Pitfall in Assessing the Size of Tumor Phantoms on Mammograms

CARLOS A. RUBIO[1], GUNILLA SVANE[2], GABRIELA ILESCU[2], BÖRKUR ADALSTEINSSON[2], TOOMAS MATHIESEN[2], MARIA THOLIN[2], MINORU MACHIDA[2], EUGENIA COLON CERVANTES[1], LARS MATTSSON[3] and EDWARD AZAVEDO[2]

[1]Departments of Pathology, and [2]Mammography, Karolinska University Hospital and Institute, Stockholm, Sweden; [3]Department of Production Engineering, The Royal Institute of Technology, Stockholm, Sweden

Abstract. *Background: Tumor size is crucial for clinical management and prognosis of breast malignancies. Materials and Methods: The gold standard-size of 12 tumor phantoms was assessed at The Department of Production Engineering. Subsequently, with a conventional ruler, seven experienced mammographers measured the largest diameter of the 12 devices in two independent trials. Results: In the first trial, 30% (n=25) of the 84 values given by the seven mammographers failed to recreate the gold standard size by >1 mm and in the second, by 37% (31/84). Size was overestimated (>1 mm) in 9.5% (n=8) of 84 measurements in the first trial, and in 15.5% (14/84) in the second. Conversely, size was underestimated (>1 mm) in 20% (n=17) of 84 measurements in the first trial, and in 21% (18/84) in the second. Neither the age of the participants, nor their years of experience improved the obtained results. Discussion: The method used here raised doubts concerning the ability of discriminating size among subgroups of T1 breast tumors in mammograms. According to the TNM staging system, T1 tumors (≤2.0 cm in greatest dimension) are subdivided into T1mic: microinvasion (≤0.1 cm), T1a (>0.1 cm but not more than 0.5 cm), T1b (>0.5 cm but not more than 1.0 cm) and T1c (>1.0 cm but not more than 2.0 cm in their greatest dimension). Since the TNM staging system for breast tumors is important in therapeutic decision making, it is crucial to develop a more reliable method for tumor size assessment.*

Breast cancer is the most frequent cancer in Sweden, affecting 7,000 females yearly; it accounts for 30% of all female malignancies (1). The preliminary diagnosis is made by anamnesis, inspection, palpation, mammography, ultrasound, magnetic resonance imaging (MRI), and is confirmed by invasive methods such as aspiration cytology, core biopsies or surgical biopsies (2-5).

The worldwide-accepted TNM staging system (6, 7), takes into account tumor size and lymph node and distant metastasis. The size of the primary breast tumor is crucial in planning therapeutic strategies for tumor cure. Tumors measuring no more than 2 cm across (with/without lymph node metastasis) are classified as T1 tumors. T2 tumors are those measuring more than 2 cm, but no more than 5 cm across. When of size more than 5 cm, breast tumors are classified as T3 (6, 7).

To measure for tumor size, several methods have been applied, including physical palpation, mammography, ultrasound, MRI and positron emission mammography (PEM) (2-5, 8-18). In one survey, the tumor size was recorded either by a pathologist on histological sections, by a surgeon on resected material, by a radiologist on x-ray mammography or by a clinician following clinical palpation (2). However, it has been demonstrated that several different diagnostic methods have different accuracies in tumor size assessment. In some subgroups of patients, the over- and underestimation can be even greater than 1 cm (12, 14).

In previous studies on tumor size, 12 tumor phantoms were carefully measured at the Department of Production Engineering, The Royal Institute of Technology, Stockholm, Sweden (19). Once the gold standard was established, 18 senior pathologists and 4 senior surgeons were asked to measure the 12 tumor phantoms. Results showed disparate inter- and intra-observer variations in size assessment in two independent trials (19). In a second test, seven senior colonoscopists were asked to measure the 12 phantoms in tandem colonoscopic examinations performed in a colon phantom (20). Results also showed disparate inter- and intra-observer variations in the size assessment of tumor phantoms in both colonoscopies. In a third test, three senior pathologists (from three different countries), using

Correspondence to: C.A. Rubio, MD, Ph.D, Department of Pathology, Karolinska Institute and University Hospital, 17176, Stockholm, Sweden. Fax: +46 851774524, e-mail: Carlos.Rubio@ki.se

Key Words: Breast, mammograms, tumor phantoms, tumor size.

photocopies, measured the largest actual size of 148 endoscopically-removed colorectal polyps (21). The results again showed disparate inter- and intra-observer variations in size assessment. Even digitalized computed tomography (CT) failed to recreate the gold standard size of phantom images (22).

