SHORT COMMUNICATION

To what degree is digital imaging reliable? Validation of femoral neck shaft angle measurement in the era of picture archiving and communication systems

J D WILSON, MRCSEd, W EARDLEY, MRCSEd, S ODAK, MRCS and A JENNINGS, FRCS (TR & ORTH)

Trauma and Orthopaedics, University Hospital of North Durham, County Durham and Darlington Foundation Trust UK

Objectives: This study investigates the reliability of femoral neck shaft angle (NSA) measurements made with the software and images available in routine clinical practice.

Methods: Using the Centricity Enterprise™ (GE Healthcare Pty Ltd Piscataway, NJ) picture archiving and communication system (PACS), the NSA of the proximal femur was measured from anteroposterior radiographs of adult hips. 3 independent observers, using a standardised technique, performed a total of 120 measurements.

Results: The Pearson's correlation coefficient for the intraobserver agreement was 0.98 ($p < 0.01$) and for interobserver measurements 0.86 ($p < 0.01$). Bland–Altman plots revealed the limits of intraobserver agreement to be $\pm 2.5^\circ$, but interobserver limits of agreement to be $\pm 6^\circ$. The intraclass correlation coefficient (ICC) was also calculated. The interobserver ICC was 0.62 (0.42–0.78, 95% confidence interval (CI); $p < 0.001$). The intraobserver ICC was 0.98 (0.95–0.99, 95% CI; $p < 0.001$).

Conclusion: PACS software has many advantages, but when using systems that can display angle measurements to one-tenth of a degree caution must be exercised to ensure that reliability of these measurements is not overestimated. We found that in the context of measuring the NSA of the proximal femur the reliability of the measurement, even under the best conditions, is only $\pm 6^\circ$ for different observers.

Received 14 June 2009
Revised 15 March 2010
Accepted 25 March 2010

DOI: 10.1259/bjr/29690721

© 2011 The British Institute of
Radiology

Measurements for pre-operative templating have traditionally been performed on analogue plain radiographs [1]. In recent years, picture archiving and communication systems (PACS), such as Centricity Enterprise™ Web version 1.0 (GE Healthcare Pty Ltd Piscataway, NJ) have become widespread and are the standard method for viewing radiological and other images in many centres [2]. The software allows image manipulation, including magnification, as well as linear and angle measurement [2].

Awareness of proximal femoral geometry is important in the pre-operative planning of osteotomy, arthroplasty or fracture fixation [3]. Pre-operative planning forces the surgeon to think in three-dimensions and has been shown to reduce operation time and the frequency of complications [4–6]. In addition, the femoral neck shaft angle (NSA) has implications for femoral neck fracture risk [7].

The introduction of PACS has given clinicians the opportunity to use software tools capable of measuring angles to one-tenth of a degree and distances to one-hundredth of a millimetre. Can these very precise tools be used reliably in clinical practice? This study aims to

evaluate the reliability of measurements of the femoral NSA performed using PACS software.

Methods and materials

A power calculation was performed using SamplePower 2™ (SPSS Inc. Chicago, IL). This showed that for a sample size of 30 per group the power would be in excess of 90% for a paired two-tailed *t*-test and for a one-way ANOVA with alpha set at 0.05. A standard deviation (SD) of 5.7° was used based on a pilot study measuring the NSA in the same population using an identical technique. This calculation was performed with the minimal clinically important difference set at 5° .

The measurement of the proximal femoral NSA was performed retrospectively on 15 anteroposterior (AP) pelvis radiographs, including both hips. The sample was identified using a quasi-random selection to generate a number of films used in everyday practice. This process involved searching the centricity database for all pelvis radiographs on an arbitrary date. Consecutive films from this date were included in the study if the image was adequate to perform the measurements on.

The sample was taken from the hospital database that serves a mixture of urban and rural populations in north-east England, with a catchment population of approximately 250 000.

Address correspondence to: J D Wilson, Care of North West Deanery, 3 Piccadilly Place, Manchester, M1 3BN. E-mail: jameswilson@doctors.org.uk

These non-standardised films were used to reflect the images that would be available in actual clinical practice. Radiographs from both genders and a broad range of ages were included. The films were judged adequate to perform measurements on if both proximal femora were included, no implants were present and no deforming hip pathology was visible. Films demonstrating subtle changes of osteoarthritis were allowed if the contours of the femoral head and neck could still be clearly visualised.

