Computer-Aided Evaluation of Radiologist's Reproducibility and Subjectivity in Mammographic Density Assessment

Ilijan Tomaš[1], Zdenka Kotoromanović[1], Nenad Belaj[1], Damir Margaretić[1], Zdravko Ivezić[2], Miroslav Katić[3], Lada Zibar[4], Dario Faj[1], Damir Štimac[2] and Mate Matić[1]

[1] University »Josip Juraj Strossmayer«, University Hospital Osijek, Department of Oncology and Radiotherapy, Osijek, Croatia
[2] University »Josip Juraj Strossmayer«, University Hospital Osijek, Department of Radiology, Osijek, Croatia
[3] University »Josip Juraj Strossmayer«, Department of Physics, Osijek, Croatia
[4] University »Josip Juraj Strossmayer«, University Hospital Osijek, Clinic of Internal Medicine, Osijek, Croatia

ABSTRACT

Mammographic density is an independent risk of breast cancer. This study has evaluated the radiologists' reproducibility and subjectivity in breast density estimation and in order to decrease the radiologists' subjective errors the computer software was developed. The very good reproducibility existed in the strong correlation with the first and the second mammogram assessment after three month period for each radiologist (correlation coefficient 0.73–1, $p<0.001$). The strong correlation was present in the case of all 5 radiologists when compared among themselves and compared with software aided MDEST-Mammographic Density Estimation (correlation coefficient 0.651–0.777, $p<0.001$). Detected differences in glandular tissue percentage determination occurred in the case of two experienced radiologists, out of 5 (one radiologist with more than 5 year experience and one with more than 10 year experience, $p<0.01$), but in the case of breast type determination (American College of Radiology-ACR I-IV), the detected difference occurred in one radiologist with the least experience (less than 5 years, $p<0.001$). It can be concluded that the estimation of glandular tissue percentage in breast density is rather subjective method, especially if it is expressed with absolute percentage, but the determination of type of breast (ARCI-IV) depends on the radiologist's experience. This study showed that software aided determination of glandular tissue percentage and breast type can be of a great benefit in the case of less experienced radiologists.

Key words: *mammography, breast density, subjectivity, software*

Introduction

Mammography is the most important diagnostic method used in the early breast cancer detection. Mammographic density which is determined by having mammogram X-rays read by trained radiologists is an independent risk factor of breast cancer. The increased radiological density has been reported to be associated with up to 4–6 time increase in the risk of breast cancer[2,3]. Breast density is closely related to tumour size, lymph node involvement, and lymphatic or vascular invasion in screen-detected cancers[4]. The computerized image analysis tool (software program) can provide a consistent and reproducible estimation of dense area percentage on the routine clinical mammograms, collected in screening, so this analysis contributes to understanding of the relationship between mammographic density and breast cancer risk, detection and prognosis, prevention and treatment of breast cancer[5,6]. It turned out when estimating the density percentage and the type of breast, besides the subjective error, that there was a difference between digital mammogram assessment and the assessment of the mammogram film which was later digitized[7].

In this study we examined reproducibility and subjectivity of 5 radiologists in MDEST (Mammographic Density Estimation), who had different experience. The radiologists visually estimated the glandular tissue percentage and the type of breast (American College of Radiology-ACR I-IV). In addition we developed software in or-

Received for publication May 1, 2012

der to reduce the subjective errors in breast density assessment. The estimations of radiologists were compared to the computer-aided assessment.

