Inter-rater and Intrarater Reliabilities of the Identification of a "Gothic Arch" in the Acetabulum of Children With Cerebral Palsy

Stacey Miller, BSc, MRSc,*† Eva Habib, BSc,† Jeffrey Bone, MSc,‡ Emily Schaeffer, PhD,†§
Brian W. Yang, MD,‖ Jodie Shea, BS,¶ Ava Maleki, BSc,†
Benjamin J. Shore, MD, MPH, FRCSC,#
and Kishore Mulpuri, MBBS, MS(Ortho), MHSC(Epi), FRCSC†§

Background: Progressive hip displacement in children with cerebral palsy (CP) is monitored by measuring migration percentage (MP) on anteroposterior (AP) pelvis radiographs. Accurate measurement of MP requires the lateral margin of the ossified acetabulum to be identified for the placement of Perkin's line. It has been suggested that when there is an erosion of the acetabular rim, described as a gothic arch, the midpoint of the arch be used for the placement of Perkin's line. However, this requires that there be agreement on what constitutes a gothic arch. The purpose of this study was to evaluate the inter-rater and intrarater reliabilities of identifying a gothic arch on pelvis radiographs.

Methods: An online survey with 100 AP pelvis images (200 hips) of children with CP was sent to international experts. Participants were asked to identify which hip(s) had a gothic arch (left, right, both, and neither). The Fleiss κ statistic for inter-rater reliability was calculated. Eight weeks later, the images were shuffled and redistributed to calculate intrarater reliability.

Results: The initial survey was completed by 10 participants with 9 participants completing the second survey. The average inter-rater κ value was 0.18 [95% confidence interval (CI), 0.14-0.23] and 0.19 (95% CI, 0.14-0.24) for the 2 surveys, respectively. Among the pediatric orthopaedic surgeons subgroup, the κ values were 0.06 (95% CI, 0.02-0.1) and 0.08 (95% CI, 0.03-0.13). The average intrarater reliability κ value was 0.61 (95% CI, 0.2-1), ranging from 0.32 to 0.86.

Conclusions: There were poor inter-rater and moderate intrarater reliabilities in identifying a gothic arch on AP pelvis radiographs in children with CP. Further characterization and clarification of what constitutes a gothic arch are required. The lack of agreement on the identification of a gothic arch may negatively impact the measurement of MP and referrals to a pediatric orthopaedic surgeon.

Key Words: cerebral palsy, hip migration, inter-rater reliability, intrarater reliability, Reimer's migration percentage

From the Departments of *Physical Therapy; ‡Obstetrics and Gynaecology; §Orthopaedic Surgery, University of British Columbia; †Department of Orthopaedics, BC Children's Hospital, Vancouver, BC, Canada; ‖Hospital for Special Surgery, New York, NY; ¶Boston Children's Hospital; and #Department of Orthopaedic Surgery, Boston Children's Hospital, Harvard Medical School, Boston, MA.

One in 3 children with cerebral palsy (CP) will have hip displacement.[1–3] Left untreated, the hip can dislocate causing pain and decreased quality of life.[4–6] Evidence supports the use of systematic hip surveillance to identify children with hip displacement and allow for timely orthopaedic management.[7,8] Guidelines for hip surveillance, which provide recommendations for frequency of surveillance and define the criteria for referral to a pediatric orthopaedic surgeon, are readily available.[7,9–11] As hip displacement cannot be detected on clinical examination of the hip, supine anteroposterior (AP) pelvis radiographs, using standardized positioning, are recommended.

