


Intra- and interobserver reliability for predicting hip preservation versus hip arthroplasty utilizing plain radiographs with comparison of surgeon specialization

Kyle Schultz ¹, Jeff Osborne², Karen Nelson³, Vishnu Potini²,
Chaoyang Chen², Andrew Aljuni^{4,5}, Asheesh Bedi⁶, James Bookout^{4,5},
Michael Yusaf^{2,7}, and Shariff K. Bishai^{2,4,5,8*}

¹Ascension Genesys Regional Medical Center, Ascension Genesys Hospital, 1 Genesys Parkway, Grand Blanc, MI 48439, USA,

²Department of Orthopedics and Sports Medicine, Detroit Medical Center – Harper Hospital, 3990 John R, Box 137, Detroit, MI 48201, USA,

³Henry Ford Health System, Henry Ford Macob Hospital – Clinton Township, 15855 19 Mile Rd, Clinton Twp., MI 48038, USA,

⁴Associated Orthopedists of Detroit, PC, 24715 Little Mack Avenue Suite 100, St. Clair Shores, MI 48080, USA,

⁵Oakland University William Beaumont School of Medicine, 586 Pioneer Dr., Rochester, MI 48309, USA,

⁶MedSport. University of Michigan, 4008 Ave Maria Dr A-1000, Ann Arbor, MI 48105, USA,

⁷Center for Advanced Orthopedics, 3100 Cross Creek Pkwy, Auburn Hills, MI 48326, USA and

⁸Michigan State University College of Osteopathic Medicine, 909 Wilson Rd, Room B305, East Lansing, MI 48824, USA.

*Correspondence to: S. K. Bishai. E-mail: skbishai@yahoo.com

Submitted 25 May 2019; Revised 28 December 2019; revised version accepted 9 January 2020

ABSTRACT

Surgeon subspecialty training and practice landscape are formative in diagnostic evaluation and treatment recommendations. Varying recommendations can have substantial impact on patients' care pathways and outcomes. We investigated intra- and interobserver reliability of treatment predictions for total hip arthroplasty (THA) between surgeons performing arthroplasty and/or arthroscopic hip preservation surgery. Anterior–posterior (AP) hip radiographs cropped to include the lateral sourcil, medial sourcil and foveal region of 53 patients with Tönnis Grade 0–3 were evaluated by five surgeons (two performing arthroplasty, two performing arthroscopic hip preservation and one performing both interventions). Surgeons predicted THA versus no THA as the treatment for each image. Predictions were repeated three times with image order randomized, and intra- and interobserver reliability were calculated. Surgeons were blinded to patient characteristics and clinical information. Interobserver reliability was 0.452 whereas intraobserver reliability ranged from 0.270 to 0.690. Arthroscopic hip preservation surgeons were more likely to predict THA (36.9%) than arthroplasty surgeons (32.7%), $P = 0.041$. Intra- and interobserver reliabilities of surgeons predicting THA versus no THA based on an AP hip radiograph were average at best. Arthroscopic hip preservation surgeons were more likely to predict THA than arthroplasty surgeons. Subjective surgeon interpretation can lead to variability in recommendations to patients; potentially complicating care pathways.

INTRODUCTION

The treatment of hip pain in the setting of early degenerative changes continues to be an extensively investigated topic. Hip arthroscopy has been gaining popularity for treatment of intra- and extra-articular causes of hip pain

with an increase of 250% between 2007 and 2011 [1, 2]. As surgical volume continues to grow, the indications and outcomes need to be continually evaluated. Several studies have demonstrated the importance and reliability of measurements for the diagnosis of femoroacetabular

impingement (FAI) as well as predictive recommendations for treatment success [3–7]. Precise evaluation of osteoarthritis (OA) within the hip is also paramount to successful outcomes as patients have been reported to have significantly increased likelihood of conversion to total hip arthroplasty (THA) when there were degenerative changes at time of arthroscopy [4–6, 8–10].