Material and Methods

This study was conducted at the Oncology and Radiotherapy Department and at Radiology Department, Clinical Hospital Osijek. The standard mammograms were made using MAMOMAT 1000 by Siemens, on the x-ray FERANIA film. We selected 240 mammograms out of mammogram archive, which did not have any focal abnormalities, but showed different breast density due to the different percentage of fibro glandular tissue. The mammograms were coded with the letters and the numbers, from A1-A9, B1-B9... all the way till Y9. Five radiologists were asked to evaluate breast density, writing in a special designed table the breast type (I–IV), put down the percentage of fibro glandular tissue next to the coded mammograms. Two out of 5 had more than 10 year working experience (radiologists C and D), two have more than 5 years of experience (radiologists B and E) and one had less than 5 years (radiologist A). The breast density was classified by each radiologist following four point ACR scale: 1. almost entirely fat, 2. scattered fibro glandular densities, 3. heterogeneously dense and 4. extremely dense[8].

The radiologists were asked to determine the percentage of glandular tissue and note it on the visual scale from 1–100. Every radiologist assessed every mammogram out of 240 twice, after three month period, putting data in the table. The radiologists assessed mammograms alone, without time limit. The obtained data helped us to establish criteria used by 5 radiologists for breast density estimation, expressed in the percentage and estimation of breast type. We evaluated the radiologists' reproducibility for each of them, comparing their first mammogram and their second mammogram assessment. The software program used to determine breast density on digital mammograms can be seen on www. Mamography analyzer (May 2007), language: Croatian, technologies: COM, Microsoft Visual Studio 6.0.

Mammograms were digitized, images were displayed on the computer screen and the thresholds were set by an observer, defining the edge of the breast and the edge of dense breast tissue. The areas defined were then measured by the computer and the percentage area of the image occupied by the dense tissue was calculated. Web mammograms readings are organized into database on PC. The statistical analysis was performed using software package SPSS 17.0 (SPSS inc. Chicago, IL, USA) the data were expressed in numeric (percentage) or ordinal categories (types). Normal data distribution was considered in the case of skewness less than 1. Depending on how normal distribution was, the descriptive statistics used $\overline{X}\pm SD$ and median rage to present data. Paired t-test was used for comparison analysis of the same mammogram assessment (percentage, numeric variable, normal distribution), done by the same radiologist on two separate occasions. If the distribution differed from normal, the data was compared using Wilcoxon signed rank test. Friedman's test was used to compare the assessments repeated for the same mammogram by 5 radiologists, whereas Wilcoxon signed rank test was used for post hoc analysis. The correlations with variables were examined using Pearson's correlation test for normally distributed data and in the case of data distributed different from normal, Kendall tau correlation coefficient was used. $P<0.05$ was considered statistically significant.

Results

The table 1 shows the breast density estimation, done by the same radiologist and expressed in glandular tissue percentage and breast type (ACR I-IV) for the same mammogram on two different occasions within 3 month interval. Out of 5 radiologists, included in the study, two radiologists had the statistically significant difference between their first and their second mammogram assessment, when determining glandular tissue percentage, whereas other 3 radiologists assessing mammograms had the same finding. The statistically significant difference occurred in the case of radiologist B and radiologist C ($p<0.001$). When the same radiologist was asked to determine the type of breast (I-IV), assessing mammogram twice within the 3 month interval, we could see the statistically significant difference between these two mammogram assessments occurred only in the case of the ra-

TABLE 1
DIFFERENCE BETWEEN 1st AND 2nd MAMMOGRAM ASSESSMENT FOR EACH RADIOLOGIST, ESTIMATING PERCENTAGE OF GLANDULAR TISSUE AND TYPE OF BREAST BEFORE AND AFTER THREE MONTH PERIOD

Radiologist	Percentage of glandular breast tissue			Type of breasts		
	1st assessment	2nd assessment	p	1st assessment	2nd assessment	p
A	15 (5–85)*	25 (5–75)	0.072	2±1**	2±1	<0.001
B	21.87±16.07*	20 (5–85)	<0.001	1 (1–4)*	1 (1–4)	0.174
C	25 (10–80)*	32.47±13.34	<0.001	2±1**	2±1	0.669
D	15 (0–90)*	15 (0–90)	0.236	1 (1–4)*	1 (1–4)	1.000
E	36.07±19.69**	35.09±21.45	0.104	2±1**	2±1	0.152

