99359_Auto_Edited.docx

WORD COUNT 3298
TIME SUBMITTED 25-SEP-2024 01:25PM
PAPER ID 111925883

Name of Journal: *World Journal of Experimental Medicine*
Manuscript NO: 99359
Manuscript Type: ORIGINAL ARTICLE

Retrospective Study

HIPPO AI: Correlating automated radiographic femoroacetabular measurements with patient-reported outcomes in developmental hip dysplasia

AI Hip Correlations with Patient Reported Outcomes

Ahmed Alshaikhsalama, Holden Archer, Yin Xi, Richard Ljuhar, Joel E. Wells, Avneesh Chhabra

Abstract

BACKGROUND

Hip dysplasia (HD) is characterized by insufficient acetabular coverage of the femoral head, leading to a predisposition for osteoarthritis. While radiographic measurements such as the Lateral Center-Edge Angle (LCEA) and Tönnis angle are essential in evaluating HD severity, patient-reported outcome measures (PROMs) offer insights into the subjective health impact on patients.

AIM

To investigate the correlations between machine-learning automated and manual radiographic measurements of HD and PROMs with the hypothesis that Artificial intelligence (AI)-generated HD measurements indicating less severe dysplasia correlate with better PROMs.

METHODS

Retrospective study evaluating 256 hips from 130 HD patients from a hip preservation clinic database. Manual and AI-derived radiographic measurements were collected and PROMs such as the Harris Hip Score (HHS), iHOT-12, SF-12, and Eq-VAS survey were correlated using Spearman's rank-order correlation.

RESULTS

The median patient age was 28.6 years (range 15.7- 62.3 years) with 82.3% of patients being women and 17.7% being men. The median interpretation time for manual readers and AI ranged between 4 - 12 minutes per patient and 31 seconds, respectively. Manual measurements exhibited weak correlations with HHS, including LCEA (r = 0.18) and Tönnis angle(r = -0.24). AI-derived metrics showed similar weak correlations, with the most significant being Caput-Collum-Diaphyseal (CCD) with iHOT-12 at r = -0.25 (P = 0.042) and CCD with SF-12 at r = 0.25 (P = 0.048). Other measured correlations were not significant (P > 0.05).

CONCLUSION

This study suggests AI can aid in HD assessment, but weak PROM correlations highlight their continued importance in predicting subjective health and outcomes, complementing AI-derived measurements in HD management.

Key Words: Hip Dysplasia; Patient reported outcome measures; Deep-learning; Artificial intelligence; Radiographs; Lateral Center Edge Angle

Alshaikhsalama A, Archer H, Xi Y, Ljuhar R, Wells JE, Chhabra A. HIPPO AI: Correlating automated radiographic femoroacetabular measurements with patient-reported outcomes in developmental hip dysplasia. *World J Exp Med* 2024; In press

Core Tip: In this study, we compared an artificial intelligence (AI) tool measuring anteroposterior hip radiographs against manual readers for assessing hip dysplasia (HD) associations with patient-reported outcome measures (PROMs). The AI tool, HIPPO, efficiently generated radiographic measurements but showed poor correlations with PROMs, highlighting its current limitations in predicting clinical outcomes solely from radiological data. This indicates that while AI can aid radiographic assessments, PROMs remain crucial for capturing subjective patient experiences. The findings underscore the importance of integrating PROMs as an additional element in the clinical decision-making processes for HD, while also incorporating efficient radiographic assessment by AI tools.

INTRODUCTION

Acetabular or hip dysplasia (HD) is a developmental condition that is characterized by a shallow or upsloping acetabulum that can be accompanied by femoral head incongruency[1]. HD often presents in the pediatric and adult population with symptoms of hip pain and/or instability. When left untreated, it can lead to hip

osteoarthritis (OA) due to stress overload, shear forces, and improper mechanics progressively affecting joint cartilage[2]. Several conservative and surgical treatment options currently exist; among them, the most used modalities include physical therapy and lifestyle modifications, periacetabular osteotomy (PAO), hip arthroscopy, and total hip arthroplasty. The treatment modality chosen depends upon the time of discovery, symptom severity, and status of the hip labrum and cartilage, and functional disability[3-5].