Migration percentage (MP) is accepted as the most valid and reliable method of measuring hip displacement from an AP pelvis radiograph.[12] MP is defined as the percentage of the ossified femoral head lateral to Perkin's line (Fig. 1).[13] Perkin's line is drawn at the lateral margin of the ossified acetabulum, perpendicular to Hilgenreiner's line, a horizontal line drawn between the right and left triradiate cartilages. A referral to a pediatric orthopaedic surgeon is frequently recommended once the MP is > 30%, whereas surgical interventions are often recommended once the MP is > 40%.[9–11,14,15] Accurate and reliable measurement of MP is therefore important for informed decision making.
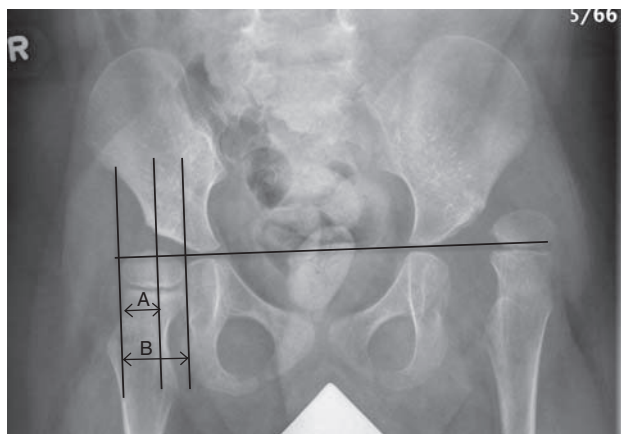
FIGURE 1. Measurement of Reimer's migration percentage (MP). MP = A/B × 100%.

Identification of the lateral margin of the acetabular rim is critical for the placement of Perkin's line and, thus, for accurate calculation of MP. In a typically developing hip, the radiologic roof of the acetabulum, the sourcil, extends horizontally to a clearly defined, everted lateral margin.[16] The term gothic arch has been used to describe the erosion of the superior articular margin of the acetabular rim.[17] It is suggested that eccentric pressure from a displaced femoral head results in inhibition of ossification of the superolateral aspect of the cartilaginous acetabulum that produces a shape similar to a gothic arch.[17,18] The true acetabular rim is described as being just below the tip of the arch.[17] The gothic arch phenomenon is otherwise poorly characterized in the literature.

When measuring MP in the presence of a gothic arch, it has been recommended that Perkin's line be placed at the midpoint of the arch as this may be a more valid representation of the coverage of the femoral head.[2,11,18] However, this requires that there be consensus on what constitutes a gothic arch. To the best of our knowledge, no studies have evaluated the inter- or intrarater reliability of identifying the presence of a gothic arch. The objective of this study was to evaluate the inter-rater and intrarater reliabilities of identifying a gothic arch on AP pelvis radiographs of children with CP.

METHODS

A retrospective chart review was conducted on patients aged 0 to 19 years with a diagnosis of CP at a Canadian and an American pediatric tertiary care center between January 2007 and December 2017. Children were identified through a database of children with CP treated in orthopaedic clinics at these centers. Authors from both centers selected AP pelvis radiographs they felt may have a gothic arch and images with normal acetabular development to act as controls. One author made the final selection of images from those identified at the 2 centers and did not complete the survey. Additional data collected included patient diagnosis, Gross Motor Function Classification System (GMFCS) level, age at radiograph, sex,

and the date of imaging. In total, 120 deidentified AP pelvis radiographs of children with CP were obtained. An online survey with 100 of these images (200 hips), from children across GMFCS levels II to V (II: 1; III: 9; IV: 34; V: 56), between the ages of 2.1 and 14.6 years (mean, 7.6 y) was developed. Of these, 65 images were identified as possibly having ≥ 1 hips with a gothic arch by the study authors. The study was reviewed and approved by the Research Ethics Board at the University of British Columbia (H17-02976).

An invitation to complete an online survey was sent to 19 participants, including members of the American Academy of Cerebral Palsy and Developmental Medicine (AACPDM) Hip Surveillance Care Pathway Committee and other experts known to the authors. Participants were asked to identify their clinical role and answer questions related to their practice and experience. They were then asked to identify which hip(s) had a gothic arch: left, right, both, or neither (survey 1). Issues with image resolution and size were reported for 5 radiographs. These radiographs were removed from the data analysis and the subsequent intrarater reliability survey. To assess intrarater reliability, the order of the remaining 95 images was shuffled and the survey was redistributed 8 weeks later to the same respondents (survey 2). An additional 2 images were removed from this survey on the basis of participant feedback. At the conclusion of the follow-up survey, participants were asked to describe what radiographic features constitute a gothic arch.

