The British Journal of Radiology, 79 (2006), S134–S140

The reliability of measuring physical characteristics of spiculated masses on mammography

[1]M P SAMPAT, PhD, [3]G J WHITMAN, MD, [3]T W STEPHENS, MD, [4]L D BROEMELING, PhD, [5]N A HEGER, PhD, [2]A C BOVIK, PhD and [1]M K MARKEY, PhD

Departments of [1]Biomedical Engineering and [2]Electrical and Computer Engineering, The University of Texas, Austin, TX 78712, [3]Division of Diagnostic Imaging and [4]Department of Biostatistics and Applied Mathematics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030 and [5]Department of Biological Sciences, Cameron University, Lawton, OK 73505, USA

ABSTRACT. The goal of this study was to assess the reliability of measurements of the physical characteristics of spiculated masses on mammography. The images used in this study were obtained from the Digital Database for Screening Mammography. Two experienced radiologists measured the properties of 21 images of spiculated masses. The length and width of all spicules and the major axis of the mass were measured. In addition, the observers counted the total number of spicules. Interobserver and intraobserver variability were evaluated using a hypothesis test for equivalence, the intraclass correlation coefficient (ICC) and Bland-Altman statistics. For an equivalence level of 30% of the mean of the senior radiologist's measurement, equivalence was achieved for the measurements of average spicule length ($p<0.01$), average spicule width ($p=0.03$), the length of the major axis ($p<0.01$) and for the count of the number of spicules ($p<0.01$). Similarly, with the ICC analysis technique ''excellent'' inter-rater agreement was observed for the measurements of average spicule length (ICC=0.770), the length of the major axis (ICC=0.801) and for the count of the number of spicules (ICC=0.780). ''Fair to good'' agreement was observed for the average spicule width (ICC=0.561). Equivalence was also demonstrated for intraobserver measurements. Physical properties of spiculated masses can be measured reliably on mammography. The interobserver and intraobserver variability for this task is comparable with that reported for other measurements made on medical images.

DOI: 10.1259/bjr/96723280

Computer-aided detection (CAD) systems have been developed to aid radiologists in interpreting mammograms [1–5]. However, it is widely acknowledged that current CAD systems detect microcalcifications more accurately than they detect masses, including spiculated masses. One reason for this is that calcifications are typically much denser than the surrounding tissue, whereas there is less contrast between masses and the parenchyma. Moreover, from an image processing perspective, calcifications are easier to detect because they can be simply modelled as impulse functions. In comparison, spiculated masses are difficult to model because of the great variability in their physical characteristics. The lack of statistical information on the physical properties of spiculated masses makes it difficult for engineers to create mathematical models of these abnormalities. For instance, there is no quantitative record of the physical characteristics of spiculated masses, such as the typical length of spicules. This information would be beneficial for the design of CAD algorithms (*e.g.* [6]), even though radiologists may not consciously use such information in detecting or characterizing lesions.

All radiological measurements are subject to interobserver and intraobserver variability. A number of statistical methods are available to quantify interobserver and intraobserver agreement. The Bland-Altman technique, intraclass correlation coefficient (ICC), kappa statistic and regression analysis are some of the most frequently used methods. While we are unaware of any studies that have focused on the reliability of measurements of mammographic lesions, several studies have assessed the observer variability of rating data, as opposed to measurement data, in mammographic interpretation. For example, considerable interobserver variability has been reported in describing mammographic masses using the BI-RADS^{TM} lexicon [7, 8]. By comparison, many studies have evaluated the interobserver and intraobserver variability of measurements in non-mammography medical imaging applications (*e.g.* [9–15]).

In this paper, we present the results of a study in which two experienced radiologists measured the parameters of spiculated masses on mammography. We demonstrate that the observer variability for this task is

Mehul P. Sampat supported by dissertation fellowship W81XWH-04-1-0406 from the Department of Defense (DoD) Congressionally Directed Medical Research Program (CDMRP). Seed grant funding from the U.T. Center for Biomedical Engineering.

Address correspondence to: Mia Markey, Department of Biomedial Engineering, The University of Texas at Austin, Austin, TX 78712, USA. E-mail: mia.markey@mail.utexas.edu

comparable with what has been reported in other medical imaging measurement studies.

Materials and methods

Data set

The images used in this measurement study were obtained from the Digital Database for Screening Mammography (DDSM), http://marathon.csee.usf.edu/Mammography/Database.html [16]. The DDSM is the largest publicly available data set of digitized mammograms. The entire database consists of 2620 cases and each case consists of four mammograms: a craniocaudal (CC) and mediolateral oblique (MLO) view of each breast. The mammograms were obtained from three institutions [16]. Along with the digitized mammograms, the DDSM contains ''boundary'' files of the abnormalities. The outlines of the abnormalities as indicated by a radiologist are stored in ''chain code'' in these files. From this ''chain code'', borders of the abnormalities can be reconstructed.

