Fractures to Guide "Malleolus First" Fixation. Augustus 2019, AOA SA/NT, Adelaide, Australia.

D.W.G. Langerhuizen, M.B. Bergsma, C.A. Selles, R.L. Jaarsma, N.W. Schep, J.N. Doornberg. Diagnosis of Dorsal Screw Penetration after Volar Plating of the Distal Radius: Intra-Operative Dorsal Tangential Views versus Three-Dimensional Fluoroscopy. Augustus 2019, AOA SA/NT, Adelaide, Australia.

D.W.G. Langerhuizen, M.B. Bergsma, C.A. Selles, R.L. Jaarsma, N.W. Schep, J.N. Doornberg. Diagnosis of Dorsal Screw Penetration after Volar Plating of the Distal Radius:

Intra-Operative Dorsal Tangential Views versus Three-Dimensional Fluoroscopy. November 2019, Traumadagen, Amsterdam, The Netherlands.

POSTER PRESENTATIONS (PRESENTER)

D.W.G. Langerhuizen, S.J. Janssen, Q. van der Vliet, K. Raskin, M. Ferrone, F.J. Hornicek, J.H. Schwab, S. Lozano-Calderon. Metastasectomy versus Intralesional Resection in the Treatment of Bone Metastases from Renal Cell Carcinoma. October, 2015, MGH Clinical Research Day, MGH, Boston, Massachusetts, USA.

N.R. Paulino Pereira, R.B. Beks, **D.W.G. Langerhuizen**, S.J. Janssen, F.J. Hornicek, M. Ferrone, J.H. Schwab. Are allogeneic blood transfusions associated with decreased survival after surgical treatment for spinal metastases? October, 2015, MGH Clinical Research Day, MGH, Boston, Massachusetts, USA.

D.W.G. Langerhuizen, B. Hayat, W.H. Mallee, D.T. Meijer, G.M.M.J. Kerkhoffs, B. Geerts, R.L. Jaarsma, J.N. Doornberg. Artificial Intelligence and Deep Learning Algorithm for Automated Segmentation of Computed Tomography Scans of Intra-Articular Fractures. February, 2019, ORS Annual Meeting, Austin, Texas, USA.

D.W.G. Langerhuizen, B. Hayat, W.H. Mallee, D.T. Meijer, G.M.M.J. Kerkhoffs, B. Geerts, R.L. Jaarsma, J.N. Doornberg. Artificial Intelligence and Deep Learning Algorithm for Automated Segmentation of Computed Tomography Scans of Intra-Articular Fractures. September, 2019, ORS Annual Meeting, Denver, Colorado, USA.

D.W.G. Langerhuizen, Traumaplatform AI Collaborative. Machine Learning to Predict the Probability of a Distal Radius Fracture in Patients Presenting to the Emergency Department with Sustained Wrist Trauma: A Clinical Prediction Rule to Indicate further Radiographic Evaluation. November 2019, Traumadagen, Amsterdam, The Netherlands.