In the present work, seven experienced mammographers were asked to assess the size of tumor phantoms on mammograms.

Materials and Methods

Tumor phantom devices. Twelve artificial tumor phantoms of different size were created with papier-mâché.

Measuring tumor phantoms at The Royal Institute of Technology. The 12 tumor phantoms were measured at The Royal Institute of Technology with the aid of low-force contacting metrology, at a temperature of 20˚C±1˚C. Held between the finger tips, each artificial tumor was rotated in a gap between two parallel metal surfaces of a micrometer screw. The distance between the surfaces was reduced until the largest diameter of the tumor phantom caused slight friction when turned around in the gap. A series of measurements was performed in random order of the 12 artificial devices. The micrometer screw (Mitutoyo Digimatic MDC-25MJT, Kawasaki, Kanagawa, Japan) has a certified uncertainty of 0.0016 mm. Only the tumor phantom with the largest diameter was measured with a calliper as its size exceeded the micrometer screw measurement range. The Luna caliper (Luna AB, Alignsås, Sweden) has 0.1 mm uncertainty. The procedure was repeated every second day and after five measurements, the average and standard deviations for each sample was calculated. The size obtained by these measurements was regarded as the gold standard.

Measuring tumor phantoms on mammograms. Tumor phantoms were haphazardly placed on an x-ray plate, each device being coded from #1 to #12 with lead granules. In trial 1, the seven mammographers measured the largest diameter of the tumor phantoms directly on the mammogram, starting from tumor phantom #1 through phantom #12, using a conventional millimeter ruler. In trial 2, the seven mammographers measured once again the largest diameter of the phantom devices two weeks later, starting this time from tumor phantom #6, down to phantom #1, followed by phantom #12 down to phantom #7. Size values by mammographers deviating by >1 mm from the gold standard were regarded as errors in assessing correct tumor size.

Statistical analysis. Each measurement was compared to the absolute value provided by The Royal Institute of Technology (considered the gold standard size) and a percentage value was calculated. The mean of the percentages obtained in the first and the second trials for each different mammographer/tumor phantom pair was calculated. The Pearson's correlation coefficient (r) was also applied to investigate the existence of a possible linear association between the age, and the years of experience as mammographer. Statistical significance was defined as $p<0.05$.

Results

Out of the seven participants measuring the tumor phantoms, three are females and the remaining four males. The age of the participants ranged from 42 to 69 years (Table I).

Measurements at The Royal Institute of Technology. The result are presented in Table I. The Table shows that the standard deviation for measurements of the largest diameter of the 12 devices was ≤0.05 mm when using the micrometer screw and ≤0.3 mm for the calliper. The difference in size in the 5 measurements was non-significant.

Measurements on mammograms. Trial 1: Table I shows that 29.8% (n=25) of the 84 values given by the seven mammographers failed to reproduce the gold standard measurements exactly.

Trial 2: Results in Table II show that 36.9% (n=31) of the 84 values given by the seven mammographers, failed to reproduce gold standard measurements.

Individual performance in size assessment: The performance of individual participants in size assessment is summarized in Table III. From the Table, it may be deduced that when compared to trial 1, two mammographers improved their performance in trial 2, one had similar success in both trials, whereas the remaining four mammographers gave lower values in trial 2 than in trial 1.

Age and gender of the mammographist and performances in size assessment: As shown in Table III, neither the age nor the gender of the mammographist influenced the performance in assessing correct size of tumor phantoms.