Hip radiographs are the current gold standard for the initial screening and assessment of HD[6].There are a multitude of validated diagnostic radiographic measurements employed to assist the diagnosis of HD. Among them, Lateral Center Edge Angle (LCEA) is most commonly used, as measured on a standing anteroposterior (AP) pelvis radiograph[7]. Additionally, the Tönnis angle and extrusion index are also commonly used in clinical practice[8]. Following radiographic assessment, advanced imaging such as MRI or CT can be used for pre-operative planning and further assessment of the health of the labrum or hyaline cartilage[9].

While a diagnosis of HD is established by a combination of clinical presentation, examination findings and radiographic measurements, patient reported outcome measurements (PROMs) are equally important to illustrate the perception of patients' subjective hip health status[10]. These are gleaned from different surveys administered at the time of clinical presentation, such as the Harris Hip Score (HHS), International Hip Outcome Tool (iHOT-12), Visual Analog Scales for Pain (VAS), Eq-VAS (health status), and SF12 (quality of life), among others. Each patient reported outcome survey provides a different evaluation of the patient's condition. For instance, the Harris Hip score is a reliable indicator for patient function, while iHOT-12 provides a good indication for quality-of-life changes[11-13].

PROMs have become increasingly important in evaluating indications for treatment and prognosis for HD patients.[14-16] Despite their common use in the clinical evaluation of patients with HD and pain, the International Hip-related Pain Research Network meeting in 2018 ruled that more studies are needed to further evaluate the

usefulness of PROMS.[17] Thus, it is important to examine the relationships between validated radiographic HD measurements and PROMs[11]. One prior study evaluated the by Takegami *et al*[18] evaluated the relationship between manual individual radiographic parameters with the patient-reported outcome measurements in Japanese patients. However, it is time consuming to routinely measure the above-described parameters, let alone control for the associated inherent reader variance and need to remember how to obtain such parameters. If these measurements could be automatically produced by machine learning using artificial intelligence (AI), the clinical note and/or radiographic interpretation report could be auto-populated. In addition, the correlations between radiographic parameters and PROMs can be studied in a more standardized manner and for longitudinal data collection. To that end, AP radiographic measurements can be auto-evaluated by HIPPO software, which is a validated AI hip measurement tool validated in a European study and Conformite Europeenne (CE) certified [ImageBiopsy Lab Inc. (Vienna, Austria)].[19] Yet, it is not known how these standardized deep-learning software generated measurements obtained in the US population correlate with their PROMs data. Additionally, it is not known if a validated AI tool can assist in predicting PROMs data and providing comprehensive evaluation for HD patients.

Our hypothesis was that AI-generated HD measurements indicating less severe dysplasia correlate with better PROMs. Thus, the aim was to assess the correlation between AI-derived hip measurement and initial PROMs in a consecutive series of patients. This is the first study to evaluate manual and AI measures of radiographs in patients with hip dysplasia and associate radiographic findings with preoperative PROMs data.

MATERIALS AND METHODS

IRB approval was received for retrospective use of a longitudinally gathered patient registry data and surveys. Anonymous survey data involving PROMs was collected in our institutional hip preservation practice. All HIPAA regulations were followed.

Patients

Using our anonymized electronic database of patients who visited the institutional hip preservation clinic, we identified 325 hips from 276 patients with a complete radiographic series from December 2016 to December 2021. Each patient had a reference final HD diagnosis based on consensus radiographic opinions of an independent fellowship trained musculoskeletal radiologist and hip preservation surgeon using the 4-view radiographic series (AP pelvis, 45° Dunn, Frog-leg lateral, and false profile views) and clinical findings. Only patients with a concordant final diagnosis of hip dysplasia were included in this study, resulting in 256 hips from 130 patients. Six of the 136 patients did not return an output from HIPPO (Figure 1). The hips with prior surgical interventions or avascular necrosis were excluded. Patient demographic data including age, gender, and BMI were extracted from the electronic health records. Additionally, dates of the patient's first office visit and survey, along with the dates and details of any surgeries were collected. The surveys were obtained at the time of the initial clinic visit when the radiograph was obtained to avoid delay between imaging and initial PROM survey.

