87733_Auto_Edited.docx

WORD COUNT	3956	TIME SUBMITTED	24-OCT-2023 11:59AM
		PAPER ID	103684858

Name of Journal: *World Journal of Orthopedics*
Manuscript NO: 87733
Manuscript Type: ORIGINAL ARTICLE

Basic Study

Automated decision support for Hallux Valgus treatment options using anteroposterior foot radiographs

Konrad Kwolek, Artur Gądek, Kamil Kwolek, Radek Kolecki, Henryk Liszka

Abstract

BACKGROUND

Assessment of the potential utility of deep learning with subsequent image analysis to automate the measurement of hallux valgus and intermetatarsal angles from radiographs to serve as a preoperative aid in establishing hallux valgus severity for clinical decision-making.

AIM

To investigate the accuracy of automated measurements of angles of hallux valgus from radiographs for further integration with the preoperative planning process.

METHODS

The data comprises 265 consecutive digital anteroposterior weightbearing foot radiographs. 181 radiographs were utilized for training (161) and validating (20) a U-Net neural network to achieve a mean Sørensen–Dice index (SDI) >97% on bone segmentation. 84 test radiographs were used for manual (computer assisted) and automated measurements of hallux valgus severity determined by hallux valgus (HVA) and intermetatarsal angles (IMA). The reliability of manual and computer-based measurements was calculated using the interclass correlation coefficient (ICC) and

standard error of measurement (SEM). Inter- and intraobserver reliability coefficients were also compared. An operative treatment recommendation was then applied to compare results between automated and manual angle measurements.

RESULTS

Very high reliability was achieved for hallux valgus angle (HVA) and intermetatarsal angle (IMA) between the manual measurements of three independent clinicians. For HVA, the interclass correlation coefficient (ICC) between manual measurements was 0.96-0.99. For IMA, ICC was 0.78-0.95. Comparing manual against automated computer measurement, the reliability was high as well. For HVA, absolute agreement ICC and consistency ICC were 0.97, and standard error of measurement (SEM) was 0.32. For IMA, absolute agreement ICC (AA-ICC) was 0.75, consistency ICC (C-ICC) was 0.89, and SEM was 0.21. Additionally, a strong correlation (0.80) was observed between our approach and traditional clinical adjudication for preoperative planning of hallux valgus, according to an operative treatment algorithm proposed by EFORT.

CONCLUSION

The proposed automated and AI-assisted determination of hallux valgus angles based on deep learning holds great potential as an accurate and efficient tool, with comparable accuracy to manual measurements by expert clinicians. Our approach can be effectively implemented in clinical practice to determine the angles of hallux valgus from radiographs, classify the deformity severity, streamline preoperative decision-making prior to corrective surgery.

INTRODUCTION

Hallux valgus (HV) is a foot deformity that affects a considerable percentage of the population[1, 2]. It is a complex positional deformity of the first ray that leads to altered joint mechanics, dysfunction, and progressive pain. The technique of weightbearing

dorsoplantar radiographs was standardized and determined in the AOFAS research committee report[3-6]. Orthopedic surgeons frequently use radiographic angles to make clinical decisions for patients with symptomatic hallux valgus (HV)[7-9]. Various radiographic measurements used in hallux valgus treatments were discussed[3, 10]. The reliability of radiographic measurements in HV was also studied[11]. Through the use of WBCT scans, it has been demonstrated that up to 87% of hallux valgus cases exhibit metatarsal bone pronation, emphasizing the intricate multiplanar nature of this deformity. This metatarsal pronation explains the perceived metatarsal bone shape and the misalignment of the medial sesamoid bone in radiological studies, which has been recognized as a significant factor contributing to recurrence following treatment. As a result, distal metatarsal articular angle has proved unreliable, demonstrating a poor interobserver agreement[12, 13]. Further research is needed to develop effective approaches for addressing the rotational deformity in individuals with HV[2, 14, 15].