Results were analyzed to determine consistency between raters when assessing images for the presence of a gothic arch. For both surveys, the overall agreement for each image (both left and right hips) and the Fleiss[19] κ statistic for inter-rater reliability were calculated. The κ values were interpreted using the Landis and Koch[20] guidelines outlined as follows: values <0.00 indicate poor agreement, 0.00 to 0.20 indicate slight agreement, 0.21 to 0.40 indicate fair agreement, 0.41 to 0.60 indicate moderate agreement, 0.61 to 0.80 indicate substantial agreement, and 0.81 to 1.00 indicate almost perfect agreement. Given the difficulty in providing parametric standard errors for the Fleiss κ statistics, confidence intervals (CIs) were on the basis of 1000 bootstrap resamples. All analyses were repeated restricting the images rated to the first 30 to assess whether potential burnout affected the overall estimates. Intrarater reliability was assessed through Cohen κ for each rater and 95% CIs were calculated.[21] All analyses were done using R version 3.5.3.[22]

RESULTS

Participant Demographics

Ten participants completed the initial survey: 6 pediatric orthopaedic surgeons, 1 pediatric radiologist, and 3 physical therapists. The participants had between 8 and 35 years (median, 18.5 y) of experience reviewing radiographs of children with CP. Participants reported seeing an average of 328 patients annually, ranging from 100 to 800 children per year. When assessing the self-reported

TABLE 1. Inter-rater Reliability Results for Survey 1 and Survey 2 for all Images

	Survey 1 (10 Participants)		Survey 2 (9 Participants)	
	Fleiss κ	95% Bootstrap CI	Fleiss κ	95% Bootstrap CI
All participants, all images	0.18	0.14-0.23	0.19	0.14-0.24
All participants, first 30 images	0.16	0.09-0.23	0.11	0.04-0.17
Orthopaedic surgeons only, all images	0.06	0.02-0.1	0.08	0.03-0.13
Orthopaedic surgeons only, first 30 images	0.04	−0.02 to 0.10	−0.01	−0.08 to 0.06

CI indicates confidence interval.

level of comfort assessing radiographs in children with CP, 9 participants reported "very comfortable" and 1 participant (physical therapist) reported being "comfortable." Responses from 1 additional orthopaedic surgeon were eliminated after the participant reported to the authors that the survey was completed in reference to a gothic arch as described by Bombelli.[23] One physical therapist did not complete the follow-up survey (survey 2).

Interobserver Level of Agreement

The average inter-rater κ value was 0.18 (95% CI, 0.14-0.23) across all participants for the initial survey and 0.19 (95% CI, 0.14-0.24) for the follow-up survey, indicating slight agreement (Table 1). Among only the pediatric orthopaedic surgeons, the κ values were 0.06 (95% CI, 0.02-0.1) and 0.08 (95% CI, 0.03-0.13), respectively, also indicating slight agreement. To assess a potential burnout effect, κ values were calculated for the first 30 images in each survey. Agreement remained only slight with κ values of 0.16 (95% CI, 0.09-0.23) and 0.11 (95% CI, 0.04-0.17). In reviewing individual images, there was >80% agreement on both surveys 1 and 2 that a gothic arch was present in 3 hips. These 3 hips were from children at GMFCS level IV at a mean age of 6.64 years (SD, 4.38). There was >80% agreement on both surveys 1 and 2 that a gothic arch was absent in 33 hips. These radiographs were from children at GMFCS levels III (3), IV (10), and V (20) at a mean age of 6.99 years (SD, 3.91).