The study was performed on standard workstations available in clinical areas with a minimum screen size of 48 cm (19 inches). To optimise the measuring process the "full-screen" view was used and the image was magnified by a factor of two to maximise resolution [8].

The projected NSA was defined as the angle subtended between the femoral neck axis and the femoral shaft axis on the AP radiograph (Figure 1). The mid-point of the femoral neck at its narrowest point and the centre of the femoral head were identified. No circle overlay tool was available with the Centricity™ software. Therefore, the centre of the femoral head was found by halving its maximum diameter using the linear measuring feature. The neck axis was then generated by extending a line through these points. The femoral shaft axis was generated by identifying the mid-point of the diaphysis at two different points distal to the flare of the lesser and greater trochanters, again using the linear measuring feature. The intersection of these two axis was then determined using the angle-measuring feature.

Figure 1. Anteroposterior radiograph of right proximal femur showing the technique for measuring the neck shaft angle.

3 different surgeons performed the measurements on each hip of the 15 radiographs to compare interobserver error. Each surgeon had different levels of experience in orthopaedics, but all had a minimum of 2 years training. They were all familiar with the PACS software. The measuring protocol was explained, but no specialist training was offered. One of the surgeons performed the measurements twice to assess intraobserver error. This gave a total of 120 measurements. Statistical analysis was performed using SPSS 15 (SPSS Inc., Chicago, IL).

During the measuring process the observers were blinded to other measurements. For the repeated measurements performed by one observer, a 2 week period elapsed between the first and second set of measurements so it was unlikely that individual results could be remembered.

The data were subject to a series of statistical tests to assess the reliability of the measurements. The mean NSA values were compared. A paired samples *t*-test was used for the observer making two observations. A one-way ANOVA was used where more than two observations were being compared. Correlation coefficients were also calculated for inter- and intra-observer error and the levels of agreement were further investigated using a Bland-Altman plot [9]. The limits of agreement were calculated and 95% confidence intervals (CI) for the limits of agreement were identified [10]. The intraclass correlation coefficient (ICC) was calculated between and within observers [11].

Results

15 AP pelvic radiographs were used, giving a total of 30 hip measurements per observer. The mean age was 47 years (range, 15–79 years). Five men and ten women were included. This difference was tolerated as the study compares repeated measurements on the same image and therefore the subjects are perfectly matched.

The mean NSAs measured on two separate occasions by the same observer were 132.8° and 133.0°. A paired samples *t*-test confirms that there is no significant difference between these values ($p=0.27$).

The mean NSA measurements of the three separate observers are shown in Table 1. To judge the significance of the difference between the means of the observations, a one-way ANOVA was performed. This gave a *p*-value of 0.026, which confirmed that there was a significant difference between measurements made by different observers. This was then characterised in the post-hoc analysis (Table 2).

Figure 2 demonstrates the scatter of the intra-observer variability. The Pearson's correlation coefficient for the

Table 1. Mean neck shaft angle (NSA) measurements of the three observers with confidence intervals (CI)

Observer	Mean	95% CI for mean	
		Lower bound	Upper bound
1	133.0	130.9	135.2
2	134.0	131.8	136.2
3	130.3	128.6	132.0
Total	132.4	131.3	133.6

Table 2. Inter-observer differences. *Post hoc* analysis using the Bonferroni principle showed a significant difference between the mean neck shaft angle measurements of Observer 2 and Observer 3, but no significant difference between the pairings of Observer 1 and 2 or pairing 1 and 3

Pair a and b	Mean difference in NSA (a,b)	p-value
Observer 1 and 2	-1.0°	1.000
Observer 1 and 3	2.8°	0.159
Observer 2 and 3	3.8°	0.027

NSA, neck-shaft angle.

two measurements was 0.98 ($p < 0.01$). The correlation coefficient for the inter-observer measurements was 0.86 ($p < 0.01$). Both these Pearson's coefficients show high levels of correlation. However, it is expected that measurements of the same images will have a strong correlation, and a better representation of the agreement between the observers is shown in Figure 3 as advocated by the Bland-Altman plot [9].