* Wilcoxon signed rank test, ** Paired t-test

TABLE 2
DIFFERENCE BETWEEN 1ST AND 2ND ASSESSMENT OF THE PERCENTAGE OF GLANDULAR TISSUE AND TYPE OF BREAST AMONG EACH OF THE FIVE RADIOLOGISTS

Radiologists	Percentage of glandular breast tissue				Type of breasts			
	Test value of 1st assessment	p	Test value of 2nd assessment	p	Test value of 1st assessment	p	Test value of 2nd assessment	p
A and B	z=–1.131	0.258	z=–5.165	<0.001	z=–9.827	<0.001	z=–8.778	<0.001
A and C	z=–4.096	<0.001	z=–10.226	<0.001	t=8.8896	<0.001	z=–4.919	<0.001
A and D	z=–1.790	0.073	z=–0.862	0.388	z=–8.725	<0.001	z=–7.013	<0.001
A and E	z=–11.918	<0.001	z=–10.608	<0.001	t=0.0000	1.000	z=–2.750	0.006
B and C	z=–4.405	<0.001	z=–8.331	<0.001	z=–2.490	0.013	z=–4.355	<0.001
B and D	z=–2.733	0.006	z=–3.980	<0.001	z=–0.590	0.555	z=–1.474	0.142
B and E	z=–12.754	<0.001	z=–9.547	<0.001	z=–9.680	<0.001	z=–9.068	<0.001
C and D	z=–2.120	0.034	z=–8.375	<0.001	z=–1.827	0.068	z=–2.642	0.008
C and E	z=–9.392	<0.001	z=–2.672	0.008	t=–9.165	<0.001	z=–6.465	<0.001
D and E	z=–11.772	<0.001	z=–11.266	<0.001	z=–9.610	<0.001	z=–8.361	<0.001

z – Wilcoxon signed rank test, t – Paired t-test

diologist A ($p<0.001$) but all other four radiologists had the matching assessments.

The table 2 shows statistically significant difference among all 5 radiologists, when asked to assess mammograms for the first time and then for the second time. The breast density was expressed as percentage ($\chi^2=312.771$, $p=0.001$ Friedman's test). This statistically significant difference for the first mammogram assessment was evident using post hoc analysis between radiologist A, and C, E, the difference occurred between radiologist B and C, E, as well as between radiologist D and E. The statistically significant difference was evident in the case of the second mammogram assessment and breast density was expressed in percentage ($\chi^2=212.324$, $p<0.001$ Friedman's test). This difference was detected using post hoc analysis among the radiologists A and B, C, E, as well as among B and C, D, E, then among C and D, E, and between D and E.

In the case of breast density expressed as breast type (I-IV), there was a statistically significant difference among all 5 radiologists when assessing mammograms for the first time ($\chi^2=211.346$, $p=0.001$ Friedman's test). This statistically significant difference became evident using post hoc analysis and occurred for the first assessment among radiologists A and B, C, D, then among B and C, D, as well as between C and E and between D and E radiologists ($p<0.001$). The statically significant difference existed among all 5 radiologists, assessing the mammograms for the second time and determining the breast types ($\chi^2=167.045$, $p<0.001$, Friedman's test). This difference, spotted by post hoc analysis was present between radiologist A and group of the radiologists B, C, D, E and between radiologist B and C, E and between radiologists D and E.

Since the study confirmed the radiologists' subjectivity in breast density assessment, we developed a software package to assess mammogram breast density more objectively. Using software program we compared the radiologists' breast density estimation expressed in percentage and noticed that there was a statistically significant difference between breast density assessment done by the radiologists and software program. In the case of the first mammogram assessment it was $p<0.001$ for the radiologists B, C, E and in the same group of radiologists the difference was $p<0.001$ for the second mammogram assessment when compared to the software aided assessment.