Years of experience in diagnostic mammography and performance of size assessment: Results in Table III show there to be no difference in assessing the correct gold standard size between mammographers with >20 years of experience (range= 21-32 years) and those with ≤16 years' experience (range= 2-16 years).

Discussion

In this survey, mammographers failed to recreate the gold standard size of tumor phantoms by >1 mm in 30% of the measurements in Trial 1, and 37% in Trial 2. Thus, the experience gained with the method in Trial 1 was of no help in improving the performance of the readings in Trial 2, 14 days later.

It may be argued that the errors in assessing the size of tumor phantom by mammographers were related to the use of a conventional ruler. However, in previous work (23), we measured by microscopy, the thickness of the collagenous band in collagenous colitis by the aid of three different methods: a) by histological estimations, b) using a calibrated micrometric ocular scale, and c) by semi-automatic

Table I. *Results of measurements of 12 tumor-phantoms (Trial 1) obtained by seven mammographers. The gold standard size was calculated at the Department of Production Engineering, The Royal Institute of Technology, Stockholm.*

Tumor phantom #	Gold standard	Measurements by mammographers (in mm)						
		A	B	C	D	E	F	G
1	8.5	8	8	8	8	8	7	8
2	13.4	14	14	14	14	9	12	12
3	24.8	25	25	26	26	26	25	24
4	18.7	19	19	19	20	20	19	18
5	8.4	8	9	9	8	9	7	8
6	18.9	19	19	19	19	19	16	17
7	16.8	17	12	18	17	18	16	17
8	16.3	16	16	17	16	17	15	16
9	10.9	11	10	11	10	12	10	10
10	10.2	9	10	11	10	10	8	9
11	16.6	15	16	16	16	16	15	15
12	27.7	28	25	28	26	28	27	27

Table II. *Results of measurements of 12 tumor-phantoms (Trial 2) obtained by seven mammographers. The gold standard size was calculated at the Department of Production Engineering, The Royal Institute of Technology, Stockholm.*

Tumor phantom #	Gold standard	Measurements by mammograpers (in mm)						
		A	B	C	D	E	F	G
1	8.5	8	13	8	8	8	7	8
2	13.4	13	14	15	14	14	12	12
3	24.8	26	26	26	26	26	15	24
4	18.7	20	20	19	20	20	17	18
5	8.4	8	14	9	9	9	7	8
6	18.9	19	18	19	19	19	16	17
7	16.8	17	17	17	17	17	17	17
8	16.3	16	16	16	16	17	15	16
9	10.9	11	11	11	11	12	11	10
10	10.2	9	10	10	10	10	8	9
11	16.6	15	16	16	16	16	14	15
12	27.7	26	28	28	28	28	17	26

micrometric measuring using a Soft Imaging System (Cell B, Olympus, Tokyo, Japan). The results also showed substantial intra- and inter-observer variations in size evaluation in two independent trials (23). Thus, even when applying more precise methods of size assessment, such as calibrated micrometric ocular scales or semi-automatic micrometric measurements, it was difficult to obtain accurate values when pathologists were confronted with the same histological sections, 14 days apart.

Table III. *Age and years of experience of seven mammographers and failure to recreate the gold standard size by >1 mm in two independent trials. The gold standard size was calculated at the Department of Production Engineering, The Royal Institute of Technology, Stockholm.*

Mamographer	Age (years)	Gender	Years of experience in mamography	Trial 1 (% error)	Trial 2 (% error)
A	57	M	21	8.3	33
B	42	F	2	16	41
C	61	F	32	16	16
D	61	M	16	25	16
E	62	M	28	41	25
F	69	F	25	66	83
G	42	M	15	25	25

M: Male; F: female.

It should be mentioned that mammographer F had 25 years of experience in reading mammograms (Table III). Hence, some mammographers, more than others, under- or overestimated the size of tumor phantoms. Notably, neither the age of the mammographers, nor their years of experience with mammogram readings reduced the errors in recreating the gold standard size.

The present study showed substantial intra- and inter-observer variations in estimating the size of phantom tumors on mammograms. The most plausible explanations for these negative findings might be lack of mental concentration, mental fatigue (due to work-overload), or both.