Intraobserver Level of Agreement

The average intrarater reliability κ value was 0.61 (95% CI, 0.2-1) indicating moderate agreement, ranging from 0.32 (fair) to 0.86 (almost perfect; Table 2). Three

TABLE 2. Intrarater Reliability for Each Rater (and Overall) Between Survey 1 and Survey 2

Rater	κ	95% CI
1	0.56	0.44-0.68
2	0.5	0.35-0.65
3	0.52	0.37-0.66
4	0.83	0.75-0.91
5	0.32	0.16-0.47
6	0.55	0.43-0.67
7	0.42	0.32-0.52
8	0.82	0.74-0.91
9	0.86	0.78-0.94
Overall	0.61	0.2-1

CI indicates confidence interval.

participants (2 surgeons and a physical therapist) had κ values over 0.81, or almost perfect agreement. Inter-reliability analysis of these 3 raters showed substantial agreement (Table 3).

DISCUSSION

Inter-rater reliability of the identification of a gothic arch in hip radiographs of children with CP was found to be poor even among international experts. Overall, intrarater reliability was moderate. These findings suggest that further characterization and clarification on what constitutes a gothic arch is required.

Before the survey distribution, no instructions were provided regarding what constitutes a gothic arch. In previous studies of MP reliability that reported using the midpoint of the gothic arch, it was noted that the raters developed consensus on what constitutes a gothic arch before measuring.[18,24] However, there are no established guidelines available in the literature that distinguish when acetabular changes are clearly a gothic arch, thus no guidelines were provided to participants. Establishing consensus regarding the definition of what constitutes a gothic arch may improve our findings.

Inter-rater agreement was >80% that a gothic arch was present for 8 hips in survey 1 and 8 hips in survey 2. When results were combined, there were 3 hips in both surveys where the agreement was >80% for the presence of a gothic arch. These radiographs (Fig. 2) were in children between the ages of 3.1 and 11.6 years, indicating that a gothic arch may be present from a young age. Not surprising, there was greater agreement among participants on the absence of a gothic arch.

Although overall intrarater reliability was moderate, 3 participants (2 orthopaedic surgeons and 1 physical therapist) had excellent intrarater reliability. Their inter-rater reliability had a moderate to substantial agreement, well above that of the entire group. Further review of their results may offer some insight into common characteristics of a gothic arch. When asked what radiographic features constitute a gothic arch, these 3 respondents described "a double line" or "an indent" on the

TABLE 3. Subgroup Inter-rater Reliability Analysis of 3 Raters With Almost Perfect Intrarater Agreement

	Fleiss κ	95% Bootstrap CI
Survey 1	0.64	0.55-0.73
Survey 2	0.60	0.5-0.67

CI indicates confidence interval.

FIGURE 2. Radiographs of 3 hips with >80% agreement by all participants that a gothic arch is present on both surveys.

lateral aspect of the acetabulum or an "accentuated and medialized" acetabular roof.

The clinical implications of this study remain unclear. The impact of poor inter-rater reliability of identifying a gothic arch on the measurement of MP was not evaluated in this study and, therefore, our findings should not be interpreted as a lack of reliability of measuring MP. Previous reports have described the inter-rater and intrarater reliabilities of measuring MP in children with CP as good to excellent.[12] Most recently, Shore et al[25] reported high reliability in measuring MP among 15 international experts. However, they noted that MP exceeded a SD of 10% in 8 of 50 hips and the only factor that was associated with this variability was the presence of a gothic arch (6/8 hips). Participants in this study were instructed to use the midpoint of the gothic arch. These findings suggest that the identification of a gothic arch and subsequent placement of Perkin's line can be challenging even for experts and can impact the measurement of MP. Further study on the impact of poor inter-reliability in identifying a gothic arch on the reliability of MP is required. The AACPDM Hip Surveillance Care Pathway recommends that a child be referred to a pediatric orthopaedic surgeon when a gothic arch is observed on imaging.[11] The authors of the pathway noted that the observation of a gothic arch on a radiograph indicates significant acetabular dysplasia is present, thus hip migration is likely significant. The inability to accurately identify a gothic arch may impact referrals. Overestimating what constitutes a gothic arch may result in children being unnecessarily referred to a pediatric orthopaedic surgeon. Screening programs aim to identify all cases and false-positives can be expected. However, given the degree of uncertainty found here, the presence of a gothic arch

cannot be used to initiate referrals. Assuming the midpoint of the gothic arch should be used to measure MP, failure to identify a gothic arch may result in a low MP. If MP is under-reported, then there may be a delay in referral to an orthopaedic surgeon. Our results suggest that individual clinicians cannot be expected to accurately identify a gothic arch as part of screening for hip surveillance and that perhaps the AACPDM guidelines should be modified accordingly.