In this study, observers primarily worked with a region of interest (ROI) from each image, although the full mammogram was always available to them. The ROI was defined such that the central mass and all spicules were visible. In particular, the ROI was taken as the smallest rectangle in which the boundary specified in the DDSM database could be inscribed, plus 500 pixels in each direction.

For this study, the MLO views of cases of spiculated masses were randomly selected from the DDSM. Cases were selected from a single scanner, and we confirmed that a range of density ratings, subtlety ratings and pathology were represented by the sample. A set of 12 cases was used for observer training and measurements were collected using a second, distinct set of 21 cases. The characteristics of the measurement set are summarized in Table 1. A list of the DDSM cases numbers and the ROI images used in this study are available on our website: http://www.bme.utexas.edu/research/informatics/.

Observer training and measurement protocol

The physical measurements were made by two experienced radiologists. For the rest of the paper, we will refer to these radiologists as R1 and R2. Radiologist R1 was the senior radiologist and had more experience in breast imaging. R1 also trained as a breast-imaging fellow for 1 year and has been reading mammograms since 1990. Currently, radiologist R1 reads 7000 mammograms per year. Radiologist R2 was trained as a breast-imaging fellow for 1 year and has been reading mammograms since 1994. Currently, radiologist R2 reads 3000 mammograms per year.

Table 1. Properties of the two sets of images used in this study. A set of 12 cases was used for observer training and measurements were collected using a second, distinct set of 21 cases

	Total number of cases	No. of malignant cases	Minimum density	Mean density	Maximum density
Training set	12	10	1	2	3
Measurement set	21	21	1	2	3

We used the ROI Manager plugin of NIH ImageJ to enable the radiologists to measure physical properties of spiculated lesions on mammograms (Figure 1). Using a straight line tool, the radiologists marked the principal axes of the central mass, the width of each spicule at its base where it meets the mass, and traced each spicule in order to measure their length. Since the resolution of the images was known, the pixel measurements were converted into physically meaningful quantities (*e.g.* millimetres). In addition, the radiologists counted the spicules associated with each lesion. The measurements were made on ROIs, but the radiologists were allowed to view the full mammogram at any time and could adjust the display as desired (*e.g.* zoom). The images were displayed on a standard laptop computer in a darkened room and the radiologists were allowed unlimited time for the measurement task. All of the images with the radiologists' markings overlaid are available on our website, as described earlier.

Measuring spiculated lesions is not part of routine clinical practice. Thus, we conducted a training stage in which the radiologists discussed the results of measurements made independently on a training set of images. Each radiologist independently measured the properties of a training set of 12 spiculated masses. Their markings were overlaid on the original ROIs (Figure 2a) and they discussed areas of agreement and disagreement in their measurements.

Following the training phase, the two radiologists independently measured the properties of 21 images of spiculated masses. There was no overlap between the training and the measurement sets. Because of time and scheduling constraints, these measurements were carried out in two sessions. In the first session (a few weeks after the training session), the properties of 12 images were measured and in the second session (a few months after the training session) the properties of the remaining 9 images were analysed. To assess the intraobserver variability, one radiologist (R1) re-measured the first set of 12 images after an interval of 5 months. Thus, a total of 21 images were used for the analysis of the interobserver agreement and a set of 12 images was used to compute the intraobserver agreement.

Statistical analysis

We believe that it is important to assess the degree of agreement using multiple statistical methods. This view is also shared by Luiz et al [17] who noted that for the analysis of measurement studies it is desirable to report the degree of agreement using multiple statistical methods as no method is foolproof and each has its own limitations.

The degree of agreement between the measurements of radiologists R1 and R2 was evaluated using a hypothesis test for equivalence, the intraclass correlation (ICC) coefficient [18], and Bland-Altman statistics [19, 20]. In testing for equivalence, the null hypothesis is that

Table 2. Summary statistics and the results of the hypothesis test for equivalence between the measurements of radiologists R1 and R2. The null hypothesis was that the two radiologists are not equivalent. Thus, if we reject the null hypothesis, the measurements of the two radiologists are deemed equivalent