Key angles utilized in clinical practice to establish the severity of HV are the hallux valgus angle (HVA) and the intermetatarsal angle (IMA)[8, 12, 13, 16-18]. Intra- and inter-observer agreement for radiographic measurement of HVA/IMA is reportedly good using various digital techniques[11, 17, 19]. The HVA is between the longitudinal axes of the first metatarsal and the proximal phalanx (PP). The IMA is between the longitudinal axes of the first and second metatarsal bones (Figure 1). Many methods were used to facilitate and accelerate manual or computer-assisted determination of the longitudinal axes of the first, second metatarsal (1,2MT) and the hallucial proximal phalanx (PP) bones, *e.g.*, establishing reference points, to make measurements more repeatable[7, 20].

Traditionally, these angles were manually measured on hard-copy radiographs. Nowadays, computer-assisted measurement methods are being developed, that reduce the measurement error of HVA[21-23]. New possibilities for radiographic images analysis have emerged thanks to recent advances in clinical applications of deep learning[24-29]. This study is among the first forays into automated HVA/IMA measurements from radiographs.

Kwolek *et al* introduced an algorithm for the automatic recognition of radiographs of the hallux valgus using U-Net neural network with promising outcomes[30]. A study on hallux valgus measurement with a deep convolutional neural network based on landmark detection has been discussed by Li *et al*[31]. In contrast to our approach based on the toe bones segmentation and reference points estimation, their method is based on a small number of landmark points. Moreover, their database contains mainly radiographs without hallux valgus (almost 50%) or with small deformation, i.e. only 5/340 (1,5%) radiographs have IMA > 16° (severe hallux valgus deformation).

In this study, we significantly expanded algorithms to automate HV assessment from foot radiographs[30]. The necessary bones (first, second metatarsal, and hallucial PP) were segmented and labelled by a U-Net to set reference points and calculate HVA/IMA automatically. Expert clinicians also determined these angles manually, with outcomes being compared later. Moreover, our algorithm was evaluated only on patients' radiographs who subsequently underwent hallux valgus surgery. Our dataset contains a considerable percentage of radiographs with severe forefoot deformations including toe overlap, severe pronation, and sesamoid dislocation. Our bone segmentation-based algorithm is sufficiently robust to handle even such challenging circumstances anatomy as toes overlapping.

Classification systems

Traditional classification methods rely upon weightbearing anteroposterior radiographs to determine the severity of HV based on the HVA, and IMA (Figure 1)[17]. More than 100 different operative techniques were described for the correction of HV[32-34]. The overall clinical picture together with the degree of deformity determine the surgical decisions made. A suitable intervention is selected by considering the overall clinical picture along with the degree of deformity, potential degenerative changes of the first metatarsophalangeal joint, size, and shape of the metatarsal, and joint congruency.

Our algorithm is based on operative treatment algorithm proposed by EFORT (Figure 2)[35]. A convolutional neural network (CNN) was trained to segment bones,

with subsequent image analysis to automatically estimate angles and recommend appropriate surgical decisions. Digital radiographs were managed using a picture archiving and communication system and the IMPAX software suite.

MATERIALS AND METHODS

Algorithm outline

The measurements of the HVA/IMA were performed automatically on bones segmented and labelled by the U-Net neural network (Figure 3)[36]. To achieve this, the U-Net was first trained using anteroposterior foot radiographs and corresponding images with manually segmented and labelled bones. By providing automatically segmented and labelled bones, the required reference points were likewise automatically measured and the HVA/IMAs ultimately calculated.

The U-Net was trained only on right feet radiographs to reduce the cost and time of model training. Radiographs with the left feet were mirrored and then incorporated into the database. At the angle measurement stage, the segmented images with the left feet were back-mirrored to perform the measurements on the feet in the original orientation.

Dataset

133 patients were randomly selected between 2014 and 2021. A total of 265 pre-operative (unilateral or bilateral) anteroposterior feet radiographs were sourced from the electronic database of the authors' institution (demographics in Table 1). Inclusion criteria were: available weight-bearing radiographs, sole indication: symptomatic hallux valgus. Exclusion criteria were: no available weight-bearing radiographs, prior osteotomies, radiographs with severe osteoarthritis and first metatarsophalangeal joint deformation, and/or severe, *e.g.*, rheumatoid forefoot deformations or Charcot diabetic foot, visible plates, and other artificial elements distorting the image of the bone. Radiographs were obtained using standard radiology equipment Eidos RF439 and

Luminos DRF unit and digitally transmitted via a picture archiving and communication system.