If we compare the radiologists' assessment of the types of breast (ACR), sorted out in four categories with software aided assessment of the breast types, the statistically significant difference occured for the first assessment in the group of radiologists A, C, E and for the second assessment among radiologists no. A, C, E, but in the case of radiologists B and D there was no statistically significant difference between their assessment and software aided ones, either for the first or second mammogram assessment (Table 3).

Once when we spotted the differences in mammogram assessments between radiologists and software aided ones, we determined to what extent the radiologists' first and second assessments were matching. We did it for each radiologist individually, then to what extent they matched each other assessments and at last how close their assessment was to software aided ones. The matching was expressed with correlation factor. There was a statistically significant strong correlation with both mammogram assessments for type of breast determination among all five radiologists ($p<0.001$, table 4).

The table 5 shows to what extent individual radiologist's assessments matched after the first and second assessments and whether they matched the software aided ones. The strong correlation was found with the first and the second mammogram assessment for the breast type for each radiologist ($p<0.001$). There was also a statistically strong correlation with the radiologists' first and

TABLE 3
DIFFERENCE BETWEEN RADIOLOGIST'S ESTIMATION OF PERCENTAGE OF GLANDULAR BREAST TISSUE AND TYPE OF BREAST AND SOFTWARE FOR 1ST AND 2ND MAMMOGRAM ASSESSMENT

Radiologist and software	Percentage of gladular breast tissue				Type of breasts			
	Test value of 1st reading	p	Test value of 2nd reading	p	Test value of 1st reading	p	Test value of 2nd reading	p
A-software	z=–0.043	0.966	z=–1.221	0.222	t=11.127	<0.001	t=8.072	<0.001
B-software	t=–2.523	0.012	z=–5.219	<0.001	z=–0.258	0.796	z=–1.236	0.216
C-software	z=–2.753	0.006	t=11.778	<0.001	t=2.184	0.032	t=2.853	0.005
D-software	z=–1.247	0.212	z=–0.759	0.448	z=–0.525	0.643	z=–0.262	0.793
E-software	t=20.007	<0.001	t=17.595	<0.001	t=12.130	<0.001	t=10.873	<0.001

z – Wilcoxon signed rank test, t – Paired t-test

TABLE 4
CORRELATION BETWEEN THE 1ST AND 2ND ASSESSMENT OF THE TYPE OF BREAST AMONG FIVE RADIOLOGISTS

Radiologists	Correlation coefficient of 1st reading	p	Correlation coefficient of 2nd reading	p
A and B	τ=0.671	<0.001	τ=0.709	<0.001
A and C	r=0.666	<0.001	r=0.678	<0.001
A and D	τ=0.638	<0.001	τ=0.671	<0.001
A and E	r=0.761	<0.001	r=0.792	<0.001
B and C	τ=0.668	<0.001	τ=0.679	<0.001
B and D	τ=0.812	<0.001	τ=0.719	<0.001
B and E	τ=0.732	<0.001	τ=0.719	<0.001
C and D	r=0.748	<0.001	τ=0.596	<0.001
C and E	τ=0.653	<0.001	r=0.709	<0.001
D and E	τ=0.731	<0.001	τ=0.687	<0.001

τ – Kendall tau correlation coefficient, r – Pearson's correlation test

TABLE 5
CORRELATION WITH THE 1ST AND 2ND ASSESSMENT IN DETERMINING THE TYPE OF BREAST FOR EVERY RADIOLOGIST AND EVERY RADIOLOGISTS' CORRELATION WITH SOFTWARE

Radiologist	Correlation for radiologist		Correlation radiologist with software			
	Correlation coefficient	p	Correlation coefficient of 1st reading	p	Correlation coefficient of 2nd reading	p
A	r=0.802	<0.001	r=0.651	<0.001	r=0.689	<0.001
B	τ=0.819	<0.001	τ=0.687	<0.001	τ=0.720	<0.001
C	r=0.731	<0.001	r=0.654	<0.001	r=0.660	<0.001
D	τ=1.000	<0.001	τ=0.739	<0.001	τ=0.732	<0.001
E	r=0.886	<0.001	r=0.777	<0.001	r=0.780	<0.001

τ – Kendall tau correlation coefficient, r – Pearson's correlation test

the second mammogram assessment (for all 5) of breast types when compared to software aided one ($p<0.001$).