Total number of cases	Delta	Major axis	Spicule width	Spicule length	Number of spicules
		R1's mean=3.78	R1's mean=0.278	R1's mean=2.44	R1's mean=17.57
		R2's mean=3.73	R2's mean=0.221	R2's mean=2.39	R2's mean=18.48
21§	δ=0.30 * mean of R1's measurement	**$p<0.01$**	**$p=0.03$**	**$p<0.01$**	**$p<0.01$**
21§	δ=0.25 * mean of R1's measurement	**$p<0.01$**	$p=0.18$	**$p<0.01$**	**$p<0.01$**
21§	δ=0.20 * mean of R1's measurement	**$p<0.01$**	$p=0.54$	**$p<0.01$**	**$p=0.01$**

§One observer measured minor axis by mistake, so that image was removed for the major axis calculation only.

the measurements of the two radiologists are not equivalent and the alternative hypothesis is that they are equivalent [21]. Note that the more familiar paired *t*-test for a null hypothesis of equal values *vs* an alternative hypothesis of not equal values is not an appropriate test for establishing equivalence. Failing to reject a null hypothesis does not prove that the null hypothesis is correct; in particular, a failure to reject the null hypothesis can arise from a lack of power. Thus, a hypothesis test specifically intended for assessing equivalence was used.

The test statistic (t) for assessing equivalence is

$$t = \frac{\sqrt{n}(\bar{x} \pm \delta)}{s} \quad (1)$$

where $\bar{x}$ and s are the mean and standard deviation, respectively, of the differences between the measurements of the two readers. The value of δ is computed as a factor multiplied by the mean of the more experienced reader's measurements. In this study, the factor was 0.20, 0.25, or 0.30. The variable δ accounts for the expected variability in the measurements made by the two radiologists. A smaller value of δ implies stricter criteria for demonstrating that the measurements of the two radiologists are equivalent.

The ICC coefficient is also used to report the degree of agreement between multiple readers. A number of different models can be used for computing the ICC value [18]. In this study, to report the interobserver agreement, a two-way random model was used since the set of images is a random subset of images from the class of mammographic images and the radiologists are also randomly selected from the population of radiologists.

$$ICC = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)}, \quad (2)$$

where k denotes the number of readers, n denotes the number of images, MS_R is the mean square error between images, MS_E is the residual mean square error and MS_C is the mean square error between readers. For the computation of the intraobserver agreement, a two-way mixed model was used as the set of images is considered as a random subset of images from all mammographic images, but the measurements are made by a single radiologist and thus the rater (radiologist) is considered as a fixed effect in the ICC model [18].

Different guidelines exist for the interpretation of ICC, but one reasonable scale is that an ICC value of less than 0.40 indicates poor reproducibility, ICC values in the range 0.40 to 0.75 indicate fair to good reproducibility, and an ICC value of greater than 0.75 shows excellent reproducibility [22].

Bland-Altman analysis (also known as the method of differences) has been proposed for measuring the degree of agreement [19, 20]. In this method, the differences in the measurements made by two readers are plotted against the average values of these measurements. According to Bland and Altman [19, 20], if 95% of the differences are within ± 1.96 standard deviations of the mean of the differences, then this denotes good agreements between the two sets of measurements. These limits are also known as the ''limits of agreement''. Note that hypothesis testing for equivalence and the ICC method provide quantitative measures of the agreement between the measurements whereas the Bland-Altman analysis technique provides a qualitative assessment.

Results

The interobserver and intraobserver variability of measurements of spiculated masses was evaluated using a hypothesis test for equivalence, the ICC coefficient and the Bland-Altman technique. For an equivalence level of 30% of the mean of R1's first measurement (Table 2), equivalence was achieved between R1's and R2's measurements (N=21) for average spicule length ($p<0.01$), average spicule width ($p=0.03$) and the count of the number of spicules ($p<0.01$). For comparing the major axis measurements, one case was removed since

Table 3. Interobserver agreement. Intraclass correlation (ICC) coefficients for the measurements made by radiologists R1 and R2.

Total number of cases	Major axis	Spicule width	Spicule length	Number of spicules
21§	ICC=0.801	ICC=0.561	ICC=0.770	ICC=0.780

§One observer measured minor axis by mistake, so that image was removed for the major axis calculation only.

