Templating in uncemented total hip arthroplasty—on intra- and interobserver reliability and professional experience

Nils J. Strøm[1], Are Hugo Pripp[2], Olav Reikerås[1]

[1]Orthopaedic Department, Oslo University Hospital, Rikshospitalet, N-0027 Oslo, Norway; [2]Oslo Center of Biostatistics and Epidemiology, Research Support Services, Oslo University Hospital, Rikshospitalet, N-0027 Oslo, Norway

Contributions: (I) Conception and design: O Reikerås; (II) Administrative support: None; (III) Provision of study materials or patients: NJ Strøm, O Reikerås; (IV) Collection and assembly of data: NJ Strøm; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Prof. Olav Reikeras, MD, PhD. Orthopaedic Department, Oslo University Hospital, Rikshospitalet, N-0027 Oslo, Norway. Email: nilsst@ous-hf.no; olav.reikeras@ous-hf.no.

Background: This study examines the intra-, and interobserver reliability of digital templating in uncemented total hip arthroplasty (THA), and assesses whether these values are dependent on professional experience.

Methods: Three independent observers retrospectively examined digital X-rays of 34 consecutive hips scheduled for uncemented THA. These were templated using templating software. Evaluations were carried out on two occasions at least 6 weeks apart. Findings were compared to each surgeon's own findings, and then to the other surgeons' findings. Data underwent statistical analysis to assess and describe reliability.

Results: The intraobserver reliability of the method was found to be good. The intra-class correlation coefficient (ICC) for individual surgeons ranged from 0.81 to 0.87 for acetabular components and 0.74 to 0.91 for femoral components. However, it was somewhat lower for neck length with kappa statistics (κ) from 0.41 to 0.51 with agreement in about 70% of the cases. Interobserver reliability was similar, with an ICC of 0.87 for the acetabular component and 0.79 for the femoral component, but somewhat lower for neck length with κ of 0.27 and agreement in 41% of the cases. We found no association between increasing experience and increasing precision, as the least experienced observer showed the highest intraobserver reliability.

Conclusions: The reliability of digital templating of uncemented THA is good for acetabular and femoral components, but inferior for neck length. Precision does not rely on professional experience. Digital templating provides surgeons with a valuable tool for preoperative planning, but cannot supersede the intraoperative assessment and final decision.

Keywords: Digital templating; arthroplasty; hip; intraobserver; interobserver; reliability; experience

Submitted Nov 15, 2016. Accepted for publication Jan 04, 2017.
doi: 10.21037/atm.2017.01.73
View this article at: http://dx.doi.org/10.21037/atm.2017.01.73

Introduction

Templating has been described as an integral part of preoperative planning in total hip arthroplasty (THA), and has been unanimously advocated from Charnley until the present (1-5). It involves the use of a visual representation of the prosthesis in question, which is combined with X-rays of the patient's hip. The objective is to predict optimal sizes of the components to obtain correct offset, hip stability and equal limb length.

Digital radiography has led to development of templating software. However, reliable measurements presume accurate magnification of the hip on the radiograph and identical magnification of the template, and controversy exists whether digital templating is more precise than conventional templating. Some studies have reported a higher accuracy for acetate templating (6,7), whereas other studies reported better accuracy for digital templating (8,9).

There is a tendency towards digitalization of radiology, and acetate X-rays are no longer produced in our department.

Several studies have been published, that aim to evaluate the accuracy, precision and reliability of preoperative templating in THA (10-14). The results are predominantly reported as favourable, and the recommendation to utilize the method is strong (15). However, most studies report on at least partially cemented prostheses, which is less dictated by anatomical constraints than uncemented prostheses. Inaccuracies during templating can therefore be amended, or hidden, by cement. Also, size gaps in cemented stems series are usually larger than in uncemented stem series. Templating cemented stems is therefore more forgiving. The literature suggests that a ±1 size estimation of the component sizes is adequate for a templating method (6,10,11,14,16-19). We would argue, however, that the clinical value of evaluating and reporting ±1 size is limited, especially for uncemented components, as a change in 1 size from template to surgery is likely to affect both leg length and offset. Hence, we asked (I) what are the intra-, and interobserver reliability of preoperative digital templating in uncemented THA? And, (II) are these values dependent on the experience of the surgeon?