The Bland-Altman plot shows that the measurements are distributed symmetrically about the mean indicating no systematic bias. The Shapiro-Wilk test was performed to test for normality of the distribution of the differences. This showed that there was no significant difference from a normal distribution ($p = 0.33$). The SD of the differences from the mean was 2.9°. The reliability of the measurement should lie within ± 2 SD [9]. Therefore, one would expect the difference between measurements from 2 different observers to lie within the limits of agreement 95% of the time. There is a 95% chance that a separate measurement performed by another observer will lie $\pm 6^\circ$ (5.8°) from the first measured value.

Following the formulae stated by Bland and Altman [9], the CIs for the estimate of the limits of agreement can be calculated [10]. This shows that there is a 95% chance

Figure 2. Scatter plot showing the correlation between repeated measurements of neck shaft angle (NSA) by the same observer.

Figure 3. Bland-Altman plot showing agreement between observers and the mean neck shaft angle (NSA). The horizontal reference lines represent the line of agreement and ± 2 standard deviations from this value.

that the limits of agreement lie between $\pm 4^\circ$ and 7.6° . Therefore, even the most optimistic interpretation would be that the limits of agreement were $\pm 4^\circ$ and may be as large as $\pm 7.6^\circ$.

The intra-observer agreement was assessed using the same method. The plot is shown in Figure 4. For repeated measurements by the same observer the limits of agreement lie within $\pm 2.5^\circ$ (95% CI = $\pm 1.7^\circ$ – 3.3°).

The ICC was calculated [12] (using a one-way, random effects ANOVA model [11]). For the interobserver calculation, the ICC was 0.62 (0.42–0.78, 95% CI) for single measures ($p < 0.001$). The intra-observer calculation revealed an ICC of 0.98 (0.95–0.99, 95% CI), which was also significant ($p < 0.001$).

Figure 4. Bland-Altman plot for intra-observer agreement of the neck shaft angle (NSA). The horizontal reference lines represent the line of agreement and ± 2 standard deviations from this value.

Discussion

This is the first paper to test and describe the reliability of this measuring technique using the digital technology made available by PACS. As these systems become more widespread in routine clinical practice, the accuracy, utility and reliability of the tools available with PACS software requires assessment.

Interobserver error was found to be much larger than the intra-observer error. Even with specific instructions to generate the necessary axes, individual observers still identified different points. This limits the reliability of repeated measurement regardless of how precise the software can measure an angle superimposed on an image.

In this study, the repeated measurements were performed on digital images using the same software and standardised computer interfaces. The cause for the interobserver error in this context was variation in cursor positioning during the measuring process. Small variations in positioning are magnified because the bony landmarks, which generate the two axes, are close together. These axes need to be extrapolated far from the landmarks to measure the NSA. Challenges to the reliable measurement of the femoral NSA have been identified by other authors. A study by Sanfridsson et al [8] showed that proximal femoral NSA measurements had the highest variance (SD 1.9°) of 9 angular measurements generated from radiographic views of the lower limb (mean SD of the other 8 angles was 0.6°).

A further cause for variation is the digital image when viewed on a monitor. The position of the crosshair is dictated by screen pixels and the measuring point can only occupy one pixel, it cannot be positioned between pixels. For example, finding the exact mid-point of the femoral shaft may lead to a choice of two cursor positions either side of exactly halfway. The operator has to make a decision at this stage about which pixel to choose. Therefore, even if two observers had identified exactly the same points across the shaft and neck, four mid-points would still need to be identified. Potentially, this could lead to 16 (2⁴) different combinations of cursor positions. Thus, even with a standardised technique, differences can arise between observers.

Measuring assistance tools (MATs) that locate the centre of a circular structure can reduce the variation in measurements of angles and distances. However, the effect of using these tools to measure the proximal femoral NSA has not been shown to be statistically significant [8]. Also, a study comparing digital with conventional techniques for assessing lower limb alignment found no significant difference between the mean values or variation of the measurements [13]. Other studies have investigated the reliability of digital measurements at other skeletal sites, such as in juvenile scoliosis, and found that levels of correlation between repeated angle measurements were low (correlation coefficient, 0.60) [14].

This study had a number of limitations. Using 15 radiographs of 30 hips means that the data were not entirely independent. There will be correlation between the hips for the same patient. This could lead to the variation being underestimated. Also, intra-observer reliability was only calculated for one observer. This estimate of the intrarater variability would be more generalisable if more observers had repeated the measurements.