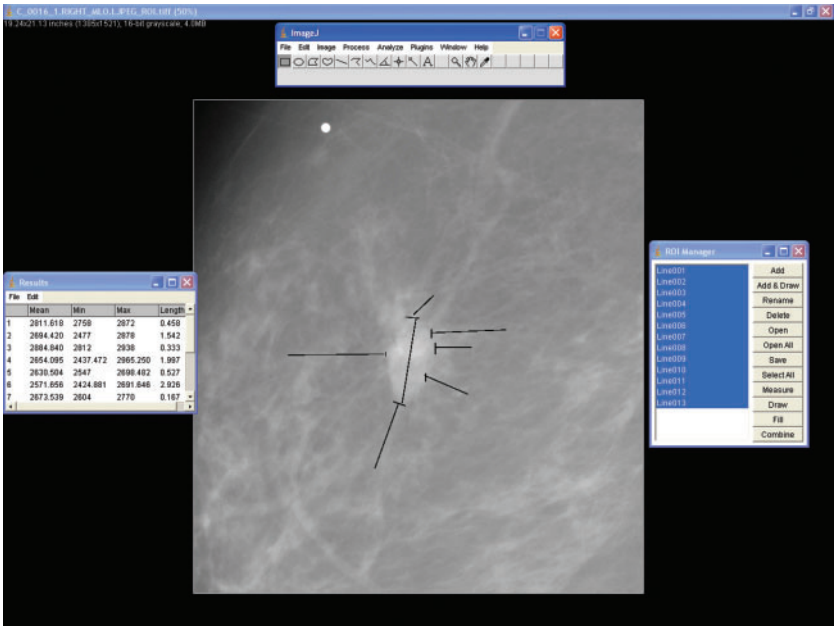
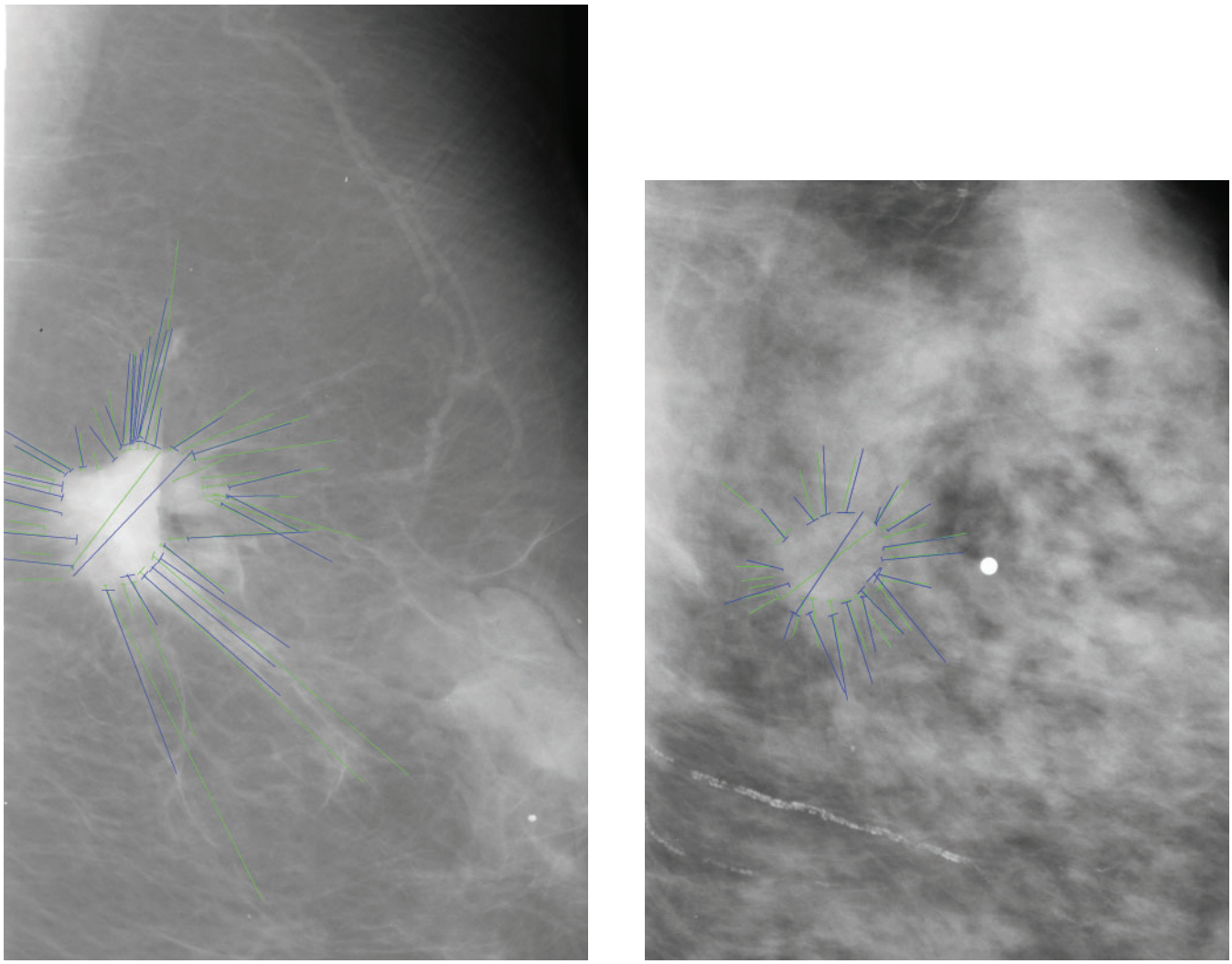


Figure 1. NIH ImageJ Interface for obtaining measurements of key characteristics of spiculated masses.