This study had limitations. After the original survey, 5 images had to be eliminated from analysis because of difficulties viewing and an additional 2 images were eliminated from the follow-up survey. There was no ability to study participants to manipulate the image size or contrast. This may have impacted the ability of raters to make a clinical decision as these features are typically available in clinical practice settings. It was also noted that decisions about a gothic arch may not be made on the basis of a single radiograph and that, if there was doubt, repeat imaging may be completed. For this reason, 1 participant wished to state "unable to determine" when evaluating the images. The importance of patient positioning should not be underestimated. In Figure 3, 2 images that are taken 6 months apart illustrate how the position of the pelvis can result in the lateral edge of the acetabulum being well defined or suggestive of a gothic arch. Further characterization regarding how AP pelvic tilt can affect the appearance of the gothic arch needs to be explored. There may have been selection bias when choosing the images. Although images were selected by 2 authors at different institutions, they may have been a poor representation of what is felt to represent a gothic arch by the study participants. Finally, by not asking
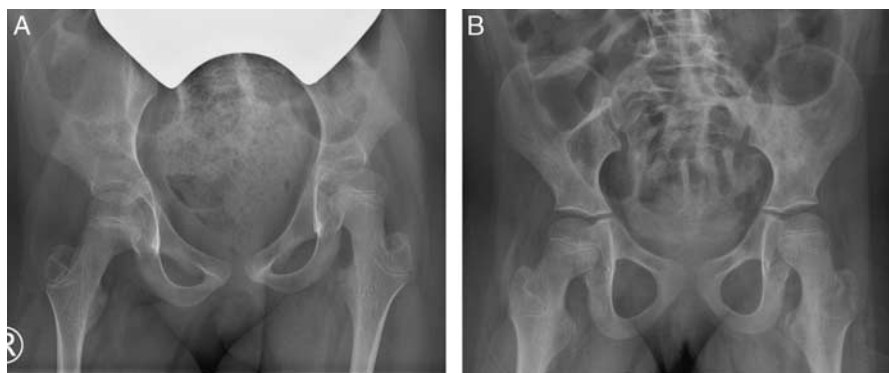
FIGURE 3. Anteroposterior pelvis radiographs of the same child taken 6 months apart suggesting the presence of a gothic arch on the left hip in image (A) that disappears in image (B).

participants to measure MP, the clinical significance of poor inter-reliability of identifying a gothic arch remains unclear.

Acetabular morphology varies widely in children with CP. This study indicates that experts in the care of hip displacement in children with CP do not agree on what constitutes a gothic arch. Future work should include consensus building on the defining criteria of a gothic arch and determining whether the child's age, GMFCS level, acetabular index, or MP are factors that influence or predict the presence of a gothic arch. The use of computed tomography scans may be a helpful tool in future efforts to describe a gothic arch and to validate the optimal placement of Perkin's line.