Patient Reported Outcome Measures (PROMs)

The patients were surveyed at the time of their initial office visit, which included HHS, iHOT-12, SF-12, and Eq-VAS as shown in Table. 1. Survey data was obtained using an online REDCap form and was retrieved into an excel document for each of the included deidentified study patients. Each survey result was manually calculated and normalized to 100% by two medical students under the training and supervision of the senior orthopedic hip specialist.

Manual measurements

Tonnis grade of hip OA was evaluated in all cases by the senior orthopedic surgeon. Manual HD measurements were obtained as a control for the AI measurements.

Measurements were taken for each patient by three readers under the supervision and training of a senior MSK radiologist. The three readers underwent extensive training under the MSK radiologist and were assessed for accuracy on a series of training images before obtaining the measurements for the study. The study measurements were then averaged and correlated with PROMs (Table 1). Time required to assess these measurements was recorded using a stopwatch from the time images were loaded on IntelliSpace Picture Archiving and Communication System (IPACS, Philips, Best, Netherlands) to completion of the reads using a built-in measurement tool. Measurement data from the AI algorithm and manual measurements with their detailed inter-reader and inter-modality correlations between manual measurements and AI was published and showed good to excellent inter-method reliability for common HD landmarks including LCEA and Tonnis angle[19].

AI measurement tool – HIPPO

'HIPPO' is an AI deep-learning software [ImageBiopsy Lab Inc. (Vienna, Austria)] that automatically locates anatomical landmarks on AP full leg standing radiographs. Using these landmarks, the tool measures various radiographic parameters. These parameters are LCEA, Tönnis Angle, Sharp Angle, Caput-Collum-Diaphyseal (CCD) angle and pelvic obliquity (Table 2, Figure 2). The software accepts images in DICOM format and returns a DICOM compatible AI report. When the software returns an error report or does not return a report at all, a software failure is indicated. A software failure could be due to errors in the software itself or anatomical subtleties in the radiograph that heavily affected how the software interprets the images. All images in the study were securely transferred to the PACS server at our institution, and from there were pushed to a local installation of the AI software. Measurements were then downloaded onto an excel document after being processed through the software (Windows 11, Microsoft, Redmond, WA). In our study, the median HIPPO reading time per patient was 41 seconds.

Statistical analysis

Descriptive statistics were calculated for patient demographics. All hip measurements were on per-hip level while PROMs except HHS were on per-patient level. Therefore, one hip from each patient was selected when comparing hip measurements to iHOT-12, SF-12, and Eq-VAS. The hip with the worst mean LCEA score from the 3 readers was selected. Correlations between hip measurements and HHS were calculated on the same selected hips. Spearman's rank correlation coefficients were reported with corresponding 95% confidence intervals. Hypothesis tests for non-zero correlation were conducted at a 0.05 significance level. P-values were adjusted for False Discovery Rate (FDR) *via* the Benjamini & Hochberg method for each PROM. Correlation coefficients were interpreted as negligible: 0-0.1, weak: 0.1-0.39, moderate 0.4-0.69, strong: 0.7-0.89 and very strong: 0.9-1[25]. With 80% power to detect a correlation of at least 0.26 at 0.05 significance level, the study needed 130 patients before adjustments for multiple comparisons.

RESULTS

Patients

Descriptive statistics were calculated for appropriate demographic factors. The median patient age was 28.6 years with a maximum of 62.3 years and a minimum of 15.7 years. 82.3% of patients were women and 17.7% were men. The BMI ranged from 17 kg/m^2 to 38 kg/m^2, with 24 kg/m^2 as the median. An orthopedic surgeon classified the hips according to the Tönnis grade. The median Tönnis grade was 0 with the majority (204 hips, 79.7%) having Tönnis grade 0, 51 hips (19.9%) with Tönnis grade 1, and 1 hip (0.4%) with Tönnis grade 2.