(a) (b)

Figure 2. Examples of the measurements made by radiologists R1 and R2 before and after the training stage. The measurements made by R1 are shown in blue and those made by R2 are shown in green. (a) Example measurements made by the two radiologists during the training phase. (b) Example measurements made by the two radiologists during the measurement phase.

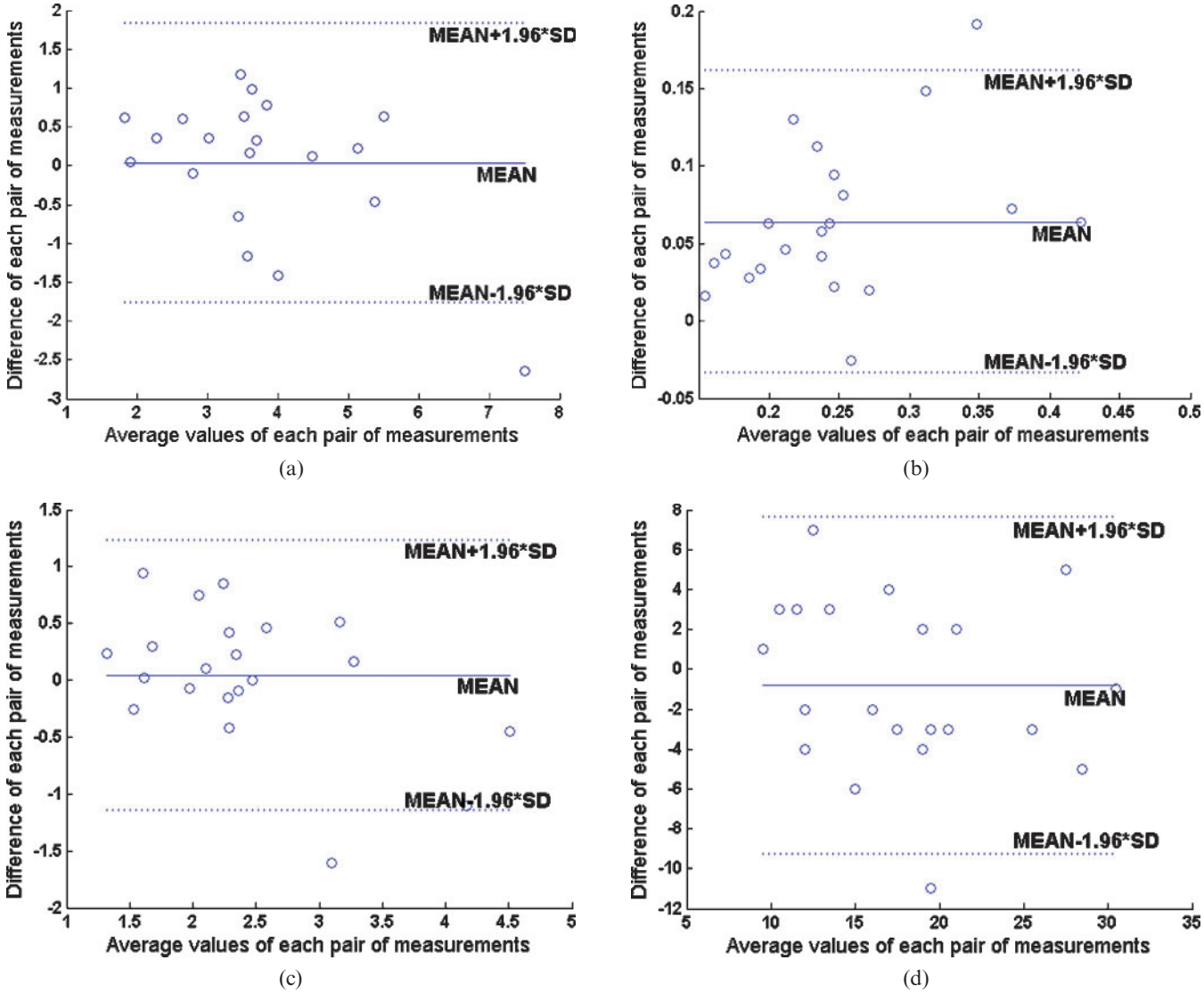


Figure 3. Bland-Altman analysis for the interobserver agreement for each of the four physical characteristics that were measured by radiologists R1 and R2. The parameters measured were: (a) major axis of the spiculated masses, (b) the width of the spiculations, (c) the length of the spiculations and (d) the number of spiculations.