The results obtained raise doubts concerning the ability to discriminate size among T1 breast tumor subgroups of the TNM classification (6, 7), when applying the traditional method of size assessment in mammograms. According to the TNM staging system for breast tumors, T1 tumors are those measuring 2.0 cm or less in their greatest dimension (3). The TNM staging system recommended sub-dividing T1 breast tumors into: T1mic: microinvasion 0.1 cm or less in greatest dimension, T1a: tumors more than 0.1 but not more than 0.5 cm in greatest dimension, T1b: tumors more than 0.5 cm but not more than 1.0 cm in greatest dimension, and T1c: tumors more than 1.0 cm but not more than 2.0 cm in greatest dimension. Since this staging system is crucial in therapeutic decision making several questions arise: Are the methods used in assessing size of T1 breast tumors with the aid of a millimeter ruler in analog mammograms reliable in discriminating the minute size differences between the T1 tumor subgroups? Should the method of assessing breast tumors in analog mammograms with the aid of a millimeter ruler be abandoned? If the answer to this question is yes, then which alternative methods should be applied for measuring T1 breast tumor subgroups?

One possible alternative method to assess the size of breast tumors in analog mammograms could be that future guidelines concerning the TNM classification of breast tumors should include use of 1:1 translucent templates with the maximum size allotted for T1a, T1b and T1c tumors. The translucent templates could then be placed on suspected tumors appearing on the mammogram to enable standardized measurements of T1a, T1b and T1c breast tumors worldwide.

References

1 Haukka J, Byrnes G, Boniol M and Autier P: Trends in breast cancer mortality in Sweden before and after implementation of mammography screening. PLoS One *6*: e22422-e22428, 2011.

2 Lehtimäki T, Lundin M, Linder N, Sihto H, Holli K, Turpeenniemi-Hujanen T, Kataja V, Isola J, Joensuu H and Lundin J: Long-term prognosis of breast cancer detected by mammography screening or other methods. Breast Cancer Res *13*: R134-R143, 2011.

3 Allen S A, Cunliffe W, Gray J, Liston J E, Lunt L G, Webb L A and Young JR: Pre-operative estimation of primary breast cancer size: A comparison of clinical assessment, mammography and ultrasound. Breast *10*: 299-305, 2001.

4 Aebi S, Davidson T, Gruber G and Cardoso F: Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of Oncology 22: Supplement 6, vi12-vi24: 2011.

5 Chagpar AB and McMasters KM: Trends in mammography and clinical breast examination: a population-based study. J Surg Res *140*: 214-219, 2007.

6 The American Joint Committee on Cancer (AJCC) staging system http://cancernet.nci.nih.gov

7 Breast. *In*: AJCC Cancer Staging Manual. 7th ed. Edge SB, Byrd DR, Compton CC, Fritz A, Greene F, and Trotti A (eds.). New York, NY: Springer, pp. 347-376, 2010.

8 Fasching PA, Heusinger K, Loehberg CR, Wenkel E, Lux MP, Schrauder M, Koscheck T, Bautz W, Schulz-Wendtland R, Beckmann MW and Bani MR: Influence of mammographic density on the diagnostic accuracy of tumor size assessment and association with breast cancer tumor characteristics. Eur J Radiol *60*: 398-404, 2006.

9 Azavedo E, Zackrisson S, Mejàre I and Heibert Arnlind M: Is single reading with computer-aided detection (CAD) as good as double reading in mammography screening? A systematic review. BMC Med Imaging *12*: 22-35, 2012.

10 Hofvind S, Geller B, Vacek P, Thoresen S and Skaane P: Using the European guidelines to evaluate the Norwegian Breast Cancer Screening Program. Eur J Epidemiol *22*: 447-455, 2007.

11 Meier-Meitinger M, Rauh C, Adamietz B, Fasching PA, Schwab SA, Haeberle L, Hein A, Bayer CM, Bani MR, Lux MP, Hartmann A, Wachter DL, Uder M, Schulz-Wendtland R, Beckmann MW and Heusinger K: Accuracy of radiological tumour size assessment and the risk for re-excision in a cohort of primary breast cancer patients. Eur J Surg Oncol *38*: 44-51, 2012.