Manual measurements

Measurement data from the AI algorithm and manual measurements showed good to excellent inter-method reliability for common HD landmarks including LCEA and

Tonnis angle. The median read time for manual readers ranged between 4 and 12 minutes per patient[19].

Manual hip measurements vs PROMs

The largest estimated correlation coefficients were between LCEA and HHS (0.18 (0.00, 0.35)), Tonnis Angle and HHS (-0.24 (-0.40, -0.06)), CCD and SF-12 (0.19, (0.01, 0.36)), and CCD and iHot-12 (-0.19, (-0.36, 0.00)); however, these weak correlations were not significant at a 0.05 Level after adjustment for multiple comparisons (Table 3). No other significant correlation was observed between the remaining manual measurements and PROMs. A scatter plot is shown in Figure 3.

AI hip measurements vs PROMs

CCD were significantly correlated with iHot12 and SF12, but the correlation strength was weak (CCD *vs* iHot12: -0.25 (-0.42, -0.07), adj. P = 0.042; CCD *vs* SF12: 0.25 (0.07, 0.42), adj. P = 0.048). Other notable correlations of similar magnitude were estimated for Obliquity and Eq-VAS (-0.22, (-0.39, -0.4)), as well as Tonnis angle and HHS (-0.20, (-0.36, -0.02)); however, these estimates were not significant at a 0.05 Level after adjustment for multiple comparisons (Table 4, Figure 4).

PROMs

HD patients before intervention had an average survey scores of 69% EqVAS suggesting moderate pain[26] and 63% SF-12, which is slightly above the depression threshold[27]. They also had 61% iHot-12, which is nominally above the acceptable symptom threshold (PASS) of 59% indicating the patients had a greatly affected quality of life[28], and 62% HHS, which is poor function as defined by the standard less than < 70%[29].

DISCUSSION

This study aimed to evaluate the correlation between AI-generated radiographic measurements and patient-reported outcome measures (PROMs) in individuals with hip dysplasia (HD). Our findings suggest that while there is a presence of weak correlations between certain AI-derived radiographic measurements and PROMs, these relationships did not achieve statistical significance after adjustments for multiple comparisons. This indicates that the current capacity of AI, specifically the HIPPO deep-learning software, to predict clinical outcomes based on radiological data is limited, although not entirely negligible.

HIPPO is a novel tool for acquiring rapid hip measurements, successfully processing most cases with notable efficiency as reported previously[30]. Where manual readers required a median time of 6 minutes and 48 seconds per hip, and trained radiologists require on average 83 seconds per AP hip radiograph, the AI completed the same task in an average of 41 seconds, highlighting a significant reduction in time and cost per radiograph[19, 30]. In this study, HIPPO AI found a significant association between the CCD angle and iHot-12/SF-12 PROMs compared to manual readers. An elevated CCD angle (Coxa Valga) has been associated with hip dysplasia, although it is a less commonly used measurement diagnostically[6, 31]. While the exact reason for this significant association is not known, the authors hypothesize that the difficulty of measuring CCD among manual readers compared to a standardized AI tool introduced sufficient variation to prevent an observed association[32]. These findings further highlight the importance of standardization in assessment and interpretation of radiographic measurements. The results of this study differ from those of Takegami *et al*[18], where the LCEA angle in 108 Japanese HD patients was independently associated with the Japanese Orthopaedic Association's hip disease questionnaire. However, the end point PROMs examined in our study were different and applied to a heterogeneous U.S. population, limiting direct comparison. Despite the potential for AI to streamline clinical workflow, our study highlights the difficulty and current unfeasibility of correlating radiographic findings with patient-centric outcomes such as PROMs. Although HIPPO is efficient at measuring, it may require more training to

recognize patterns that better match patients experience. This highlights an area where AI can develop to become more clinically meaningful.