REFERENCES

1. Soo B, Howard JJ, Boyd RN, et al. Hip displacement in cerebral palsy. *J Bone Joint Surg Am.* 2006;88:121–129.
2. Hagglund G, Lauge-Pedersen H, Wagner P. Characteristics of children with hip displacement in cerebral palsy. *BMC Musculoskelet Disord.* 2007;8:101.
3. Connelly A, Flett P, Graham HK, et al. Hip surveillance in Tasmanian children with cerebral palsy. *J Paediatr Child Health.* 2009;45:437–443.
4. Ramstad K, Terjesen J. Hip pain is more frequent in severe hip displacement: a population-based study of 77 children with cerebral palsy. *J Pediatr Orthop B.* 2016;25:217–221.
5. Jung NH, Pereira B, Nehring I, et al. Does hip displacement influence health-related quality of life in children with cerebral palsy? *Dev Neurorehabil.* 2014;17:420–425.
6. Ramstad K, Jahnsen RB, Terjesen T. Severe hip displacement reduces health-related quality of life in children with cerebral palsy. *Acta Orthopaedica.* 2017;88:205–210.
7. Hagglund G, Alriksson-Schmidt A, Lauge-Pedersen H, et al. Prevention of dislocation of the hip in children with cerebral palsy; 20 year results of a population-based prevention programme. *Bone Joint J.* 2014;96-B:1546–1552.
8. Kentish M, Wynter M, Snape N, et al. Five year outcome of state-wide hip surveillance of children and adolescents with cerebral palsy. *J Pediatr Rehabil Med.* 2011;4:205–217.
9. Miller SD, Mayson TA, Mulpuri K, et al. Developing a province-wide hip surveillance program for children with cerebral palsy: from evidence to consensus to program implementation: a mini-review. *J Pediatr Orthop B.* 2019. Doi: 10.1097/BPB.0000000000000707.
10. Wynter M, Gibson N, Willoughby KL, et al. Australian hip surveillance guidelines for children with cerebral palsy: 5-year review. *Dev Med Child Neurol.* 2015;57:808–820.
11. O'Donnell M, Mayson T, Miller S, et al. Hip surveillance in cerebral palsy care pathway [American Academy of Cerebral Palsy and Developmental Medicine Web site]. September 2017. Available at: https://www.aacpdm.org/publications/care-pathways/hip-surveillance. Accessed November 13, 2019.
12. Pons C, Remy-Neris O, Medee B, et al. Validity and reliability of radiological methods to assess proximal hip geometry in children with cerebral palsy: a systematic review. *Dev Med Child Neurol.* 2013;5:1089–1102.
13. Reimers J. The stability of the hip in children: a radiological study of results of muscle surgery in cerebral palsy. *Acta Orthop Scand.* 1980;184:1–100.
14. Hagglund G, Lauge-Pedersen H, Persson M. Radiographic threshold values for hip screening in cerebral palsy. *J Child Orthop.* 2007;1:43–47.
15. Dobson F, Boyd RN, Parrott J, et al. Hip surveillance in children with cerebral palsy. Impact on the surgical management of spastic disease. *J Bone Joint Surg [Br].* 2002;84-B:720–726.
16. Robin J, Graham HK, Selber P, et al. Proximal femoral geometry in cerebral palsy: a population-based cross-sectional study. *J Bone Joint Surg Br.* 2008;90:1372–1379.
17. Roach JW, Hobatho MC, Baker KJ, et al. Three-dimensional computer analysis of complex acetabular insufficiency. *J Ped Orthop.* 1997;17:158–164.
18. Parrott J, Boyd RN, Dobson F, et al. Hip displacement in spastic cerebral palsy: repeatability of radiologic measurement. *J Pediatr Orthop.* 2002;22:660–667.
19. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull.* 1971;76:378–382.
20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.
21. Cohen J. A coefficient of agreement for nominal scales. *Educ and Psychol Meas.* 1960;20:37–46.
22. R Core Team. R: a language and environment for statistical computing [R Foundation for Statistical Computing website]. 2019. Available at: https://www.R-project.org. Accessed November 13, 2019.
23. Bombelli R. The biomechanics of the normal and dysplastic hip. *Chir Organi Mov.* 1997;117–127.
24. Kulkarni VA, Davids JR, Boyles AD, et al. Reliability and efficiency of three methods of calculating migration percentage on radiographs for hip surveillance in children with cerebral palsy. *J Child Orthop.* 2018;12:145–151.
25. Shore BJ, Martinkevich P, Riaz M, et al. Reliability of radiographic assessments of the hip in cerebral palsy. *J Pediatric Orthop.* 2019;39:e536–e541.