12 Regner DM, Hesley GK, Hangiandreou NJ, Morton MJ, Nordland MR, Meixner DD, Hall TJ, Farrell MA, Mandrekar JN, Harmsen WS and Charboneau JW: Breast lesions: Evaluation with US strain imaging–clinical experience of multiple observers. Radiology *238*: 425-437, 2006.

13 Oberaigner W, Daniaux M, Geiger-Gritsch S, Knapp R, Siebert U and Buchberger W: Introduction of organised mammography screening in Tyrol: Results following first year of complete rollout. BMC Public Health *11*: 673-679, 2011.

14 Taplin S, Abraham L, Barlow WE, Fenton JJ, Berns EA, Carney PA, Cutter GR, Sickles EA, Carl D and Elmore JG: Mammography facility characteristics associated with interpretive accuracy of screening mammography. J Natl Cancer Inst *100*: 876-887, 2008.

15 Lee KY, Seo BK, Yi A, Je BK, Cho KR, Woo OH, Kim MY, Cha SH, Kim YS, Son GS and Kim YS: Immersion ultrasonography of excised nonpalpable breast lesion specimens after ultrasound-guided needle localization. Korean J Radiol *9*: 312-319, 2008.

16 Pritt B, Ashikaga T, Oppenheimer RG and Weaver DL: Influence of breast cancer histology on the relationship between ultrasound and pathology tumor size measurements. Mod Pathol *17*: 905-910, 2004.

17 Smith MF, Raylman RR, Majewski S and Weisenberger AG: Positron emission mammography with tomographic acquisition using dual planar detectors: Initial evaluations. Phys Med Biol *49*: 2437-2452, 2004.

18 Weaver DL, Rosenberg RD, Barlow WE, Ichikawa L, Carney PA, Kerlikowske K, Buist DS, Geller BM, Key CR, Maygarden SJ and Ballard-Barbash R: Pathologic findings from the Breast Cancer Surveillance Consortium: Population-based outcomes in women undergoing biopsy after screening mammography. Cancer *106*: 732-742, 2006.

19 Rubio CA, Grimelius L, Lindholm J, Hamberg H, Porwit A, Elmberger G, Höög A, Kanter L, Eriksson E, Stemme S, Orrego A, Saft L, Petersson F, De La Torre M, Ekström C, Astrom K, Rundgren A, Djokic M, Chandanos E, Lenander C, Machado M, Nilsson P and Mattsson L: Reliability of the reported size of removed colorectal polyps. Anticancer Res *26*: 4895-4899, 2006.

20 Rubio CA, Höög CM, Broström O, Gustavsson J, Karlsson M, Moritz P, Stig R, Wikman O, Mattsson L and Palli D: Assessing the size of polyp phantoms in tandem colonoscopies. Anticancer Res *29*: 1539-1545, 2009.

21 Rubio CA, Jónasson JG, Nesi G, Mazur J, Olafsdóttir E: The size of colon polyps revisited: Intra- and inter-observer variations. Anticancer Res *30*: 2419-2423, 2010.

22 Suzuki C, Matsson L, Rubio CA: Assessing polyp size by improved digitalized computed tomography (CT). Anticancer Res *28*: 1911-1915, 2008.

23 Rubio CA, Orrego A, Höög A, Porwitz A, Petersson F, Elmberger G, Glaessgen A, Eriksson E, Kanter L, Jaremko G, Egevad L, Laforga J, Liljefors M, Löfdahl B, Norman P, Larsson O, Wanat R, Wejde J, Zickert P, Björk J, Caini S, Palli D and Nesi G: Quantitative assessment of the subepithelial collagen band does not increase the accuracy of diagnosis of collagenous colitis. Am J Clin Pathol *130*: 375-381, 2008.

Received December 18, 2012
Revised February 5, 2013
Accepted February 5, 2013