In his original studies from 1972, Tönnis characterized hip OA during investigations into hip dysplasia. These investigations specifically focused on the sourcil region of the femoroacetabular articulation, allowing precise evaluation and characterization of hip arthritic disease [11, 12]. In multiple studies, Philippon further evaluated this region to determine key radiographic findings and their impact on patient outcomes. His group's landmark findings suggest patients with arthritic disease at the sourcil region, specifically Tönnis Grades 2 and 3, are 4.8 times more likely to require THA [4]. In addition, patients with <2 mm of joint space at any position along the sourcil demonstrate decreased functional outcomes and upwards of 50% conversion to THA within 3 years of a hip arthroscopic procedure [5, 6]. In 2019, a consensus project aimed at identifying best practice guidelines for arthroscopy in FAI patients demonstrated the importance of identifying pathology in this load-bearing area of the femoroacetabular articulation prior to treatment. Furthermore, they concluded arthroscopic intervention in patients with progressive arthritic disease offers little to no clinical benefit and provides a significantly increased risk for THA conversion [13]. Thus, the ability to identify radiographic disease prior to hip preservation surgery continues to be supported in the literature as vital to patient outcomes.

There are several factors involved in the construction of a well-formulated treatment plan (i.e. detailed history, clinical exam and surgeon experience); however, the appropriate use and interpretation of imaging remain keystones of an orthopedic evaluation. Several studies have supported radiographic measurements including alpha angle, lateral-center-edge angle and head-neck offset ratio as reliable between surgeons [14, 15]. Conversely, Carlisle *et al.* [16] found that although the same measurements were reliable within a single provider, they become less reliable when compared with other surgeons. Collectively, these studies consistently demonstrated low interobserver reliability for Tönnis grading [14–16]. While radiographic interpretation provides information toward decision-making, low reliability between surgeons allows for increasing opportunity for variation in treatment recommendations. In addition, as surgeons draw on personal experience, it is possible that bias from sub-specialization may further impact their interpretations and recommendations.

We aim to evaluate if there are variances between surgeons with differing areas of expertise in the treatment of hip pain. Our primary objective was to evaluate for inter- and intraobserver variability in predictive THA recommendations based on limited imaging of the hip for surgeons performing hip arthroplasty, arthroscopic hip preservation or a combination of both interventions.

MATERIALS AND METHODS

Study design

International review board approval was obtained prior to the start of our retrospective cohort study. Sixty patients were selected after either undergoing a THA or arthroscopic hip preservation procedure between 11/2014 and 8/2015. A supine anterior–posterior (AP) radiograph of the operative hip was obtained from their initial presentation to the clinic. Tönnis grading was performed by V.P. on a web-based PACS software (PaxeraUltima, Paxera Health, Newton, MA, USA) as previously described [5]. V.P. was blinded to surgical intervention for all patients prior to grading. Patients who underwent a previous hip surgery (excluding intra-articular injections) or exhibited Grade 4 Tönnis changes were excluded from the study population. Image captures were obtained from each radiograph to include only the sourcil region of the hip and were provided to the surgeons for evaluation (Figure 1).

Two surgeons performing arthroscopic hip preservation, two performing arthroplasty and one who performs both interventions participated in the study. Surgeons were notified that the patients in the study presented for a complaint related to the hip that failed conservative management, and were asked to determine, based on the image, if they would recommend a THA. The surgeons were blinded to patient sex, age, laterality, all clinical evaluations and end treatment intervention. No history, physical exam findings or any other information beyond the single cropped radiograph were provided. Each surgeon evaluated the images three times for determination of intraobserver variation. There was no direction given to the surgeons on how to evaluate the image (i.e. joint space measurements, Tönnis grading, etc.) and their predictions were based on their own diagnostic opinion.

Statistical analysis

Determination of intraobserver reliability for the three rounds of predictions for each surgeon was completed using a Fleiss Kappa model. Interobserver reliability between the five surgeons was determined using a Cohen Kappa model. To evaluate impact of surgeon sub-specialization on treatment prediction, comparison of the two surgeons practicing



Fig. 1. Example of image provided to surgeons to predict if a patient would receive a recommendation for THA. (A) Tönnis 0, (B) Tönnis 1, (C) Tönnis 2 and (D) Tönnis 3.

arthroscopic hip preservation versus the two surgeons performing hip arthroplasty were evaluated with Pearson's chi-squared and Fisher's exact tests.