Methods

The study was performed in accordance with the ethical standards of the 1975 Declaration of Helsinki as revised in 2008. Data were anonymized and treated according to the ethical standards of our institution. As the study was not interventional, the patients followed ordinary routines and could not be identified. Therefore specific ethics approval and patient consent were not required.

We reviewed the X-rays of 42 consecutive patients who underwent primary THA at our institution. The variables collected included templated implant size for acetabular, femoral and neck components. Three independent observers conducted measurements twice, separated by a time interval of at least 6 weeks to minimize recall bias. These had different levels of professional experience: (I) sixth-year resident; (II) senior chief attending surgeon and (III) chief attending surgeon. Only the 1st measurement was used to assess interobserver, i.e., between surgeons, reliability. Randomization of the X-rays, data collection and analysis were performed by independent evaluators who were not the observers.

Eight patients were excluded due to inadequate X-rays, either lacking sufficient representation of the calibration marker, or inadequate exposure of, or malrotation of the joint. This lefts 34 patients for further examination: 22 women (65%) and 12 men (35%). The age ranged from 13 to 82 years, with a mean of 51 years. The indication for THA was primary osteoarthritis in 15 (44%), avascular necrosis of the femoral head in 6 (18%), developmental dysplasia in 6 (18%), Legg-Calvé-Perthes disease in 4 (12%), and miscellaneous in 3 (9%).

All patients were planned to receive a THA using Zimmer Trilogy uncemented shell, and DePuy Corail uncemented femoral stems. The available acetabular implants ranged from 40 to 68 mm in 2 mm increments, and the femoral implants ranged from 8 to 18 in 11 size units. The neck lengths ranged from short via medium to long.

All the radiological examinations were performed digitally at the same radiological centre, using a standardized protocol. We used a calibration marker of 36 mm positioned between the patient's legs, as close to the focal point of the X-ray beam as practically possible.

Templating was performed with a digital radiograph planning software (EndoMap, Siemens, Nuremberg, Germany), which is routinely used in our clinic, and all surgeons are trained in using this software to position the templates within the anatomical borders.

Statistical analysis

We considered 0.75 as the minimal acceptable value for intra-class correlation coefficient (ICC), and expected an ICC value of 0.90, hence a sample size of n>26 for a test-retest design (k=2) was estimated (significance level =0.05 and power =0.80) (20).

The differences between measurements on same patient, the ICC and the repeatability coefficient (RC) were used to assess reliability for acetabular and femoral implant sizes. Differences between measurements were described by their mean, standard deviation (SD) and range that statistically include 95% of the observed differences (mean ±1.96 × SD). The ICC describes how strongly units in the same group resemble each other; with 1.0 as perfect and 0 as agreement just by chance.

The RC provides a more direct clinical measure on reliability. If the difference between two measurements made on a subject (i.e., a patient) is approximately normally distributed, the absolute difference between the two measurements is in the long run expected to differ by no more than the RC on 95% of occasions. On a relative basis, a lower RC indicates a better reliability.

Table 1 Intraobserver reliability for acetabular and femoral component sizes assessed by the interclass correlation coefficient, the difference (1st–2nd measurement) and the RC

Intraobserver	Agreement, ICC (95% CI)	Difference, mean (SD): range	RC	Agree 1 size	Agree 2 sizes
Acetabular component					
Surgeon 1	0.87 (0.76–0.93)	0.4 (1.5): –2.5–3.2	2.9	68	91
Surgeon 2	0.82 (0.67–0.91)	0.5 (1.3): –2.1–3.0	2.6	74	90
Surgeon 3	0.83 (0.70–0.91)	–0.3 (1.5): –3.3–2.6	3.0	59	94
Combined	0.85 (0.76–0.91)	0.2 (1.4): –2.7–3.0	2.9	67	92
Femoral component					
Surgeon 1	0.91 (0.83–0.95)	0.2 (0.7): –1.2–1.6	1.4	97	100
Surgeon 2	0.89 (0.79–0.94)	0.0 (0.9): –1.7–1.7	1.7	94	100
Surgeon 3	0.74 (0.56–0.86)	0.5 (1.4): –2.1–3.2	2.8	79	94
Combined	0.83 (0.74–0.90)	0.3 (1.0): –1.8–2.3	2.1	90	98

RC, repeatability coefficient; ICC, intra-class correlation coefficient; 95% CI, 95% confidence interval; SD, standard deviation; range, range that statistically include 95% of the observed difference; agree 1 size, percentage of cases with agreement within 1 size difference; agree 2 size, percentage of cases with agreement within 2 sizes difference.

For neck lengths, reliability was assessed using the kappa statistics (κ) (21). With complete agreement, then κ=1, and with no agreement among raters other than as expected by chance, then κ=0. The percentage of cases with agreement within 1 and 2 sizes differences by the same surgeon (intraobserver) or between the three surgeons (interobserver) was estimated.

Reliability was graphically assessed using pairplots. Pairplots is a user-developed package in Stata that plots paired data using lines to show the difference between values.

All statistical analyses were done in Stata 13 (StataCorp LP, College Station, TX, USA).

Results

We found good intraobserver reliability with an ICC from 0.81 to 0.87 for the acetabular component and from 0.74 to 0.91 for the femoral component. The combined RCs were 2.9 and 2.1 for acetabular and femoral component sizes, respectively (*Table 1*).

The interpretation is that a surgeon differs between two measurements on the same patient very rarely by more than 3 size units for acetabular and by more than 2 size units for femoral components. In 67% and 90% of the cases there were intraobserver agreements within 1 size difference for acetabular and femoral components, respectively. The intraobserver reliability is graphically illustrated in *Figures 1*,*2* and shows that in about half of the patients there was negligible difference between first and second measurement. Due to space constraints, we present plots solely for surgeon 2.

Neck length was classified in short, medium or long, and seemed to have somewhat lower intraobserver reliability with κ from 0.41 to 0.50 (*Table 2*). Agreement between 1st and 2nd measurement was about 70% for all surgeons, but approximately 40% to 50% agreement could statistically be expected just by chance. Surgeon 1, 2 and 3 had agreement within 1 size difference in 91%, 100% and 97% of the cases, respectively. The lowest κ was observed for surgeon 3 with complete agreement for 25 (74%) patients.

The interobserver reliability for acetabular and femoral component had ICC of 0.87 and 0.79 and RC of 2.7 and 2.5, respectively (*Table 3*), and thereby similar to the intraobserver reliability.

An interpretation of the RC estimates is that two surgeons differ between measurements on the same patients very rarely by more than 3 size units for acetabular and femoral component. About half of the patients (44% for acetabular and 59% for femoral) had an agreement within 1 size difference among the three surgeons. A somewhat lower interobserver reliability was found for neck length with a combined κ of 0.27. For 14 (41%) of the patients all three surgeons agreed completely and for 32 (94%) of patients the surgeons agreed within 1 size difference in neck

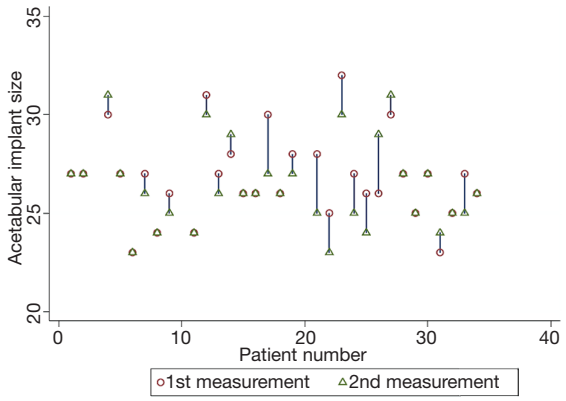

Figure 1 Intraobserver reliability on acetabular implant size assessed by a pair plot for surgeon 2. Paired data is plotted using lines to visualize the difference between values. Superimposed plots signify coinciding measurements.

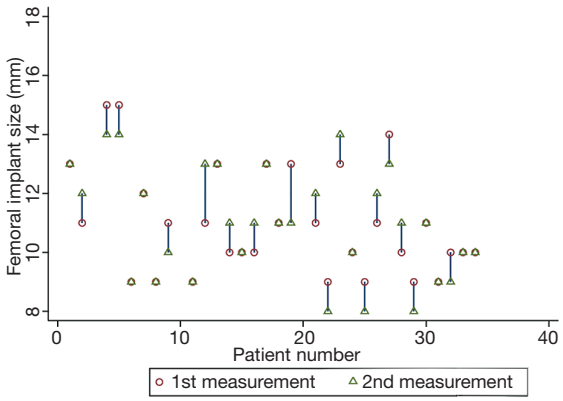

Figure 2 Intraobserver reliability on femoral implant size assessed by a pair plot for surgeon 2. Paired data is plotted using lines to visualize the difference between values. Superimposed plots signify coinciding measurements.

Table 2 Intraobserver reliability given cross-tabulation of neck length from 1st and 2nd measurement and corresponding the kappa coefficient for each surgeon

Intraobserver	Neck length, 1st measurement	Neck length, 2nd measurement			κ	Agree 1 size
		Short	Medium	Long		
Surgeon 1	Short	3	1	0	0.50	91
	Medium	3	16	0		
	Long	3	3	5		
Surgeon 2	Short	7	3	0	0.47	100
	Medium	1	10	5		
	Long	0	1	3		
Surgeon 3	Short	5	4	1	0.41	97
	Medium	1	20	2		
	Long	0	1	0		

κ, Kappa coefficient; agree 1 size, percentage of cases with agreement within 1 size difference.

length size.

We found that the least experienced surgeon, surgeon 1, had the highest ICC, i.e., was most consistent, for templating the acetabular and femoral components, and had the highest Kappa coefficient for neck length. The most experienced surgeon, surgeon 2, had intermediate ICC for templating the femoral component, and lowest ICC for the acetabular component, as well as intermediate Kappa coefficient for neck length. Hence, we did not find an association between professional experience and intraobserver reliability.

Discussion

An increasing number of studies aim to evaluate the accuracy, precision and reliability of preoperative templating in THA, and the results are predominantly reported in favour of templating (10-15). Most publications follow the example of Knight and Atwater, who used a ±1 size approximation of the component sizes in their evaluation. Intraobserver agreement of the acetabular and femoral components is reported at 60% to 85% (6,10,11,14,16-19). Within 1 size, we found that in between 67% and 90% of the cases there were intraobserver agreements for acetabular and femoral components, respectively. We would argue, however, that the clinical value of reporting ±1 size is limited, particularly for uncemented components, as a change in 1 size from template to surgery is likely to affect both leg length and offset. We therefore prefer to present our primary findings as ICC and Kappa values, but with a thorough statistical analysis. We did not investigate component placement, as this adds several magnitudes of bias.

Table 3 Interobserver reliability for acetabular and femoral component sizes assessed by the interclass correlation coefficient and the RC

Interobserver reliability	Agreement, ICC (95% CI)	RC	Agree 1 size	Agree 2 sizes
Acetabular component	0.87 (0.78–0.92)	2.7	44	88
Femoral component	0.79 (0.66–0.88)	2.5	59	88

RC, repeatability coefficient; ICC, inter-class correlation coefficient; 95% CI, 95% confidence interval; agree 1 size, percentage of cases with agreement within 1 size difference; agree 2 size, percentage of cases with agreement within 2 sizes difference.

We found good intraobserver reliability with an ICC from 0.81 to 0.87 for the acetabular component and from 0.74 to 0.91 for the femoral component. Our findings compare favourably with other reports of templating uncemented THA (4,13,14,16-19). If an agreement within 2 to 3 component sizes is clinically relevant for acetabular and femoral components, reliability can be satisfactory. Reliability for neck length on the other hand seemed low, with Kappa values from 0.41 to 0.50. This can be explained by fluctuating or individual preferences for medicalization of the acetabulum and axial stem placement. Hence, the choice of neck length is dependent on two other factors, each adding uncertainty. Varying the neck length is used perioperatively to provide optimal conditions for leg length and biomechanics around the hip joint. A possible interpretation of these findings is that there is good agreement on templating component sizes among surgeons, but poorer agreement on component placement. It follows that preoperative templating cannot supersede the intraoperative assessment and final decision.

The influence of professional experience on the precision of templating was evaluated in this study. In the current literature this is controversial, and contradictory results have been reported (10,13). While two studies found a significant influence of the professional experience on templating the components for THA, two recently published studies did not find a significant influence (22,23). However, it should be noted that in these studies, the different types of observers are not directly comparable. What constitutes an "experienced" and "less experienced" surgeon is not universally defined, and this is likely to cause some of these discrepancies. Some of the studies also included cemented as well as uncemented components, and had small numbers in some of the subgroups. In our study we templated only uncemented components that intended to be inserted with press fit. Anatomical borders are therefore more clearly defined. We found that the least experienced surgeon, surgeon 1, had the highest ICC, i.e., was most consistent, for templating the acetabular and femoral components, and highest Kappa value for neck length. The most experienced surgeon, surgeon 2, had intermediate ICC for templating the femoral component, and lowest ICC for the acetabular component, as well as intermediate Kappa coefficient for neck length. Hence, at these experience intervals, or at least, above our threshold of experience, we could not find a correlation between experience and consistency. We did not compare the templates with the components chosen during surgery, however, so it is not possible to tell how the precision correlates with the prediction of actual implant (accuracy).

There are a few limitations of our study. First, our study was carried out in only one clinic, and there are concerns whether the findings can be generalized. Second, we did not investigate in a longitudinal fashion. However, all observers were experienced in using the software, so we do not expect any significant change in measurements over time. Third, the number of patients could be higher. But our power analysis suggests that the number is sufficient to yield significant results (20). Also, our sample population had an adequate distribution of all values to make the analyses meaningful. Fourth, there is no golden standard that we could compare our results with. But this is a general and main objection to preoperative templating of joint prostheses. Fifth, we concede that radiographic interpretation relies on the clinician's experience with reading hip radiographs. Main sources of errors when measuring radiographs are errors in locating corresponding landmarks, and errors in calibration on digital radiographs. However, our findings suggest that above a certain threshold of experience, this is mainly a methodological problem, and is not likely to infer further bias to our results.

In conclusion, if an agreement within 2 to 3 component sizes is clinically relevant for uncemented acetabular and femoral components, reliability can be satisfactory. However, preoperative templating cannot supersede the intraoperative assessment and final decision. Also, professional experience is not correlated to intraobserver reliability. Hence, lack of precision is not due to lack of experience, but inherent to the method.

Acknowledgements

The authors wish to thank John Magnar Slåstad and Anne Guro Vreim Holm for their help with the study.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

Ethical Statement: As the study was not interventional, the patients followed ordinary routines and could not be identified. Therefore specific ethics approval and patient consent were not required.

References

1. Charnley J. Low friction arthroplasty of the hip. Springer-Verlag Berlin Heidelberg, 1979.
2. Crutcher JP Jr. Preoperative planning for total hip arthroplasty. Oper Tech Orthop 2000;2:102-5.
3. Della Valle AG, Padgett DE, Salvati EA. Preoperative planning for primary total hip arthroplasty. J Am Acad Orthop Surg 2005;13:455-62.
4. Gamble P, de Beer J, Petruccelli D et al. The accuracy of digital templating in uncemented total hip arthroplasty. J Arthroplasty 2010;25:529-32.
5. Marcucci M, Indelli PF, Latella L, et al. A multimodal approach in total hip arthroplasty preoperative templating. Skeletal Radiol 2013;42:1287-94.
6. González Della Valle A, Comba F, Taveras N et al. The utility and precision of analogue and digital preoperative planning for total hip arthroplasty. Int Orthop 2008;32:289-94
7. Iorio R, Siegel J, Specht LM, et al. A comparison of acetate vs digital templating for preoperative planning of total hip arthroplasty: is digital templating accurate and safe? J Arthroplasty 2009;24:175-9.
8. The B, Verdonschot N, van Horn JR, et al. Digital versus analogue preoperative planning of total hip arthroplasties: a randomized clinical trial of 210 total hip arthroplasties. J Arthroplasty 2007;22:866-70.
9. Whiddon DR, Bono JV, Lang JE, et al. Accuracy of digital templating in total hip arthroplasty. Am J Orthop (Belle Mead NJ) 2011;40:395-8.
10. Carter LW, Stovall DO, Young TR. Determination of accuracy of preoperative templating of noncemented femoral prostheses. J Arthroplasty 1995;10:507-13.
11. González Della Valle A, Slullitel G, Piccaluga F, et al. The precision and usefulness of preoperative planning for cemented and hybrid primary total hip arthroplasty. J Arthroplasty 2005;20:51-8.
12. Eggli S, Pisan M, Muller ME. The value of preoperative planning for total hip arthroplasty. J Bone Joint Surg Br 1998;80:382-90.
13. Kearney R, Shaikh AH, O'Byrne JM. The accuracy and inter-observer reliability of acetate templating in total hip arthroplasty. Ir J Med Sci 2013;182:409-14.
14. Knight JL, Atwater RD. Preoperative planning for total hip arthroplasty. Quantitating its utility and precision. J Arthroplasty 1992;7 Suppl:403-9.
15. Petretta R, Strelzow J, Ohly NE, et al. Acetate templating on digital images is more accurate than computer-based templating for total hip arthroplasty. Clin Orthop Relat Res 2015;473:3752-9.
16. Crooijmans HJ, Laumen AM, van Pul C, et al. A new digital preoperative planning method for total hip arthroplasties. Clin Orthop Relat Res 2009;467:909-16.
17. Davila JA, Kransdorf MJ, Duffy GP. Surgical planning of total hip arthroplasty: accuracy of computer-assisted EndoMap software in predicting component size. Skeletal Radiol 2006;35:390-3.
18. Kumar PG, Kirmani SJ, Humberg H, et al. Reproducibility and accuracy of templating uncemented THA with digital radiographic and digital TraumaCad templating software. Orthopedics 2009;32:815.
19. The B, Diercks RL, van Ooijen PM, et al. Comparison of analog and digital preoperative planning in total hip and knee arthroplasties. A prospective study of 173 hips and 65 total knees. Acta Orthop 2005;76:78-84.
20. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med 1998;17:101-10.
21. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37:360-3.
22. Efe T, El Zayat BF, Heyse TJ, et al. Precision of preoperative digital templating in total hip arthroplasty. Acta Orthop Belg 2011;77:616-21.
23. Jung S, Neuerburg C, Kappe T, et al. Validity of digital templating in total hip arthroplasty: impact of stem design and planner's experience. Z Orthop Unfall 2012;150:404-8.

Cite this article as: Strøm NJ, Pripp AH, Reikerås O. Templating in uncemented total hip arthroplasty—on intra- and interobserver reliability and professional experience. Ann Transl Med 2017;5(3):43. doi: 10.21037/atm.2017.01.73