The data was divided randomly into three subsets: training, validation, and testing (Figure 3B). Both the patient's right and left feet were included in these subsets. The training and validation subsets were used to train and validate the U-Net for bone segmentation, while the testing subset was used to evaluate the performance of the trained U-Net and automatically measure the HVA/IMA.

Training and validation subset

We initially applied a 71/29 percent random split between training and validating subsets to develop the U-Net. After achieving the SDI greater than the cutoff value, the final training set was established.

Testing subset

According to Zhou *et al*, the minimum number of subjects (testing subset) to estimate the agreement of the measurements between the two methods is 80[37]. Our testing subset consisted of 84 randomly selected anteroposterior foot radiographs. Apart from the input radiographs, the testing subset also contained manually segmented radiographs to evaluate the quality of bone segmentation by the U-Net network. We calculated the SDIs using radiographs with automatically and manually segmented bones. There were no duplicate patient radiographs between the training, validation, and testing subsets. The validation radiographs were used to select the best neural network model, and validate the performance of the selected network during its training. The HVA/IMA were estimated only on the testing radiographs.

Anonymization and manual labelling

The input radiographs were anonymized (Figure 3A) and stored in .png image format with lossless compression. Radiographs were digitally anonymized with unique IDs. To train a U-Net network that would achieve high bone segmentation accuracy, bones

were manually annotated on original high-resolution radiographs. Initially, seventy radiographs were manually segmented and labelled by the first author in Adobe Photoshop. The radiographs with labelled bones were randomly split into a set of 50 training and 20 validation images. Manual segmentation of bones on radiographs is a very time-consuming task with considerable effort necessary to properly separate the border of bone from surrounding soft tissue. Considering the current understanding of pronation and variable shape of the first metatarsal head in hallux valgus deformation described by Wagner *et al*[14], the first metatarsal head and the sesamoid bones were delineated carefully and precisely by a foot surgeon to achieve precise measurements of the HVA/IMA[11]. The complex structure of bones in anteroposterior feet radiographs makes automated segmentation (delineation) particularly difficult[38]. Radiographs are contaminated by noise, artifacts, insufficient contrast, resolution, and/or intensity. These factors made preparing the dataset and developing the algorithm particularly challenging.

Various bone segmentation strategies were considered during algorithm development. We started with a binary segmentation of bones with bone extraction[30]. However, this approach produced clinically unreliable results in cases of cross-over toe with higher HVA. To overcome these difficulties, and simplify the algorithm to achieve robust automated separation of each required bone even in "difficult" radiographs, we established main regions on each foot radiograph *via* multi-class segmentation (Figure 3A). This approach allowed us to select and process just the three bones forming the HVA/IMA (1,2MT, and hallucial PP) and exclude all remaining structures and radiograph background. Considering that the region of interest on a given foot may have varying aspect ratios (height to width), some images were padded vertically with rows of black pixels to standardize the image size to 768x1024 pixels without changing the resolution (Figure 4).

U-Net training and validation.

Radiographs Pre-processing. The radiographs were prepared as described above to training a U-Net neural network[36]. We designed a U-Net neural network for bone segmentation that operates on grey images sized 768x1024px (Figure 4). In contrast to the U-Net proposed by Ronneberger *et al* our network is symmetric one, i.e. the input image size is equal to output map size, it performs multi-class segmentation, and relies on the Dice loss and score for training and evaluation, respectively. The accuracy of the bone segmentation was evaluated using SDI which is the most used metric in medical image segmentation[37, 39]. We assumed that the threshold SDI should have a mean greater than 97%, with a minimum value greater than 92%. SDIs were determined only for the three bones required to estimate HVA/IMA. During U-Net training, the calculated SDI was used to check whether U-Net training should be stopped or continued on an extended training subset containing more images. After U-Net training was complete, the SDI was checked on the testing subset to verify whether the U-Net achieved the required generalizability. The initial training subset consisted of 50 anteroposterior foot radiographs with corresponding bone masks and labels. The initial training set was increased by 10 images after each training round until the threshold SDI was achieved on the validation subset (Figure 3C). The validation subset was fixed during training and used to compare the segmentation abilities of networks trained on incrementally larger training subsets. The threshold SDI was achieved on a training set of 150 radiographs with corresponding bone masks and labels. After adding an additional radiographs, the final training set consisting of 161 training images and 20 validation images (90% and 10%, respectively) was used to train the final U-Net. The dataset is available upon request.