RESULTS

Descriptive statistics

After inclusion criteria were implemented, 53 patients remained in the final study population. Twenty-six (49%) of the patients underwent a THA. The remaining 27 (51%) patients underwent an arthroscopic hip preservation procedure. Twelve patients (22.6%) exhibited Grade 0 Tönnis disease, whereas 18 patients (34.0%) demonstrated Grade 1 pathology. Twenty-one patients (39.6%) exhibited Grade 2 disease, and the remaining two patients (3.8%) demonstrated Grade 3 (Table I).

Reliability comparisons

The five surgeons evaluated the images three times and the intraobserver reliability was calculated for each surgeon with a Fleiss Kappa model. The kappa values for surgeons performing arthroscopic hip preservation were 0.572 and 0.659. The kappa values for surgeons performing hip arthroplasty were 0.668 and 0.270. Finally, the surgeon performing both procedures exhibited a kappa value of 0.690. Interobserver reliability between the five surgeons, utilizing a Cohen Kappa model, demonstrated a kappa value of 0.425 (Table II).

Intervention prediction based on specialty

Surgeons specializing in arthroscopic hip preservation predicted THA intervention 39.6% of the time, whereas

Table I. Distribution of Tönnis grade for patients included in the study

Tönnis grade	0	1	2	3
Number of patients	12	18	21	2
Percentage of patients	22.6	34.0	39.6	3.8

arthroplasty surgeons predicted THA intervention 32.7% (Figure 2). This reached significance with a Fisher's exact test ($P = 0.041$) and trended toward, but did not reach, a statistically significant level with a chi-square, Pearson's test ($P = 0.069$).

DISCUSSION

To our knowledge, this is the first study to evaluate the potential effect surgeon subspecialty has on treatment recommendation for hip preservation versus arthroplasty. Our study showed a significant difference for prediction of treatment recommendations by surgeons of varying practice demographics and procedural focus based on limited imaging. Arthroscopic hip preservation surgeons were more likely to predict THA as the recommended treatment than arthroplasty surgeons. This may be attributed to the view of the surgeon through the lens of their specialty. Arthroscopic hip preservation surgeons may be evaluating the image with the opinion that the patient has progressed beyond possibility of a successful arthroscopic procedure and would recommend an arthroplasty. Contrarily, an arthroplasty surgeon may evaluate the image with the

Table II. Kappa evaluations for intraobserver and interobserver reliability

Surgeon	1	2	3	4	5
Measure of agreement to patient's treatment selection					
First measure	0.622	0.366	0.659	0.439	0.624
Second measure	0.623	0.366	0.399	0.514	0.549
Third measure	0.548	0.182	0.434	0.55	0.548
Intraobserver reliability					
Kappa	0.572	0.668	0.270	0.659	0.690
Interobserver reliability					
Kappa	0.425				

Surgeons 1 and 4 specialize in arthroscopy, surgeons 2 and 3 specialize in arthroplasty while surgeon 5 performs arthroplasty and arthroscopy.

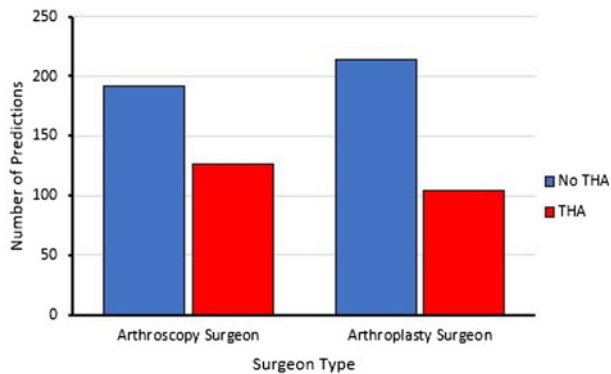


Fig. 2. Summary of THA versus no THA recommendations by specialty of surgeon. Tabulation of total recommendations for the 53 patients after three rounds of predictions.

opinion that degenerative changes have not progressed to a point requiring arthroplasty.

Previous studies outside the fields of arthroplasty and hip preservation have identified differences in treatment recommendations by varying subspecialties. Childs *et al.* [17] demonstrated hand fellowship-trained surgeons operatively treated multi-fragment intra-articular distal radius fractures more frequently than nonhand fellowship-trained surgeons. In addition, the hand fellowship-trained surgeons were also more likely to perform additional procedures at index treatment than their counterparts. Another study identified significant differences in arthroplasty versus open reduction internal fixation recommendations for proximal humerus fractures across multiple orthopedic subspecialties [18]. Our study shows this effect is present within the fields of arthroplasty and hip preservation as well.

