A Semi-automated Measurement Technique for the Assessment of Radiolucency

Pegg, EC [1], Kendrick, BJL [1], Pandit, HG [1], Gill, HS [2], Murray, DW [1]

[1] Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK, OX3 7LD.

[2] Department of Mechanical Engineering, University of Bath, Bath, UK, BC2 7AY

Summary

The assessment of radiolucency around an implant is qualitative, poorly defined, and has low agreement between clinicians. Accurate and repeatable assessment of radiolucency is essential to prevent misdiagnosis, minimise cases of unnecessary revision, and to correctly monitor and treat patients at risk of loosening and implant failure. The purpose of this study was to examine whether a semi-automated imaging algorithm could improve repeatability, and enable quantitative assessment of radiolucency. Six surgeons assessed 38 radiographs of knees after unicompartmental knee arthroplasty for radiolucency, and results were compared with assessments made by the semi-automated program. Large variation was found between the surgeon results, with total agreement in only 9.4% of zones and a Kappa value of 0.602; whereas the automated program had total agreement in 81.6% of zones and a Kappa value of 0.802. The software had a 'fair to excellent' prediction of the presence or absence of radiolucency, where the area under the curve (AUC) of the receiver operating curves (ROC) was 0.82 on average. The software predicted radiolucency equally well for cemented, and cementless implants (p=0.996). The identification of radiolucency using an automated method is feasible and these results indicate it could aid in the definition and quantification of radiolucency.

Keywords

Radiolucency, Knee, Measurement, Reliability

Introduction

After arthroplasty, radiolucent lines can often be observed on radiographs surrounding implanted components. The lines are normally noted within 1 year of implantation, but can develop up to 2 years after implantation [1]. There is sometimes a misconception among surgeons that the presence of radiolucency around an implant is always indicative of loosening [2-4]. However, Goodfellow *et al.* noted that the presence of a radiolucency does not always indicate that implant loosening will occur; the authors described two types of radiolucency, pathological, or physiological [1]. Pathological radiolucent lines increase in thickness with time, and are generally more than 2 mm thick and are indicative of a problem such as infection or loosening [1, 5]. Physiological radiolucent lines

are generally well defined with a sclerotic margin, non-progressive, less than 2 mm thick, and are not indicative of loosening [1, 5]. It is, therefore, important to correctly distinguish between pathological and physiological radiolucent lines, to ensure that unnecessary revision of well-fixed components does not occur [2-4].

The positive identification of a radiolucent line within a radiograph is generally performed visually by clinicians. The process can be subjective, the experience and speciality of the clinicians performing the assessment can vary, and accurate detection relies upon high quality radiographs. Surgeons are not trained formally in how to identify radiolucency at any stage of their training, and there are no guidelines on how it should be done. In practice, clinical judgement will not only be influenced by the presence of a radiolucent region on a radiograph, but also by the reported patient symptoms and by observing how the observed radiolucency changes with time.

It is important to control the angle at which the radiograph is taken, as the implant may obscure the radiolucency; fluoroscopically guided radiographs have been shown to be the most accurate measurement method [6], however, these are not always used [7]. McCaskie *et al.* examined the reliability of detecting radiolucencies within hip radiographs and found that there was wide variation in the agreement between the raters, and this was equally variable for experienced as well as inexperienced clinicians [8]. In addition, it has been shown that surgeons are unable to reliably detect radiolucent lines less than 0.7 mm thick [9].

In order to ensure the early identification of progressive radiolucency and to minimise misdiagnosis leading to unnecessary revision of implants with physiological radiolucencies, it is important that the measurement of radiolucencies is standardised and the reliability improved. Kobayashi *et al.* defined radiolucency as being "radiolucency surrounded by lines of increased density" [10, 11]. By using this definition it is possible to automate the analysis of radiolucent lines, using image processing techniques. By using a computerised process, the subjectivity of the assessor is removed, and more quantitative results can be obtained.

There is evidence that it is more difficult to distinguish between a physiological radiolucent line and a pathological one after unicompartmental knee arthroplasty (UKA) [12]. The consequence of such misdiagnosis may be the unnecessary revision of the component. It is known that the revision rate after UKA is significantly higher than total knee arthroplasty (TKA) [13], and it is possible that poor identification of radiolucency may be a factor.

For this reason, the purpose of this study was to examine the feasibility of using an automated computer program to help standardise the measurement of radiolucency after UKA. The study aimed

to (1) quantify the variability between surgical identification of radiolucency, (2) compare the results of the automated procedure with the average surgical consensus, (3) examine whether the software algorithm was robust enough to cope with cemented as well as cementless components and (4) use the software to create a quantifiable definition of radiolucency around a UKA.