An additional consideration is our patient cohort. Overall, the study's patient cohort was symptomatic, presenting with moderate pain, slightly above the depression threshold, and poor functional scores as per EqVAS, SF-12, iHot-12, and HHS, respectively[26-29]. The homogeneity of this group may have diluted the potential to discern a stronger correlation between radiographic measurements and PROMs. Including asymptomatic individuals in future studies may provide a broader spectrum of disease and potentially unveil more defined associations.

It is important to note the subjective nature of PROMs and their potential to be affected by factors beyond the HD diagnosis. For instance, while HHS mainly measures hip function, SF-12 encompasses wider quality of life and mental health parameters, which can be affected by multiple socio-economic and demographic factors[33]. Similarly, individual variability in physical fitness and factors such as hamstring strength play a role in hip stability and perceived symptoms and functionality, contributing to an observed variability in PROMs that may make it difficult to correlate any radiographic measurement, no matter the tool used[34, 35].

The weak correlations observed challenge our initial hypothesis that improvements in HD radiographic measures would linearly correlate with better PROMs. The authors do not believe that these weak correlations are due to inaccuracies in the AI measurement tool, which was previously validated by Archer *et al*[19] revealing moderate to strong associations with trained manual readers. Additionally, the vast majority of observed correlations were nonsignificant and contained similar results to the manual readers, with exception of CCA angle and certain PROMs on AI reads, thus suggesting a similar radiographic accuracy between groups as previously described. These results call into question the clinical utility of radiographic measurements alone in predicting patient-reported outcomes and highlights the complexity of HD as a disease entity. While AI can rapidly provide quantitative data valuable for initial screenings and monitoring disease progression, it should complement—not replace—

PROMs, which encapsulate the patient's subjective experience and the functional impact of the disease. PROMs remain essential for capturing the holistic impact on quality of life, guiding more personalized treatment approaches. Therefore, clinicians are encouraged to use various means of information-gathering including the use of PROMs. They capture a spectrum of patient experiences and outcomes that are not obvious through radiographic data, reinforcing their role in comprehensive care for patients with HD.

Our study has several limitations. The gender distribution in our study was predominantly female, reflecting the higher incidence of HD in women[36]. This distribution may influence the correlations observed and thus may not be generalizable to a male population. Additionally, most participants were middle-aged adults, so our results might not reflect the bone density and joint health variations found in older patients, and thus may affect the generalizability of this study[37]. Finally, the manual measurements, while performed by medical students under the supervision of an MSK radiologist, are not immune to human error. Anatomical variability might have led to inaccuracies; however, extensive training aimed to mitigate such errors, and their impact on the study's validity is considered minimal. Future studies should also incorporate prospective clinical validation studies to assess AI tools against traditional radiographic measurements, post-implementation in patient care settings. Additionally, randomized controlled trials comparing patient outcomes using AI-derived data with those using manual radiographic assessments are critical to establish the effectiveness of AI in clinical decision-making for hip dysplasia.

CONCLUSION

In conclusion, this study validated fast measurements using AI-software. Some correlations between AI-derived radiographic measurements and PROMs were seen in HD patients but these findings are mostly insignificant and weak, with most of the associations mirroring that of manual readers. Thus, at present, AI interpretations of radiographic data should be used with caution when predicting patient-reported

outcomes. The potential of AI in clinical decision-making for HD patients remains promising in providing quick and accurate radiographic hip measurements. AI software has massive potential in streamlining physician workflow and in performing measurements that can have influence on the clinical decision-making process for patients with HD. It is through these continued efforts that we may fully realize the role of AI in the management of HD, while PROMs will continue to play a crucial role in assessing the broader implications of treatment on patient quality of life.

99359_Auto_Edited.docx

ORIGINALITY REPORT

3%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet	41 words — 1%
2	mail.meddocsonline.org Internet	29 words — 1%
3	Atlal M. Abusanad, Omar Iskanderani, Marwan R. Al-hajeili, Reem Ujaimi, Rolina Alwassia. "Survival in patients with brain metastasis secondary to breast cancer from Saudi Arabia.", Journal of Clinical Oncology, 2023 Crossref	13 words — < 1%
4	academic-accelerator.com Internet	13 words — < 1%

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