The surgeon in our study performing both arthroplasty and arthroscopy surgeries exhibited the highest

intraobserver reliability. This could be attributed to their specific practice where they regularly provide recommendations for both interventions, rather than more specialized surgeons viewing the images under the influence of the specific intervention they provide. Furthermore, the low-reliability findings in our study support the importance of the full clinical picture of the patient for proper treatment recommendations and highlight imaging as a tool for diagnosis rather than a definitive investigation.

The correct recommendations of treatment have considerable impact on the patient's eventual outcome. While at times an interval arthroscopy could be considered as a temporizing or bridge procedure, studies have shown arthroscopy in the wrong patient can have a detrimental impact to the eventual arthroplasty outcome [19]. Perets *et al.* [20] found inferior subjective and objective outcomes for 35 matched, controlled patients undergoing arthroplasty after an arthroscopic procedure. While arthroscopy can lead to inferior outcomes for patients ultimately requiring THA, the potential for arthroscopy to progress the timeline to THA should also be considered.

A systematic review in 2015 found patients of increasing age or with pre-existing arthritis demonstrated an increased likelihood for rapid progression of disease [21]. In a prospective study, Gicquel *et al.* [22] showed that patients with Grade 1 Tönnis changes exhibited lower WOMAC scores, a higher rate of OA progression (54% versus 24%) and a higher rate for conversion to THA (33.3% versus 2.9%) than those with Grade 0 Tönnis changes. Previous studies have also shown that <2 mm of joint space along the sourcil as well as Grades 2 and 3 Tönnis changes are associated with decreased hip preservation surgery outcomes, high conversion rate to THA and decreased THA functional

outcomes [4–6]. Ultimately, these findings in the literature present well-supported evidence for predictive outcomes of patients with minimal or advanced arthritic disease. They, however, bring to light the complex patient that presents with Grade 1 Tönnis changes who may receive varying recommendations from surgeons. Subspecialty bias could play a role in the care pathway these patients experience as the ability to identify these early changes not only has an impact on their outcomes from a hip preservation surgery but also their progression of arthritic disease.

Secondly, interobserver reliability was found to be lower than the average intraobserver reliability suggesting there is a difference in the subjective interpretation of imaging amongst surgeons. Our findings of relatively low intra- and interobserver reliability resemble a prior study evaluating subjective evaluation of the hip, demonstrating low intraobserver reliability for grading OA ($\kappa = 0.57$) [14]. Interestingly, the same study demonstrated κ levels below 0.55 for interobserver reliability for objective measurements as well. Another study by Clohisey *et al.* [23] demonstrated low interobserver reliability in diagnosis of hip disease amongst six surgeons with a κ of 0.54. Their investigation included expanded imaging compared with our study by utilizing an AP pelvis, cross-table lateral, frog lateral, as well as a false profile view. In addition to subjective diagnosis, proposed objective structural measurements were made by each observer. Even with comprehensive imaging available, only 3 of the 15 structural measurements exhibited an interobserver κ greater than 0.5 (acetabular inclination, position of head center and Tönnis grade). Furthermore, this study was performed by surgeons within the field of hip preservation. The variability highlighted in former studies and reproduced in ours can have significant impact on an individual patient course to end treatment, as differing subjective interpretation of imaging can prompt variable recommendations for treatment and further studies.

Variations in recommendations can have cost and experience impacts for the patient. In the setting of early degenerative changes, especially in younger patients, advanced imaging is often considered to evaluate for intra- and extra-articular soft tissue pathology as well as degenerative changes. This can increase cost to the patient and health-care system as a whole. Previous studies show there are potential barriers to obtaining imaging which may limit their availability to patients which can further complicate a patient's health-care experience [4]. In addition, a complicated patient with early degenerative disease (i.e. Tönnis Grade 1) can receive conflicting recommendations from multiple surgeons. This can increase the cost to the patient, and as described above, has an impact on their eventual

outcomes based on which treatment route they decide to follow. The low interobserver reliability in our study identifies the potential for this phenomenon. Furthermore, the differences in recommendations from hip preservation and arthroplasty surgeons can have varying cost implications. As we continue to become more judicious in our attention to health care spending, further investigation into the impact of cost by surgeon recommendation based on subspecialty is warranted.