R2 inadvertently measured the minor axis; with $N=20$, equivalence was achieved for the length of the major axis ($p<0.01$). Similarly, Table 3 shows the degree of agreement between the measurements of the radiologists R1 and R2 using the ICC method. Our analysis shows that there is ''excellent'' inter-rater agreement between R1's and R2's measurements ($N=21$) for average spicule length (ICC=0.770), and the count of the number of spicules (ICC=0.780). ''Fair to good agreement'' was obtained for the average spicule width (ICC=0.561). For comparing the major axis measurements, one case was removed and with $N=20$. ''Excellent'' inter-rater agreement was observed for the length of the major axis (ICC=0.801). The interobserver agreement was also analysed using the Bland-Altman technique. Bland and Altman suggested that if 95% of the differences were within the ''limits of agreement'' then this denoted good agreement between the two sets of measurements. According to the Bland-Altman method, good

Table 4. Results of the hypothesis test for equivalence between the first and second set of measurements made by radiologist R1. The null hypothesis was that the two sets of the measurements made by radiologist R1 are not equivalent. Thus, if we obtain a *p*-value of less than 0.05 (bold type), we can reject the null hypothesis and say that the two sets of measurements are equivalent.

Total number of cases	Delta	Major axis	Spicule width	Spicule length	Number of spicules
12	$\delta=0.30$ * mean of first set of measurements	**$p=0.00$**	**$p=0.00$**	**$p=0.00$**	**$p=0.00$**
12	$\delta=0.25$ * mean of first set of measurements	**$p=0.00$**	**$p=0.00$**	**$p=0.02$**	**$p=0.02$**
12	$\delta=0.20$ * mean of first set of measurements	**$p=0.01$**	**$p=0.00$**	$p=0.26$	$p=0.09$

Table 5. Intraobserver agreement. Intraclass correlation (ICC) coefficients for the two sets of measurements made by radiologist R1

Total number of cases	Major axis	Spicule width	Spicule length	Number of spicules
12	ICC=0.951	ICC=0.896	ICC=0.852	ICC=0.641

interobserver agreement was obtained for all four parameters measures (Figure 3).

We studied the intraobserver variability based on re-measurement of 12 images by the senior radiologist R1. For an equivalence level of 30%, equivalence was achieved between R1's first and second measurements (N=12) for all properties (Table 4): average spicule length ($p<0.01$), average spicule width ($p<0.01$), length of major axis ($p<0.01$) and the count of the number of spicules ($p=0.01$). Moreover, equivalence was demonstrated even at the stricter level of 25% of the mean of R1's first measurement. The intraobserver agreement between the two sets of measurements made by radiologist R1 (N=12) using the ICC method were also very good (Table 5). The intraobserver agreement was ''excellent'' for the length of the major axis (ICC=0.951), average spicule length (ICC=0.852), and the average spicule width (ICC=0.896). ''Fair to good agreement'' was observed for the count of the number of spicules (ICC=0.641). The intraobserver agreement was also analysed using the Bland-Altman technique. Bland and Altman suggested that if 95% of the differences were within the ''limits of agreement'' then this denoted good agreement between the two sets of measurements. Figure 4 show that for all of four parameters measured, good intra-observer agreement is obtained according to the Bland-Altman technique.

Discussion

In this paper, we have shown that it is feasible to make reliable measurements of the physical properties of