Discussion and Conclusion

The study showed that three radiologists had matching results when asked to assess percentage of breast density after the first and the second mammogram assessment (A, D, E). One radiologist had least working experience, two more than 5 years. It has to be pointed out that the two radiologists (with more than 5 year working experience) did not have matching assessments (radiologists B more than 5 and C more than 10 years). The matching breast density assessments determining breast types (ACR I-IV), occurred in the group of 4 radi-

ologists (B, C, D, E) after their first and the second mammogram assessment, but in the case of the radiologist with the least working experience (A, less than 5 years) mammogram assessments differed (Table 1).

Examining the first and the second mammogram assessments of breast density percentage and breast types, it is obvious that the statistically significant differences exist among all radiologists and that there are quite many of them (Table 2).

This study showed that the percentage estimation did not depend on radiologists' mammogram assessment experience, but is more likely related to human ability to determine the percentage, whereas in the case of breast type estimation the radiologist's experience is of a great importance.

After strong subjectivity presence among radiologists who were asked to assess mammograms, the software program for breast density estimation got developed. The software expresses mathematically calculated areas of fibro glandular tissue or fat tissue. If we take software aided estimation as a point of reference for breast density estimation done by three radiologists (B, C and E), we can see that their first and second mammogram assessments did not match the software ones. In the case of software breast type estimation, taken as a point of reference, three radiologists' assessments (radiologist B, C, E) did not match the software assessments, meaning that the radiologist working experience in comparison to software assessment was irrelevant (Table 3). The difference between software aided assessment and radiologists' occurred due to the fact that the fibro glandular tissue is mostly in the middle, surrounded by the fat tissue, so using mathematical calculation fat tissue could have a bigger surface, although visually the fat tissue edge around glandular tissue is smaller and it seems to have far smaller surface, what is more visible in bigger breast.

This study differs from the others, because we tested the reproducibility of each radiologist and then the mammogram assessment accuracy between radiologists and software program, whereas in the other conducted studies was tested the percentage accuracy of the mammogram assessment between radiologists and software-aided assessment, but without radiologists' reproducibility. However, the radiologists were taken as a point of reference in comparison to software program[6,9]. The reason why in our study the correlation was lower than in the other studies can be explained with the fact that in the other studies the radiologists assessed CC and ML projection and could analyze density more thoroughly due to two projections[10]. In our study every mammogram was coded and density estimation was based on one projection, regardless of the second projection and since the mammograms were coded, the density estimation could not be related to another breast projection. The other studies expressed the percentages in ten figures 10, 20, 30%, whereas in our study the radiologists expressed the density in units, not only in ten figures[9].

There was a strong correlation with the first and 3 months later second assessment for each radiologist (correlation coefficient 0.731–1, $p<0.001$). All 5 radiologists highly correlated with themselves and in relation to software program (correlation coefficient 0.651–0.777, $p<0.001$). The strong correlation showed that all radiologists included in the study, had matching breast density estimations, but their estimations also matched software aided ones; however, the radiologists used other initial criteria. The differences were evident in the case of 2 experienced radiologists (one more than 5 years of experience and one more than 10, $p<0.001$) when estimating the percentage of glandular tissue, but in the case of breast type determination (ACR I-IV), the difference occurred in one radiologist with the least working experience (less than 5 years, $p<0.001$). The conclusion of this study is that the estimation of glandular tissue percentage in breast density is a subjective method, especially when we use absolute percentage, but the breast type determination (ACR I-IV) depends on the radiologists' working experience. The study showed that the software program could be of a great help in the estimation of glandular tissue percentage and breast type determination in case of less experience radiologists.