Limitations

There are several limitations to our study. Our findings are not tied to subjective, objective or clinical outcome measures; however, the purpose of the study is tied to predictions of treatment rather than the outcomes of interventions. Secondly, only two surgeons were included in the arthroplasty or arthroscopy groups and only one surgeon performing both interventions. Increasing the number of surgeons involved could provide further insight into variability amongst surgeons; however, we feel the inclusion of our high-volume surgeons provides a sufficient basis for our study. Our results were also consistent with those previously described in regards to imaging interpretation variability [16]. There was no control group in this study, and it may have been beneficial to have included normal hip radiographs in patients without pathology in the blinded X-rays. A previous study demonstrated 64% of their asymptomatic patients with normal radiographs were diagnosed with hip pathology by the blinded evaluating surgeons [23]. We feel we addressed this concern by including Tönnis Grade 0 patients and blinding providers to clinical information other than the patients presented for a hip complaint and failed conservative management.

In addition, our study does not include clinical information about the patients and we recognize surgeons do not make treatment recommendations on imaging alone. Decision for operative intervention includes a detailed history, clinical exam and adequate imaging in combination with a comprehensive discussion with the patient about their goals and expectations. We feel the limited information provided to the surgeons is a strength of our study as we aimed to evaluate the subjective interpretation of imaging alone. Inclusion of clinical information would have added bias to the study design and limited our ability to evaluate the variability of recommendations based on imaging interpretation. The imaging provided to the surgeons was limited, and we recognize this does not represent a typical imaging study a surgeon has in their office. We chose this region consistent with previous investigations by Philippon as the key landmark for predictive success of hip preservation surgery [4–6]. As mentioned previously,

Clohisy *et al.* [23] also found low interobserver reliability with an expanded imaging profile, thus our findings are more likely tied to actual variations between surgeons rather than a result of the limited imaging. Finally, our outcome prediction of THA versus no THA does not fully represent the range of interventions a patient may decide for or against. Patients with severe radiographic disease may be largely asymptomatic while others with relatively benign radiographic findings may be at the opposite end of the symptomology spectrum. Intricacies into each patient's expectations, goals, clinical history and exam lead to the final recommendation and the combined decision-making between the patient and surgeon. Our study did not aim to evaluate this portion of the care pathway and rather was devised to evaluate surgeons' prediction for treatment recommendation based solely on the imaging provided.

Our study provides several interesting possibilities for future directions of research. The population of patients with Grade 1 Tönnis changes who do not get a THA recommendation presents significant difficulty to their treating surgeons, as previously mentioned. Future study into the recommendations these patients received, such as arthrograms or arthroscopic procedures, and their resulting outcomes would shed further light on the care pathway the patients experience. In addition, further investigation into surgeons performing arthroplasty and arthroscopic hip preservation is required, as intraobserver reliability could consistently be higher in this population as they regularly provide both recommendations to patients. Finally, repeating our study with increasing levels of clinical information provided to surgeons could elucidate the critical levels of information necessary to show increased agreement in treatment recommendations both within and between surgeons.