This paper does not investigate the underlying reasons why observers chose different points to measure the NSA. Differences in perception of spherical and tubular structures may be important. Subtle variations in the technique employed may be accountable. For example, Observer 1 may have always found the mid-point of structures by measuring from the medial side to the centre, while Observer 2 used the lateral side and Observer 3 had an inconsistent approach. Details of perception and fine points of technique were not scrutinised, but further investigation may reveal underlying reasons for interobserver variation or that detailed training could improve reliability.

Despite these limitations, this paper demonstrates that even with image enhancement, magnification and software that can compute angles to a fraction of a degree, the level of agreement between observers is limited to $\pm 6^\circ$.

Previous studies of NSA measurement have employed methodologies that can produce results far more precisely than measurements on analogue films [15, 16]. However, these methods would not be routinely available or realistic in clinical practice. Anderson and Trinkaus [15] took direct measurements from human femora post-mortem and Husmann et al [16] used 3D CT reconstructions to derive their angles. These methodologies are far removed from the everyday assessment of images in clinical situations.

Conclusion

PACS carry many advantages that have led to their increasing prevalence within healthcare settings. However, clinicians presented with software tools that can measure angles to one-tenth of a degree must be aware that the reliability of measurements using these tools is limited. Caution must be exercised and the authors would recommend that in the context of measuring the NSA of the proximal femur, the limits of agreement for different observers, even under optimal conditions, is only $\pm 6^\circ$.

References

1. Knight JL, Atwater RD. Preoperative planning for total hip arthroplasty. Quantitating its utility and precision. *J Arthroplasty* 1992;7:403–9.
2. Johnson LJ, Cope MR, Shahrokhi S, Tamblyn P. Measuring tip-apex distance using a picture archiving and communication system (PACS). *Injury* 2008;39:786–90.
3. Kay RM, Jaki KA, Skaggs DL. The effect of femoral rotation on the projected femoral neck shaft angle. *J Pediatr Orthop* 2000;20:736–9.
4. Muller ME. Lessons of 30 years of total hip arthroplasty. *Clin Orthop Relate Res* 1992;274:12–21.
5. Haddad FS, Masri AB, Garbuz DS, Duncan CP. Classification and preoperative planning. *Instr Course Lect* 2000;49:83–96.
6. Egli S, Pisan M, Muller ME. The value of preoperative planning for total hip arthroplasty. *J Bone Joint Surg Br* 1998;80:382–90.
7. Gnudi S, Ripamonti C, Lisi L, Fini M, Giardino R, Giavaresi G. Proximal femur geometry to detect and distinguish femoral neck fractures in postmenopausal women. *Osteoporos Int* 2002;13:69–73.
8. Sandfridsson J, Svahn G, Ryd L, Ahl L, Sundén P, Jonsson K. Assessment of image post-processing and of measuring

- assistance tools in computed radiography. Evaluation of the weight bearing knee. *Acta Radiol* 1998;642–8.
9. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;307–10.
10. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8:135–60.
11. Nicholas DP. Choosing an intraclass correlation coefficient. UCLA: Academic Technology Services, Statistical Consulting Group. Available on the Internet, <http://support.spss.com/ProductsExt/SPSS/Documentation/Statistics/articles/whichicc.htm> accessed 17 December 2008.
12. Sanchez MM, Binkowitz BS. Guidelines for measurement validation in clinical trial design. *J Biopharm Stat* 1999;9:417–38.
13. Sailer J, Scharitzer M, Peloschek P, Giurea A, Imhof H, Grampp S. Quantification of axial alignment of the lower extremity on conventional and digital total leg radiographs. *Eur Radiol* 2005;15:170–3.
14. Modi HN, Chen T, Suh SW, Mehta S, Srinivasalu S, Yang J-H, et al. Observer reliability between juvenile and adolescent idiopathic scoliosis in measurement of stable Cobb's angle. *Eur Spine J* 2009;18:52–8.
15. Anderson JY, Trinkaus E. Patterns of sexual, bilateral and interpopulational variation in human femoral neck-shaft angles. *J Anatomy* 1998;192:279–85.
16. Husmann O, Rubin PJ, Leyvraz PF, de Rouguin B, Argenson JN. Three-dimensional geometry of the proximal femur. *J Bone Joint Surg Br* 1997;12:444–50.