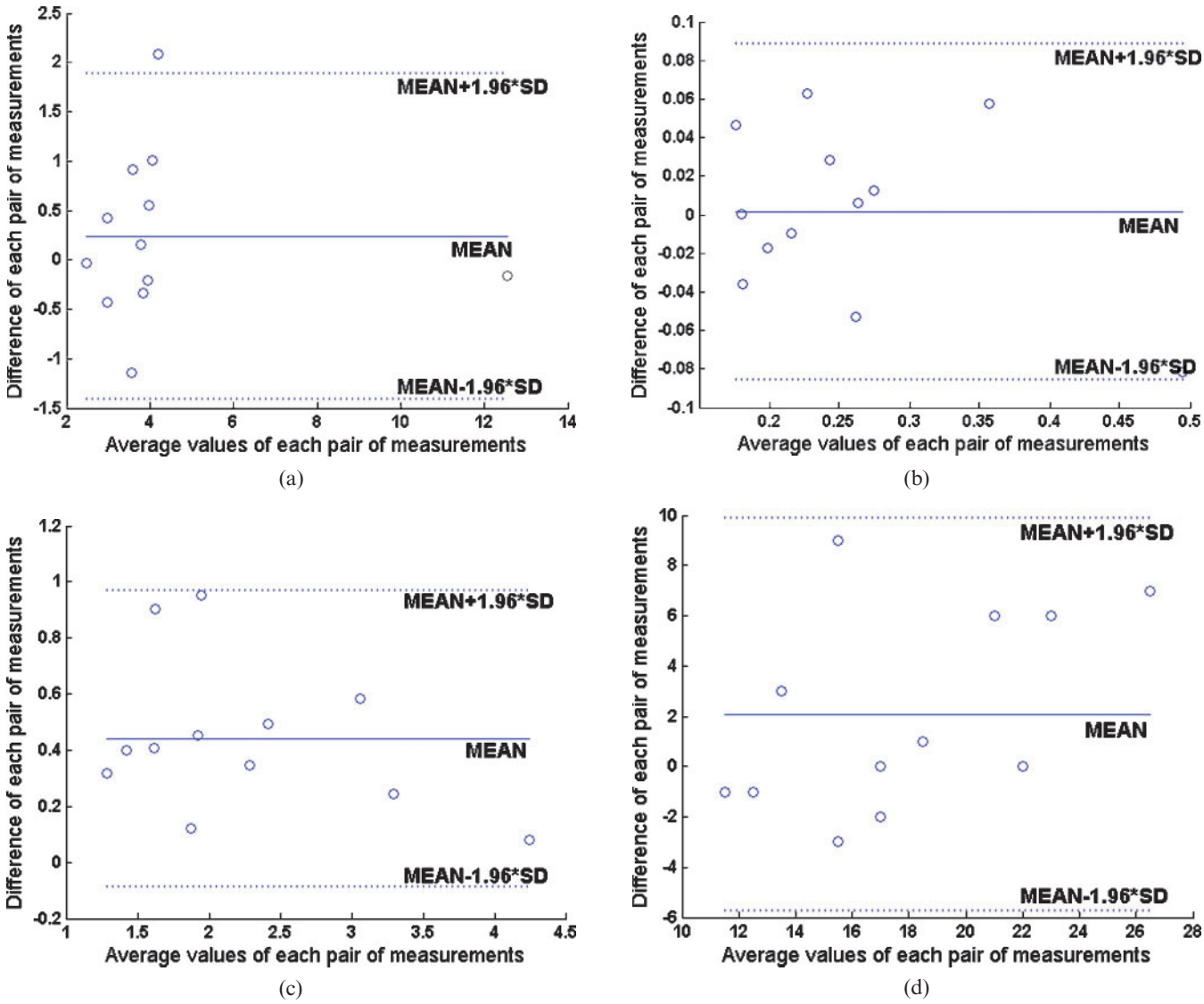


Figure 4. Bland-Altman analysis for the intraobserver agreement for each of the four physical characteristics that were measured twice by the senior radiologist R1. The parameters measured were (a) major axis of the spiculated masses, (b) the width of the spiculations, (c) the length of the spiculations and (d) the number of spiculations.

spiculated masses on mammography. The properties measured were the length and width of all spicules and length of the major axis of the central mass region. The count of the total number of spicules was also assessed.

We obtained good interobserver and intraobserver agreement in our study for the measurement of the properties of spiculated masses. We were able to demonstrate this with a hypothesis test for equivalence, the ICC and the Bland-Altman analysis. Since such a measurement task is not a part of the radiologists' regular clinical duties, the training stage was crucial for this measurement study. In the training phase, the radiologists discussed measurements that they had made independently (Figure 2a). While it was difficult for them to verbalize a consensus measurement protocol, the discussion was clearly fruitful since the data collected for the training process did not show equivalence (except for major axis), but equivalence was demonstrated for all four physical parameters in the measurement study after the training was complete.

Two interesting points are evident from a visual inspection of the marked images from the training and measurement phases of the study. First, some of the changes to their measurement protocol can be surmised; before the training, R2 typically marked spicules as being much longer than R1, but R2 marked the spicule lengths similarly to R1 after the training phase. Second, we noticed that if the two readers picked the same spicule, their measurements for that spicule were nearly identical. Thus, the primary source of variability appears to be the identification of structures as ''spicules'' rather than the task of measuring a spicule after it is located. Both of these points are observed in Figure 2, where the measurements made by the two radiologists are overlaid on the original image. Figure 2a shows the measurements made on an image during the training stage and Figure 2b shows the measurements on an image from the second set of spiculated masses.