Methodology

Image preparation

Thirty-eight fluoroscopy-screened radiographs [1] were examined from patients who had undergone UKA at the Nuffield Orthopaedic Centre in Oxford. The radiographs were from an on-going randomised control trial investigating cementless, and cemented, implant migration using Radiostereometric Analysis (RSA) in 40 patients; ethical approval was attained with informed consent to use the radiographs (Oxford REC B in 2002, reference C01-101). All patients had been implanted with the medial Oxford UKA (Biomet UK Ltd., Swindon, UK), and the radiographs were taken within 2 years of implantation. Of the 38 images, 19 were of cementless implants and 19 were of cemented components.

Manual Assessment of Radiolucency

The radiographs were visually assessed for radiolucency by six orthopaedic surgeons. The surgeons were all either orthopaedic registrars or orthopaedic consultants. The surgeons were provided with the radiographs in JPEG file format which had been exported from the PACS hospital system software (average image resolution of 183.6 DPI ± 20.5) and a Microsoft Excel-based form to enter their data. Each surgeon then assessed the radiographs for radiolucency in their own time on computers of their choice; all the computers chosen were office-based systems. The tibial tray was considered to be divided into seven zones, previously defined by Hooper *et al.* [14] (Figure 1), and the surgeon recorded whether radiolucency was present or not in each zone. In order to best reproduce the clinical situation, the surgeons were not given any guidance as to how to identify radiolucency within the zones; they were just presented with the radiographs, and a form to complete containing an image illustrating the location of the zones.

Semi-Automatic Radiolucent Area Quantification (SARAQ)

The SARAQ program was created to automate assessment of radiolucency and was written in MATLAB (Version 2010a, MathWorks Inc. Natick, MA, USA). The radiograph images were processed in three stages. Firstly, the program identified where the tibial tray was in the radiograph by iterating from a user-defined start point; secondly, the change in greyscale normal to the tray was recorded; finally the series of greyscale profiles were analysed and the presence of radiolucency

assessed. In order to get an idea of the intra-observer reliability of the SARAQ software, all radiographs were processed six times.

To locate the tibial tray within the image, a technique called active shape modelling was used, which was originally described by Cootes *et al.*[15]. The method involves first 'training' the model, where specific landmarks on a range of similar images are manually selected. Please note, training of the model need only be performed once to define the shape for the model parameters for any given type of implant; once defined, the model could be run on as many images as required without re-training the model. The training information was processed to record the expected shape and intensity profile for the particular application. For the present study, an active shape model which had been previously validated for the Oxford UKA was used [16].

The model was trained on 36 training radiographs; if necessary, images were flipped to ensure that the implanted side was on the right-hand side of the image. Fifty-three landmark points surrounding the tibial tray were then manually selected for each training image (this process took approximately 15 minutes per image); these points were then interpolated to give a total of 989 co-ordinates. All shapes were aligned using the Procrustes method [17]; the shapes were then translated so that the centroid of the shape was at the origin; rotation effects were removed by using the mean angle to the centroid, finally scaling was removed by normalising the image by the tibial width [16]. Principal component analysis was then performed on the training data shape co-ordinates, and the greyscale profiles (12 pixels long) normal to each point and the greyscale differences in the profiles which were normalised by the average profile. Once this process was completed and the shape model had been defined for the shape of the UKR, the shape model could then be applied to numerous radiographs to locate the tibial tray. Further information on the definition and validation of the model has been described in a previous publication [16]; the study demonstrated measurements taken with the shape model were significantly quicker, had equivalent accuracy and higher reliability than manual measurements.

To run the model on a radiograph, the user inputted the starting position, and then the software performed 40 iterations to find the final shape which best correlated with the training data. Radiographic magnification was calculated using the spherical femoral component as a reference value to perform the calibration between image and real-world dimensions; points circling the femoral component were found using edge detection, and then a circle was fitted using a least-squares algorithm [16].

Once the region of the tibial tray within the image had been identified, the image was cropped to just include the proximal tibia, and resized to a 700x700 pixel resolution for consistency. The greyscale profiles emanating perpendicular to the tray through the bone over a distance of 50 pixels were then

recorded. Due to the landmark point definition of the shape model, the locations of corners of the tibial tray were known and therefore the profiles could be split into the relevant zones (Figure 2). The profiles were then normalised to minimise any effects from radiographic intensity; the maximum intensity (implant region) was set to 1, and the minimum intensity to zero.

Each individual profile was examined for radiolucency. All profiles were smoothed using a local regression with a weighted linear least squares and a 1st degree polynomial model (LOWESS), to reduce scatter and emphasise trends. Radiolucency was deemed to be present where there was a dip in the greyscale intensity, followed by a rise of sufficient height (Figure 3). The height difference was required to be more than 5% of the maximum greyscale intensity within the zone; otherwise the radiolucency was not recorded. The width of the radiolucent line was defined as the distance along which the intensity was in the lower half of the intensity of the radiolucent region (i.e. the mean intensity of the dip and the peak).