Architecture and training U-Net. The neural network for bone segmentation follows the standard U-Net architecture established by Ronneberger *et al*[36]. Each U-Net encoder and decoder contains four layers (Figure 5). The validation SDI was calculated at the end of each epoch during the training of the U-Net, and the training was stopped when the SDI did not increase over 10 following epochs. This served as an early stop

technique to avoid overfitting, where the value of early stop (patience) was set to 10. The U-Net was trained using Adam optimizer with Dice loss, learning rate (LR) set to 0.0001 (with reducing LR on plateau) and batch size equal to 8. The number of epochs was set to 80, and a callback was used to save the best U-Net model and its weights. The training data was augmented using mirroring, rotations, and contrast enhancement. Training of neural networks was performed on NVIDIA A100 GPU, whereas the testing was performed on the notebook's GPU (RTX2060).

Final validation of U-Net. Before measuring HVA/IMA, the trained U-Net was evaluated on the testing subset to assess its generalizability (Figure 3D). As the average SDI was larger than 97% on the testing subset with a minimal score larger than 92%, we used the trained U-Net to segment bones on all test radiographs (Figure 3E). In image post-processing, small holes in bones segmented by the U-Net were filled using morphological operations, and artifacts such as small blobs were deleted. Our algorithm first segmented and labeled bones that it then used to automate determining reference points and HVA/IMA measurements (Figure 3F, and Figure 5). The programmer who trained the U-Net did not participate in manual measurements of HVA/IMA and did not see any results before statistical analysis.

Measurement of HVA and IMA

Automatic Determination of Reference Points and Angles. Using the anteroposterior feet radiographs, the U-Net segmented bones and labelled them with different colors (Figure 4). Using these labels our algorithm selected three bones of interest: 1,2MT, and hallucial PP. HVA/IMA were automatically measured using reference points from these bones.

According to AOFAS (Figure 1) all reference points on the 1,2MT, and hallucial PP are the metaphyseal/diaphyseal points from which a guideline had to be determined to automate measurement of bone axes[18]. The final bone split ratios were selected following various combinations of bone split ratios to obtain the points closest to the

diaphysis (Figure 5). For the 1MT, the reference points were located at 0.3 of the bone length proximal to the distal articular surface and at 0.25 of the bone length distal to the proximal articular surface. For the 2MT: 0.30 and 0.10, and for the hallucial proximal phalanx: 0.25 and 0.25, respectively.

STATISTICAL ANALYSIS

Eighty-four radiographs of patients were used to measure HVA/IMA both manually by clinicians (reference method) and automatically by our algorithm. The reliability of the measurements between these two approaches was calculated using ICC and the standard error for a single measurement (SEM). Manual measurements (HVA/IMA) were performed by: an orthopedic surgeon (O_A) with 7 years' experience and repeated at 2 mo in blinded test (O_{A1} and O_{A2}), an orthopedic surgeon (O_B) with 15 years of experience; and by musculoskeletal radiologist (R) with 15 years of experience. Interobserver and intraobserver reliability coefficients (ICC) were calculated. The observers were not aware of any clinical results. Assessment of the HVA/IMA was performed according to the guidelines of the AOFAS ad hoc Committee on Angular Measurements (Figure 1) and digital technique using Radiant/Carestream[7, 18, 40]. Our algorithm then classified the appropriate severity (Figure 2) and operative decisions were compared against the orthopedic surgeon (O_{A2}). All statistical calculations were performed using MedCalc.