FUNDING

Funds for publication were provided by the Detroit Medical Center Orthopedic Surgery Sports Medicine Fellowship.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Marin-Pena O, Tey-Pons M, Perez-Carro L *et al.* The current situation in hip arthroscopy. *EFORT Open Rev* 2017; **2**: 58–65.
2. Sing DC, Feeley BT, Tay B *et al.* Age-related trends in hip arthroscopy: a large cross-sectional analysis. *Arthroscopy* 2015; **31**(12): 2307–13.e2.
3. Nepple JJ, Carlisle JC, Nunley RM *et al.* Clinical and radiographic predictors of intra-articular hip disease in arthroscopy. *Am J Sports Med* 2011; **39**: 296–303.
4. Philippon MJ, Briggs KK, Carlisle JC *et al.* Joint space predicts THA after hip arthroscopy in patients 50 years and older. *Clin Orthop Relat Res* 2013; **471**: 2492–6.
5. Philippon MJ, Briggs KK, Yen YM *et al.* Outcomes following hip arthroscopy for femoroacetabular impingement with associated chondrolabral dysfunction: minimum two-year follow-up. *J Bone Joint Surg Br* 2009; **91**: 16–23.
6. Philippon MJ, Schroder E, Briggs KK. Hip arthroscopy for femoroacetabular impingement in patients aged 50 years or older. *Arthroscopy* 2012; **28**: 59–65.
7. Sansone M, Ahlden M, Jonasson P *et al.* Outcome of hip arthroscopy in patients with mild to moderate osteoarthritis-A prospective study. *J Hip Preserv Surg* 2016; **3**: 61–7.
8. Kester BS, Capogna B, Mahure SA *et al.* Independent risk factors for revision surgery or conversion to total hip arthroplasty after hip arthroscopy: a review of a large statewide database from 2011 to 2012. *Arthroscopy* 2018; **34**: 464–70.
9. Redmond JM, Gupta A, Dunne K *et al.* What factors predict conversion to THA after arthroscopy? *Clin Orthop Relat Res* 2017; **475**: 2538–45.
10. Davies O, Grammatopoulos G, Pollard TCB *et al.* Factors increasing risk of failure following hip arthroscopy: a case control study. *J Hip Preserv Surg* 2018; **5**: 240–6.
11. Brückl R, Hepp W, Tönnis D. [Differentiation of normal and dysplastic juvenile hip joints by means of the summarized hip factor]. *Arch Orthop Unfallchir* 1972; **74**: 13–32. [In German.]
12. Busse J, Gasteiger W, Tönnis D. [A new method for roentgenologic evaluation of the hip joint]. *Arch Orthop Unfallchir* 1972; **72**: 1–9. [In German.]
13. Radha S, Hutt J, Lall A *et al.* Best practice guidelines for arthroscopic intervention in femoroacetabular impingement syndrome: results from an International Delphi Consensus Project – Phase 1. *J Hip Preserv Surg* 2019; **6**: 1–13.
14. Mast NH, Impellizzeri F, Keller S *et al.* Reliability and agreement of measures used in radiographic evaluation of the adult hip. *Clin Orthop Relat Res* 2011; **469**: 188–99.
15. Nepple JJ, Martell JM, Kim Y-J *et al.*; for the ANCHOR Study Group. Interobserver and intraobserver reliability of the radiographic analysis of femoroacetabular impingement and dysplasia using computer-assisted measurements. *Am J Sports Med* 2014; **42**: 2393–401.
16. Carlisle JC, Zebala LP, Shia DS *et al.* Reliability of various observers in determining common radiographic parameters of adult hip structural anatomy. *Iowa Orthop J* 2011; **31**: 52–8.
17. Childs S, Mann T, Dahl J *et al.* Differences in the treatment of distal radius fractures by hand fellowship trained surgeons: a study of ABOS candidate data. *J Hand Surg Am* 2017; **42**: e91–7.
18. Gradl G, Knobe M, Pape HC *et al.* Decision making in displaced fractures of the proximal humerus: fracture or surgeon based? *Int Orthop* 2015; **39**: 329–34.
19. Konopka JF, Buly RL, Kelly BT *et al.* The effect of prior hip arthroscopy on patient-reported outcomes after total hip arthroplasty: an institutional registry-based, matched cohort study. *J Arthroplasty* 2018; **33**: 1806–12.
20. Perets I, Mansor Y, Mu BH *et al.* Prior arthroscopy leads to inferior outcomes in total hip arthroplasty: a match-controlled study. *J Arthroplasty* 2017; **32**: 3665–8.

21. Kemp JL, MacDonald D, Collins NJ *et al.* Hip arthroscopy in the setting of hip osteoarthritis: systematic review of outcomes and progression to hip arthroplasty. *Clin Orthop Relat Res* 2015; **473**: 1055–73.
22. Gicquel T, Gedouin JE, Krantz N *et al.* Function and osteoarthritis progression after arthroscopic treatment of femoro-acetabular impingement: a prospective study after a mean follow-up of 4.6 (4.2–5.5) years. *Orthop Traumatol Surg Res* 2014; **100**: 651–6.
23. Clohisy JC, Carlisle JC, Trousdale R *et al.* Radiographic evaluation of the hip has limited reliability. *Clin Orthop Relat Res* 2009; **467**: 666–76.