To the best of our knowledge, no prior study has measured the physical properties of masses on mammograms or assessed the observer variability of such a task. However, researchers have reported the interobserver and intraobserver agreement for various measurement tasks in other areas of medical imaging (*e.g.* [9–15]). Although several statistical methods can be used to report the interobserver agreement, the most common approach has been to use the ICC. The ICC values reported in prior medical imaging measurement studies range from 0.570 to 0.820; thus, the ICC values observed this study (0.561 to 0.951) are within the range defined by previous work. Thus, we have demonstrated that properties of spiculated masses can be reliably measured on mammography, within the level of interobserver and intraobserver variability typical of other measurement tasks in radiology.

References

1. Giger ML. Computer-aided diagnosis of breast lesions in medical images. Comput Sci Engineering 2000;2:39–45.
2. Giger ML, Karssemeijer N, Armato SG, III. Computer-aided diagnosis in medical imaging. IEEE Transactions on Medical Imaging 2001;20:1205–8.
3. Doi K, MacMahon H, Katsuragawa S, Nishikawa RM, Jiang Y. Computer-aided diagnosis in radiology: potential and pitfalls. Eur J Radiol 1999;31:97–109.
4. Vyborny CJ, Giger ML, Nishikawa RM. Computer-aided detection and diagnosis of breast cancer. Radiol Clin N Am 2000;38:725–40.
5. Sampat MP, Markey MK, Bovik AC. Computer-aided detection and diagnosis in mammography. In: Bovik AC, editor. Handbook of image and video processing, 2nd edn. Academic Press, 2005:1195–217.
6. Sampat MP, Whitman GJ, Markey MK, Bovik AC. Evidence-based detection of spiculated masses and architectural distortions. Medical Imaging 2005: Image Processing 2005;5747:26–37.
7. Baker JA, Kornguth PJ, Floyd CE Jr. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. AJR Am J Roentgenol 1996;166:773–8.
8. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast imaging reporting and data system: inter- and intraobserver variability in feature analysis and final assessment. AJR Am J Roentgenol 2000;174:1769–77.
9. Vos MJ, Uitdehaag BM, Barkhof F, et al. Interobserver variability in the radiological assessment of response to chemotherapy in glioma. Neurology 2003;60:826–30.
10. Erasmus J, Gladish GW, Broemeling L, Sabloff BS, Truong MT, Herbst RS, et al. Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor responses. J Clin Oncol 2003;21:2574–82.
11. Bogot NR, Kazerooni EA, Kelly AM, Quint LE, Desjardins B, Nan B. Interobserver and intraobserver variability in the assessment of pulmonary nodule size on CT using film and computer display methods. Acad Radiol 2005;12:948–56.
12. Valentin L, Bergelin I. Intra- and interobserver reproducibility of ultrasound measurements of cervical length and width in the second and third trimesters of pregnancy. Ultrasound Obstet Gynecol 2002;20:256–62.
13. de Vries M, de Koning PJ, de Haan MW, Kesselo AG, Nelemans PJ, Nijenhuis RJ, et al. Accuracy of semiautomated analysis of 3D contrast-enhanced magnetic resonance angiography for detection and quantification of aortoiliac stenoses. Investigative Radiol 2005;40:495–503.
14. Wetzel SG, Cha S, Johnson G, Lee P, Law M, Kasow DL, et al. Relative cerebral blood volume measurements in intracranial mass lesions: interobserver and intraobserver reproducibility study. Radiology 2002;224:797–803.
15. Hopper KD, Kasales CJ, Van Slyke MA, Schwartz TA, TenHave TR, Jozefiak JA. Analysis of interobserver and intraobserver variability in CT tumor measurements. AJR Am J Roentgenol 1996;167:851–4.
16. Heath M, Bowyer KW, Kopans D, Moore R, Kegelmeyer P Jr. The digital database for screening mammography. In: Proceedings of the 5th International Workshop on Digital Mammography. Madison, WI: Medical Physics Publishing, 2000.
17. Luiz RR, Szklo M. More than one statistical strategy to assess agreement of quantitative measurements may usefully be reported. J Clin Epidemiol 2005;58:215–6.
18. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1996;1:30–46.
19. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–10.
20. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol 2003;22:85–93.
21. Wellek S. Testing statistical hypothesis of equivalence. Boca Raton, FL: CRC Press LLC, 2003.
22. Rosner B. Fundamentals of biostatistics. Belmont, CA. Duxbury Press, 2005.