RESULTS

We proposed a novel automated HVA/IMA measurement method using deep learning algorithms. To measure these angles, the 1,2MT, and hallucial PP bones were automatically segmented, with reference points then automatically assigned. We obtained high interobserver and intraobserver correlations between manual measurements of HVA and IMA, and great agreement between AI (our algorithm) and clinician angle measurements (Table 2). We analyzed HVA and IMA measurement errors for each patient's radiograph finding (Figure 6, Figure 7). Standard Error of the

Mean for HVA was 0.26 and 0.16 for IMA. The accuracy of algorithmically measured angles is similar to that of orthopedic surgeons.

A decision system was developed and tested according to the EFORT operative treatment algorithm (Figure 2)[35]. Operative decisions were taken (D1- chevron, D2- chevron or scarf, D3- scarf or Lapidus) based on calculated angles. The AI decisions were compared to O2 decisions for concordance. The agreement of clinician decisions was also compared. The ratio of same pre-operative surgical decisions among AI and O_{A2} was almost 0.80 (67/84), which was higher than the ratio among clinicians (Table 2). A key achievement of our algorithm is that it saves radiologist and orthopedic surgeon's time while providing a clinically actionable HVA/IMA measurement that supports preoperative planning.

DISCUSSION

Some initial work on deep-learning radiographic and WBCT foot analysis was recently published[41-44]. While WBCT is arguably the future of hallux valgus preoperative qualifications, X-ray remains the standard as it is cheap, and widely available for symptomatic HV[38, 45]. This work is in line with emerging research and substantially improves upon our previous algorithm. As demonstrated experimentally, the proposed approach can estimate HV angles on high-resolution radiographs and classify the severity of HV as a preoperative decision-making tool. Moreover, this work may expedite novel developments in forefoot surgery. This will provide a reliable opportunity to compare preoperative and postoperative measurements and analyze the effects of surgical correction to produce better HV treatment standards.

Coughlin *et al* found that only 83.8% of IMA measurements made by physicians were within 3 degrees of concordance. AI overcomes the issue of clinician intra-observational and inter-observational reliability in terms of repeatable angular measurements of HV[19, 46].

Considering that collecting a dataset of radiograms of patients with HV who were subsequently operated is not easy, we decided to rely on HV measurements on the segmented bones. Our initial research demonstrated that such an approach permits achieving better accuracy of HVA/IMA measurements on limited numbers of radiographs compared to key point-based ones. The difficulties associated with segmentation of proximal epiphyseal of 2MT bones due to anatomical overlap inclined us to apply a simplified segmentation with the exclusion of this bone area (Figure 5 C). Consequently, the IMA measurements have an irrelevant bias (Figure 6 B, Figure 7 B).

Foot surgeons are aware that the decision to perform osteotomies or first tarsometatarsal joint (TMTJ) fusions (Lapidus procedure) depends on more than just HVA and IMA. Rather, it depends on the patients' clinical picture, concomitant deformities of the foot such as lesser toe deformities, pes planus, metatarsus adductus, first TMTJ instability, the width of the 1st MT shaft, pronation of the first ray, presence of first metatarsophalangeal joint osteoarthritis, and the surgeons' own skill level. The lateral view is also critical in evaluating the first TMTJ instability or presence of osteoarthritis which may necessitate a fusion rather than a 1st MT osteotomy. According to Lee *et al* the HVA, IMA, interphalangeal angle, sesamoid rotation angle, and first metatarsal protrusion distance are worth measuring in HV considering three-dimensional role in this deformity[11]. Presently the above-mentioned requirements may limit the applicability of our method in some cases. Nonetheless, our algorithm establishes itself as a fast and clinically effective tool in the assessment of many HV cases. In order to fully automate preoperative HV planning, further research and development remain necessary.

In future work, more radiographs will be labeled to train more advanced U-Net to distinguish bones under challenging areas better. A multi-center database of radiographs should be created. Due to recent developments and a deeper understanding of pronation, enhanced segmentation and further research on manual and automatic estimation of the distal metaphyseal/diaphyseal 1MT reference point is

necessary. We plan to train and evaluate different networks on our dataset, which will be extended to new images from other hospitals.

CONCLUSION

The proposed automated, AI-assisted determination of hallux valgus angles based on deep learning is an accurate tool that produces measurements comparable to manual measurements performed by experienced clinicians in significantly less time. Automation can be used in clinical practice to determine hallux valgus angles on X-ray images, classify the degree of deformity, and streamline preoperative decisions-making prior to HV surgery.