REFERENCES

1. OZA AM, BOYD NF, Epidemiol rev, 15 (1993) 196. — 2. BYRNE C, J Natl Cancer Inst, 89 (1997) 531. — 3. BOYD NF, LOCKWOOD GA, BYNG JW, TRITCHLER DL, YAFFE MJ, Cancer Epidemiol Biomarkers Prev, 12 (1998) 1133. — 4. AIELLO EJ, BUIST SMD, WHITE E, PORTER LP, Cancer Epidemiol Biomarkers Prev, 14 (2005) 662. — 5. PRIBIĆ S, GMAJNIĆ R, MAJNARIĆ-TRTICA L, EBLING B, VRANJEŠ Z, Coll Antropol, 34 (2010) 871. — 6. MARTIN KE, HELVIE MA, ZHOU C, ROUBIDOUX MA, BAILEY JE, PARAMAGUL C, BLANE CE, KLEIN KA, SONNAD SS, CHAN H-P, Radiology, 240 (2006) 656. — 7. CHAN HP, Automated Method for Analysis of Mammographic Breast Density – A Technique for Breast Cancer Risk Estimation. Annual report. (Michigan univ Ann Arbor, Michigan, 2004) — 8. RESTON VA, Breast imaging reporting and data system (BI-RADS), American College of Radiology, 1998. — 9. ZHOU C, CHAN HP, PETRICK N, HELVIE MA, GOODSITT MM, SAHINER B, HADJIISKI LM, ROUBIDOUX MA, Med Phys 28 (2001) 1056. — 10. YAFFE JM, Breast Cancer Research, 10 (2008) 209.

I. Tomaš

University »Josip Juraj Strossmayer«, University Hospital Center Osijek, Department of Oncology and Radiotherapy, Huttlerova 4, 31000 Osijek, Croatia
e-mail: ilijant@yahoo.com

PROCJENA REPRODUCIBILNOSTI I SUBJEKTIVNOSTI RADILOGA PRI ODREĐIVANJU DENZITETA DOJKE NA MAMOGRAFIJAMA U KOMPARACIJI SA SOFTVEROM

S A Ž E T A K

Denzitet mamografije je neovisan faktor rizika karcinoma dojke. U ovom istraživanju smo ispitivali subjektivnost i reproducibilnost radiologa pri određivanju gustoće tkiva dojke te smo stoga razvili softver koji bi pri tome smanjio subjektivne greške. Jako dobra reproducibilnost se pokazala kroz jaku korelaciju između 1. i 2. očitanja u razmaku od 3 mjeseca za svakog radiologa pojedinačno (correlation coefficient 0,731–1, $p<0,001$). Također je jaka korelacija nađena i za svih pet radiologa međusobno i u odnosu na software (correlation coefficient 0,651–0,777, $p<0,001$). Razlike vidljive u procjeni postotka nađene su kod dva iskusna radiologa (jedan više od 5 g. iskustva i jedan više od 10 g. iskustva, $p<0.001$) od ukupno 5 radilologa u određivnju postotka žljezdanog tkiva dok je razlika u određivanu tipa dojke (American College of Radiology-ARC I-IV) nađena u očitanju radiologa s najmanje iskustva (manje od 5 g, $p<0,001$). Zaključak ove studije bi bio da je određivanje denziteta dojke u postotku žljezdanog tkiva dojke subjektivna metoda osobito kada se izražava apsolutnim postotkom, a da procjena tipa dojke (ARC I-IV) ovisi o iskustvu radiologa. Pokazalo se također da softver može pomoći u procjeni postotka žljezdanog tkiva i tipa dojke osobito kod manje iskusnih radiologa.