For a particular zone, the distance (in millimetres) between the peaks of each profile defined as being radiolucent was averaged to give the average width of the radiolucent line. The width was then multiplied by the length over which the line was present, to give a radiolucent "area" in millimetres squared. The radiolucent area was then used as a score to define whether or not radiolucency was present within the zone.

Statistical Analysis

To quantify how well the SARAQ software predicted the surgical assessment of radiographs, receiver operating characteristic (ROC) analysis was performed using the pROC package in R [18] (version 2.15.1, http://www.r-project.org). The areas under the curve (AUC) results were analysed using the guidelines from Haase-Fielitz *et al.* [19] where; 0.9-1.0 indicates an excellent prediction, 0.80-0.89 good, 0.70-0.79 fair, 0.60-0.69 poor, and 0.50-0.59 no useful value. DeLong's statistical test was used to determine whether there was a statistical difference between each individual surgeon's AUC result and that of the entire combined dataset [20]. The AUCs were calculated and compared for; each individual surgeon and the combined dataset; cemented versus cementless radiographs, and each individual radiographic zone. The prevalence of radiolucency in the cemented and cementless radiographs was also examined using a Chi-squared test. The optimum SARAQ area threshold values (mm^2), above which a radiographic zone would be defined as being radiolucent, were calculated from the ROC curves for each radiographic zone using the Youden method [21].

The reliability of the manual and semi-automatic measurement methods was compared. Agreement between the surgical ratings of radiolucency was assessed using the Fleiss Kappa statistic [22] with the guidelines detailed by Landis and Koch [23]; where Kappa values <0 indicated poor agreement, 0-

0.20 slight agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, 0.61-0.80 substantial agreement, and more than 0.81 to be almost perfect agreement. The software reliability was assessed using two methods; the Fleiss Kappa statistic and the Intraclass Correlation Coefficient (ICC). The Kappa statistic was calculated to examine how reliable the SARAQ software was in detecting the presence or absence of radiolucency (binary data); the presence or absence was calculated using the threshold values calculated using the Youden method. The Intraclass Correlation Coefficient (ICC) was calculated to assess the reliability of the radiolucency area measurement (continuous data) and radiolucency thickness measurement (continuous data); a two-way mixed model with single measures was used.

Results

The AUC values for how well the SARAQ software predicted the combined surgeon results and the individual surgeon results were examined (Figure 4). The AUC values varied from 0.75 to 0.93 depending on the surgeon (Figure 5); this represents a fair to excellent prediction. The AUC of the combined dataset was 0.82 (95% CI: 0.75-0.88). When the individual surgeon ROC curve areas were compared statistically against the combined dataset, five of the six were not statistically different, but one was significantly higher ($p<0.0001$) than the combined dataset.

Out of the zones where the SARAQ software identified radiolucency, the average thickness was 1.05 mm with a standard deviation of 0.50 mm; the maximum thickness measured was 2.71 mm (Figure 6).

The ROC curves examining the results from radiographs of cemented and cementless components were similar (Figure 7). The AUC for the cemented results was 0.824 (CI: 0.740-0.908), and for the cementless results was 0.857 (CI: 0.762-0.952), and statistically no significant difference was found between the two ($p=0.996$). Of the radiographs examined, a lower incidence of radiolucency was found for the cementless components (surgeons determined radiolucency to be present in 21.2% of the cementless zones) compared with the cemented components (radiolucency present in 30.3% of the cemented zones); this difference was significant ($p = 0.028$).

The SARAQ software demonstrated varying accuracy in predicting radiolucency in different zones; AUC values ranged from 0.60 (zone 4, CI: 0.45-0.74) to 0.82 (zone 7 CI: 0.76-0.89). The estimated threshold value for the optimum cut-off for prediction of radiolucency were different for each zone (Table 1); the lowest threshold was found for zone 5 at 0.0075 mm^2, and the highest was for zone 4 at 0.0594 mm^2. The calculated threshold for the entire dataset was 0.0266 mm^2. Pair-wise comparisons showed that the AUC for zone 7 was statistically different to zone 1 ($p=0.015$), zone 4 ($p=0.006$) and zone 6 ($p=0.002$); no other differences were found between the zones.

The Kappa statistic calculated for the surgeon agreement was 0.602 ($p < 0.0001$). According to Landis and Koch's definition, this represents a substantial agreement. However, out of the 38 radiographs examined, the surgeons were in absolute agreement for only 25 zones out of the 266 measured, which is 9.4%. The Kappa statistic calculated for the SARAQ software agreement was 0.802 ($p<0.0001$) which is very close to being an 'almost perfect agreement' ($\kappa > 0.81$) as defined by Landis and Koch, and the SARAQ software was in absolute agreement for 217 zones of the 266 measured which is 81.6%. The ICC for the software thickness measurements was 0.704, and for the software area results was 0.881, representing good to excellent reliability for both measurements (perfect reliability = 1).

Discussion

Although radiographs are routinely examined for the presence of radiolucency, there is a surprising lack of guidance on how to define radiolucency. Kobayashi's definition, that a radiolucency must consist of a radiolucent line surrounded by lines of increased density [10], is clear; but it is a qualitative description which is open to interpretation. The variation in surgical assessment of radiolucency in the present study was relatively large, with the surgeons only being in total agreement for 9.4% of the measurements; however, the Kappa statistic (0.602) indicated substantial agreement according to Landis and Koch's definition. The surgeon agreement found in the present study is greater than that reported by McCaskie *et al.* [8]. McCaskie *et al.* examined radiolucency measurement reliability in different regions of hip radiographs; Kappa values ranged from -0.520 to 0.373. From these data overall, it can be concluded that there is a clinical need for a more reliable method to measure radiolucency after joint replacement.

The reliability of the SARAQ software was shown to be higher than the surgical assessment of radiolucency in terms of the Kappa statistic (0.802), and also the software was found to be in absolute agreement for 82% of the zones which was far greater than the 9.4% of absolute agreement from the surgeons. The resolution of the radiographs analysed in this study was relatively low (183.6 DPI), however this represents the quality of radiographs which would be routinely assessed clinically. The resolution was limited by the clinical requirement to minimise patient radiation levels and also compounded by the small size of the unicompartmental knee replacement; yet despite this the SARAQ software was able to reliably analyse the radiographic images, which was encouraging. The reason the software was not in perfect agreement when run multiple times was due to the variation in the manually assigned start position for the active shape model, resulting in slightly different profiles each time the software was run. This result demonstrates the SARAQ software provided improved reliability to the surgical assessment and previous work has shown the software takes only 2 minutes

to analyse one radiograph so its use clinically would be viable [16]; however, reliability is of no use without accuracy.

The SARAQ software demonstrated promising results in terms of accuracy of radiolucency assessment; the software had an average AUC value of 0.82 which represents a 'good' agreement between the surgical assessment and the software assessment. The majority of radiolucency thicknesses measured by the SARAQ software were below 2 mm, indicating these were physiological radiolucencies; this correlates well with the literature [1, 12]. Some thicker pathological radiolucencies were measured, but none exceeded 3 mm, which was in accordance with Tibrewal *et al.* [1] who also found that radiolucencies did not exceed 3 mm.

There was a concern that the effectiveness of the SARAQ software in predicting radiolucency would differ for cemented versus cementless implants due to the radio-opaque nature of the cement influencing the algorithm. However, the ROC curve results demonstrated that there was no significant difference between the AUC values for the two implant designs, and visually the ROC curves were similar (Figure 7). To positively identify a radiolucency, the software needs to detect a decrease in the brightness of the profile followed by a subsequent increase; because the cement is more radiolucent that the bone the definition will work appropriately for both cemented and cementless components. However, this study has not considered whether the algorithm would be able to detect radiolucency in cases where radiolucent bone cement had been used; in addition, if there was debonding of the cement from the implant, the algorithm may misinterpret this as a radiolucency. The results of the study also showed a greater incidence of radiolucency for the cemented components compared with the cementless components, which correlates with other studies reported in the literature [24].

It is known that for the unicompartmental tibial tray, that some zones are more likely to be radiolucent than others [5]. For this reason, the radiolucency and cut-off thresholds were assessed for particular zones around the tray. Wide variation was found in the AUC values from the ROC curves for each zone, ranging from 0.59 to 0.82; however, only zone 7 was significantly different to other zones, indicating the range in AUC was largely due to a manifestation of background intensity variability across the radiographs, rather than representing real differences in the zones.

The AUC of zone 7 was significantly higher than the other zones; zone 7 is next to the wall of the tray, and is different from the other zones in that it is not axially loaded, cement is not applied to the zone [5], and in the case of cementless components, porous titanium coating is not present on the tray wall. The difference observed may be a consequence of these reasons, alternatively it may relate to a bony response to high stresses in the corner of the tibial tray [25, 26].

The present study has several limitations. The surgeons were given no guidance on how to identify radiolucent regions around the implant therefore the study was not controlled; however, this does represent the clinical situation where surgeons have been shown to use different assessment criteria for radiolucency. In addition, the surgeons were not guided on the monitor quality on which to perform their assessment of the radiographs; and monitor quality has been shown to be influential in radiographic assessment. The SARAQ software has been assessed in this study on only one single type of implant. The Oxford UKA has a unique design with a flat tibial implant surface which enables the bone-implant interface to be clearly imaged, provided that the radiograph is properly screened. This may not be possible with other joint replacement designs; therefore, whether the SARAQ software would work equally well on other designs has yet to be confirmed. In addition, the process of selecting the landmark points for the training data is time-consuming; therefore this method may not be suitable for studies comparing many different shapes/designs. As previously mentioned, it is yet unconfirmed as to whether the algorithm is able to identify a radiolucency in a radiograph containing radiolucent cement, and it is also likely that the software is unable to distinguish between cement debonding and radiolucency. Radiolucent cement is rarely used; therefore this is not expected to be a common issue. Cases of cement debonding are more common but it is important to note that the treatment for severe radiolucency and cement debonding are the same. When the individual zones were examined separately in this study, the sample size was too low to draw any conclusions on differences between the zones, with the exception of zone 7; a larger research study is underway to refine and optimise the program for specific zones.

Nevertheless, this study is promising and indicates that automated measurements of radiolucency could be a useful addition to clinical observance. The wide variation found in the surgical assessment of radiolucency was concerning, and a digital computerised tool may be the solution to enable standardisation and quantification of clinical radiolucency assessment. Such a tool may help to ensure that patients with progressive radiolucencies can be identified early and monitored using a reliable, quantitative method.

Acknowledgements

The authors would like to thank the surgeons who took part in the study; Mr Abtin Alvand, Mr Alex Liddle, Mr George Grammatopoulos and Mr James Pegrum. Some of the authors of this work have received financial support from Biomet Inc. (the manufacturer of the Oxford Unicompartmental Knee examined in this work); however, Biomet has not provided funding specifically for this study and Biomet have had no influence on, or knowledge of, this work.

Data Accessibility

A datasheet containing all the measurement data collected from the surgeons and the software measurements for each zone of each radiograph is provided as supplementary data. The pROC package used for the ROC curve analysis is available at: http://cran.r-project.org/web/packages/pROC.

References

[1] Tibrewal S, Grant K, Goodfellow J. 1984 The radiolucent line beneath the tibial components of the Oxford meniscal knee. *J Bone Joint Surg Br*. **66-B**, 523-8.

[2] Goodfellow JW, O'Connor JJ, Dodd CAF, Murray DW. 2006 *Unicompartmental arthroplasty with the Oxford knee*: Oxford University Press, UK.

[3] Clarius M, Hauck C, Seeger J, James A, Murray D, Aldinger P. 2009 Pulsed lavage reduces the incidence of radiolucent lines under the tibial tray of Oxford unicompartmental knee arthroplasty. *Int Orthop*. **33**, 1585-90. (DOI 10.1007/s00264-009-0736-y)

[4] Mukherjee K, Pandit H, Dodd CAF, Ostlere S, Murray DW. 2008 The Oxford unicompartmental knee arthroplasty: a radiological perspective. *Clin Radiol*. **63**, 1169-76. (DOI 10.1016/j.crad.2007.12.017)

[5] Gulati A, Chau R, Pandit HG, Gray H, Price AJ, Dodd CAF, et al. 2009 The incidence of physiological radiolucency following Oxford unicompartmental knee replacement and its relationship to outcome. *J Bone Joint Surg Br*. **91B**, 896-902. (DOI 10.1302/0301-620x.91b7.21914)

[6] Mintz AD, Pilkington CA, Howie DW. 1989 A comparison of plain and fluoroscopically guided radiographs in the assessment of arthroplasty of the knee. *J Bone Joint Surg Am*. **71**, 1343-7.

[7] Kalra S, Smith TO, Berko B, Walton NP. 2011 Assessment of radiolucent lines around the Oxford unicompartmental knee replacement: sensitivity and specificity for loosening. *J Bone Joint Surg Br*. **93-B**, 777-81. (DOI 10.1302/0301-620x.93b6.26062)

[8] McCaskie AW, Brown AR, Thompson JR, Gregg PJ. 1996 Radiological evaluation of the interfaces after cemented total hip replacement. Interobserver and intraobserver agreement. *J Bone Joint Surg Br*. **78**, 191-4.

[9] Zhang YD, Putnam AW, Heiner AD, Callaghan JJ, Brown TD. 2002 Reliability of detecting prosthesis/cement interface radiolucencies in total hip arthroplasty. *J Orthop Res*. **20**, 683-7. (DOI 10.1016/s0736-0266(02)00005-0)

[10] Kobayashi S, Takaoka K, Saito N, Hisa K. 1997 Factors affecting aseptic failure of fixation after primary Charnley total hip arthroplasty. Multivariate survival analysis. *J Bone Joint Surg Am*. **79**, 1618-27.

[11] Williams HD, Browne G, Gie GA, Ling RS, Timperley AJ, Wendover NA. 2002 The Exeter universal cemented femoral component at 8 to 12 years. A study of the first 325 hips. *J Bone Joint Surg Br*. **84**, 324-34.

[12] Kendrick BJL, James AR, Pandit H, Gill HS, Price AJ, Blunn GW, et al. 2012 Histology of the bone-cement interface in retrieved Oxford unicompartmental knee replacements. *Knee*. **19**, 918-22. (DOI 10.1016/j.knee.2012.03.010)

[13] NJR, 2013 *10th Annual Report*: National Joint Registry for England and Wales.

[14] Hooper GJ, Maxwell AR, Wilkinson B, Mathew J, Woodfield TBF, Penny ID, et al. 2012 The early radiological results of the uncemented Oxford medial compartment knee replacement. *J Bone Joint Surg Br*. **94-B**, 334-8. (DOI 10.1302/0301-620x.94b3.27407)

[15] Cootes TF, Taylor CJ, Cooper DH, Graham J. 1995 Active Shape Models-Their Training and Application. *Comput Vis Image Und*. **61**, 38-59. (DOI 10.1006/cviu.1995.1004)

[16] Pegg EC, Mellon SJ, Salmon G, Alvand A, Pandit H, Murray DW, et al. 2012 Improved radiograph measurement inter-observer reliability by use of statistical shape models. *Eur J Radiol*. **81**, 2585-91. (DOI 10.1016/j.ejrad.2011.12.018)
[17] Goodall C. 1991 Procrustes methods in the statistical analysis of shape. *J Roy Stat Soc B Met*. **53**, 285-339.
[18] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. 2011 pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. **12**, 77.
[19] Haase-Fielitz A, Bellomo R, Devarajan P, Story D, Matalanis G, Dragun D, et al. 2009 Novel and conventional serum biomarkers predicting acute kidney injury in adult cardiac surgery--a prospective cohort study. *Crit Care Med*. **37**, 553-60. (DOI 10.1097/CCM.0b013e318195846e)
[20] DeLong ER, DeLong DM, Clarke-Pearson DL. 1988 Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. **44**, 837-45.
[21] Youden WJ. 1950 Index for rating diagnostic tests. *Cancer*. **3**, 32-5.
[22] Fleiss JL. 1971 Measuring nominal scale agreement among many raters. *Psychol Bull*. **76**, 378-82. (DOI 10.1037/h0031619)
[23] Landis JR, Koch GG. 1977 The Measurement of Observer Agreement for Categorical Data. *Biometrics*. **33**, 159-74. (DOI 10.2307/2529310)
[24] Liddle AD, Pandit H, O'Brien S, Doran E, Penny ID, Hooper GJ, et al. 2013 Cementless fixation in Oxford unicompartmental knee replacement: A multicentre study of 1000 knees. *Bone & Joint Journal*. **95-B**, 181-7. (DOI 10.1302/0301-620x.95b2.30411)
[25] Pegg EC, Walter J, Mellon SJ, Pandit HG, Murray DW, D'Lima DD, et al. 2012 Evaluation of factors affecting tibial bone strain after unicompartmental knee replacement. *J Orthop Res*. n/a-n/a. (DOI 10.1002/jor.22283)
[26] Chang T-W, Yang C-T, Liu Y-L, Chen W-C, Lin K-J, Lai Y-S, et al. 2011 Biomechanical evaluation of proximal tibial behavior following unicondylar knee arthroplasty: Modified resected surface with corresponding surgical technique. *Med Eng Phys*. **33**, 1175-82. (DOI 10.1016/j.medengphy.2011.05.007)

Figure captions

Figure 1. A schematic illustration of the position of the seven defined zones surrounding the tibial implant.

Figure 2. An example of the greyscale profiles from each zone of a radiograph after processing.

Figure 3. Example of a smoothed greyscale profile of a radiolucent region, an image of the radiograph from which the profile was taken is shown and the profile is illustrated by a black line. The value *h* represents the difference in height between the second peak (more radiodense line) and the minimum of the dip prior to it (the radiolucent line). The value *t* represents the thickness of the radiolucent line.

Figure 4. ROCs illustrating the diagnostic capability of the radiolucent area parameter output by the software in predicting radiolucency as defined by the average data from the six surgeons, and the curves for each surgeon individually.

Figure 5. Summary of the AUC results calculated from the ROC curves for the six surgeons; error bars represent the 95% confidence intervals.

Figure 6. Boxplot of the distribution of the thickness of radiolucent lines measured by the automated program.

Figure 7. ROC curves for results from cemented and cementless components. No significant difference was found between the ROC curve areas (p=0.97).

Tables

Zone	AUC	95% CI AUC		Threshold (mm^2)
		lower	upper	
1	0.683	0.593	0.773	0.0189
2	0.765	0.660	0.871	0.0232
3	0.713	0.611	0.814	0.0294
4	0.596	0.452	0.741	0.0532
5	0.704	0.615	0.793	0.0075
6	0.668	0.594	0.742	0.0085
7	0.821	0.757	0.886	0.0281

Table 1. Summary of the AUC values measured for different zones and the estimated threshold for radiolucency area within each zone.

Short Title

Semi-automated Radiolucency Assessment

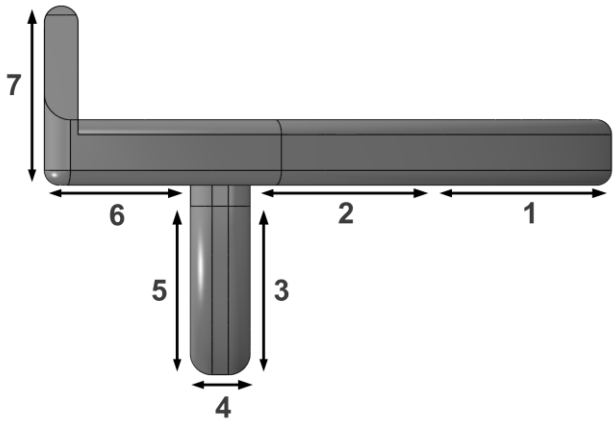

Figure 1

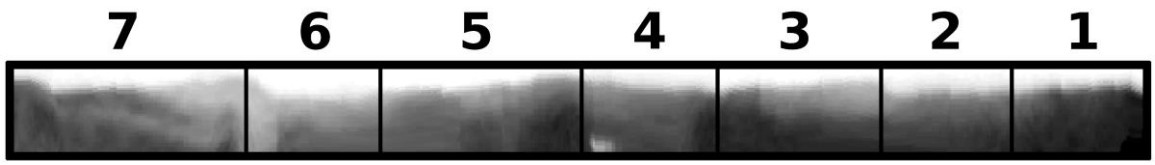

Figure 2

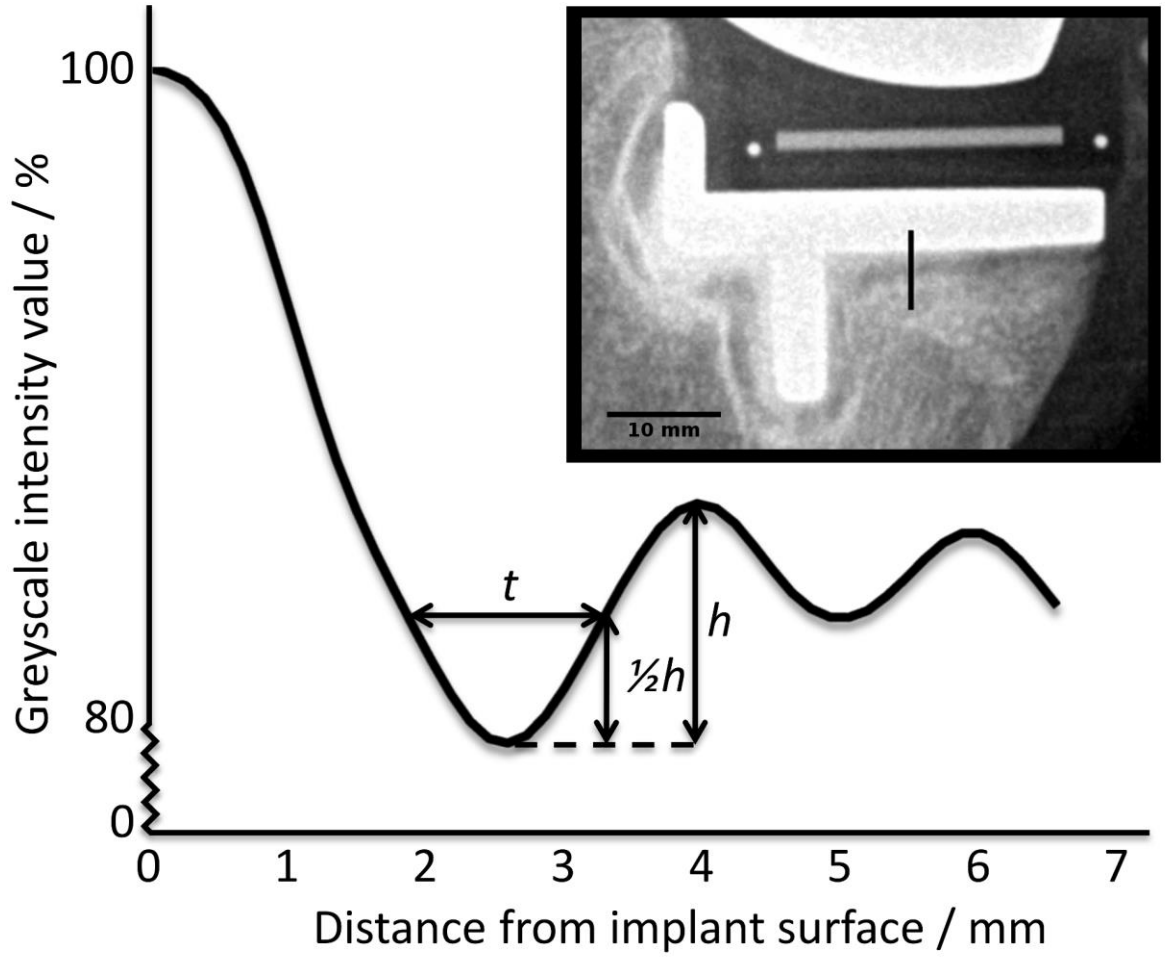

Figure 3

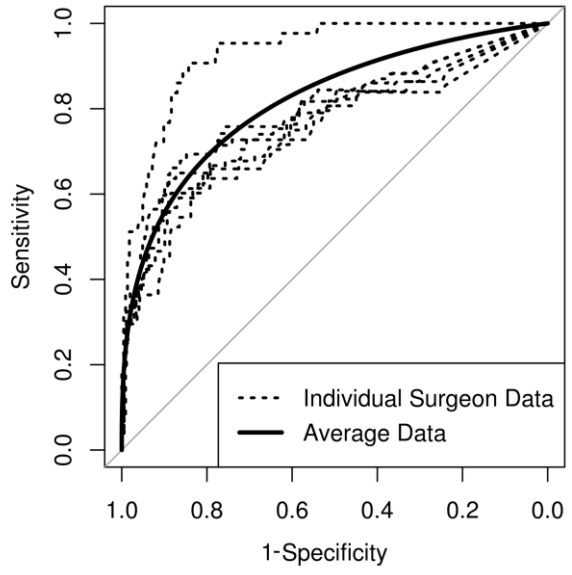

Figure 4

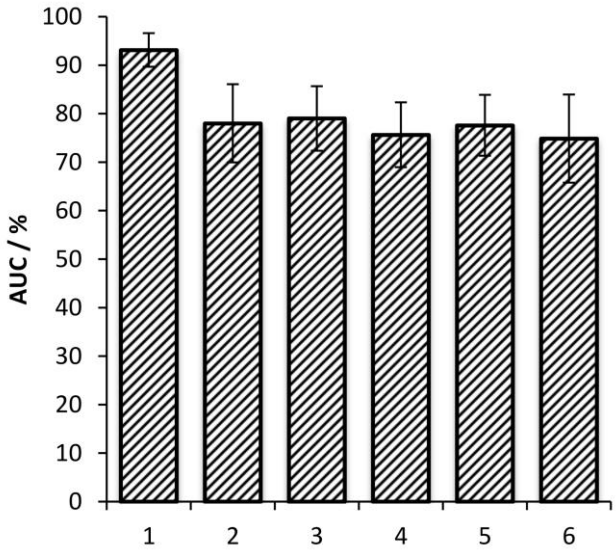

Figure 5

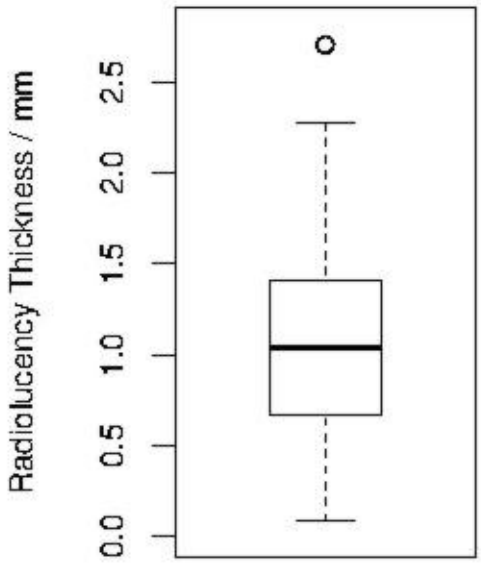

Figure 6

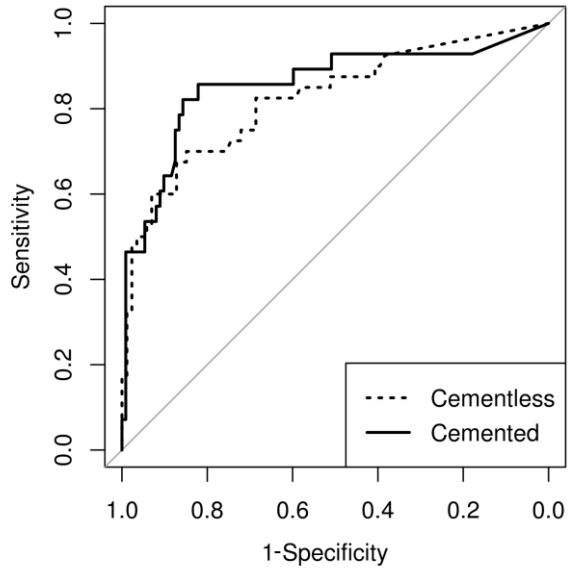
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.8
0.6
0.4
0.2
0.0
1-Specificity
Cementless
Cemented

Figure 7