ARTICLE HIGHLIGHTS

Research background

Recent advances in artificial intelligence and deep learning has spurred innovations in medical imaging modalities, resulting in enhanced visualisation possibilities. Additionally, there is a growing interest in the automation of regular diagnostic procedures alongside orthopedic measurements.

Research motivation

So far, no reliable and automated method has been developed for measuring angles of foot bones in significant deformities of the big toe from radiographs according to AOFAS. Likewise, there is no system for automated preoperative decision-making.

Research objectives

The aim of our research was to develop a robust automated method for measuring angles of hallux valgus on radiographs according to AOFAS guidelines, to determine the accuracy of this method, to compare it against expert clinician measurements, and to develop a preoperative decision-making systems.

Research methods

The bones which are necessary to determine the angles of hallux valgus, obtained on anteroposterior weight-bearing feet radiograms were segmented by a U-Net. The bone axes were determined, and then the reference points for determining the hallux valgus angles (HVA) and intermetatarsal angles (IMA) were found. The interclass correlation coefficient and standard error for single measurements were used to calculate the agreement between manual and automatic measurements. Finally, the correlation between the decisions of our algorithm and clinical adjudication for preoperative planning of hallux valgus was investigated.

Research results

The key foot bones were segmented from anteroposterior feet radiograms by the U-Net neural network with high accuracy (average Sørensen–Dice index larger than 97%). Such a precise segmentation enabled the accurate determination of bone axes and the required reference points. Excellent agreement was achieved between manual and automated measurements of both angles. For HVA, absolute agreement interclass correlation coefficient (AA-ICC) and consistency ICC (C-ICC) were 0.97, and standard error of measurement (SEM) was 0.32. For IMA, AA-ICC was 0.75, C-ICC was 0.89, and SEM was 0.21. The proposed hallux valgus treatments based on HVA and IMA measured automatically correlated well with those proposed by orthopedic surgeons performing manual angle measurements.

Research conclusions

The proposed AI-powered automation for evaluating angles of hallux valgus through deep learning is a precise, yielding measurements akin to those conducted manually by experienced clinicians. This offers promising clinical applications such as facilitating the automated determination of angles of hallux valgus from X-ray images, categorizing the extent of deformity, and recommending a specific protocol for corrective surgery.

Research perspectives

Future research will focus on automating the measurements of remaining angles and parameters of forefoot deformation along its greater clinical implementation to further enhance diagnostic accuracy and improve patient outcomes.

87733_Auto_Edited.docx

ORIGINALITY REPORT

4%
SIMILARITY INDEX

PRIMARY SOURCES

1	Pieter van der Woude, Stefan B. Keizer, Martine Wever-Korevaar, Bregje J.W. Thomassen. "Intra- and Interobserver Agreement in Hallux Valgus Angle Measurements on Weightbearing and Non-Weightbearing Radiographs", The Journal of Foot and Ankle Surgery, 2019 Crossref	40 words — 1%
2	journals.sagepub.com Internet	21 words — 1%
3	www.ncbi.nlm.nih.gov Internet	21 words — 1%
4	www.wjgnet.com Internet	18 words — < 1%
5	Jonathan Day, Cesar de Cesar Netto, Martinus Richter, Nacime Salomao Mansur et al. "Evaluation of a Weightbearing CT Artificial Intelligence-Based Automatic Measurement for the M1-M2 Intermetatarsal Angle in Hallux Valgus", Foot & Ankle International, 2021 Crossref	17 words — < 1%
6	Seung Min Ryu, Keewon Shin, Soo Wung Shin, Seungjun Lee, Namkug Kim. "Enhancement of evaluating flatfoot on a weight-bearing lateral radiograph of the foot with U-Net based semantic segmentation on the long	16 words — < 1%

axis of tarsal and metatarsal bones in an active learning manner", Computers in Biology and Medicine, 2022
Crossref

7 ds.inflibnet.ac.in
Internet
16 words — < 1%

8 journals.lww.com
Internet
13 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS