CrossMark

BREAST

Inter- and intra-observer agreement of BI-RADS-based subjective visual estimation of amount of fibroglandular breast tissue with magnetic resonance imaging: comparison to automated quantitative assessment

G. J. Wengert[1] · T. H. Helbich[1] · R. Woitek[1] · P. Kapetas[1] · P. Clauser[1] ·
P. A. Baltzer[1] · W-D. Vogl[2] · M. Weber[3] · A. Meyer-Baese[4] ·
Katja Pinker[1,4,5]

Abstract

Purpose To evaluate the inter-/intra-observer agreement of BI-RADS-based subjective visual estimation of the amount of fibroglandular tissue (FGT) with magnetic resonance imaging (MRI), and to investigate whether FGT assessment benefits from an automated, observer-independent, quantitative MRI measurement by comparing both approaches.

Materials and methods Eighty women with no imaging abnormalities (BI-RADS 1 and 2) were included in this institutional review board (IRB)-approved prospective study. All women underwent un-enhanced breast MRI. Four radiologists independently assessed FGT with MRI by subjective visual estimation according to BI-RADS.

Automated observer-independent quantitative measurement of FGT with MRI was performed using a previously described measurement system. Inter-/intra-observer agreements of qualitative and quantitative FGT measurements were assessed using Cohen's kappa (k).

Results Inexperienced readers achieved moderate inter-/intra-observer agreement and experienced readers a substantial inter- and perfect intra-observer agreement for subjective visual estimation of FGT. Practice and experience reduced observer-dependency. Automated observer-independent quantitative measurement of FGT was successfully performed and revealed only fair to moderate agreement (k = 0.209–0.497) with subjective visual estimations of FGT.

Conclusion Subjective visual estimation of FGT with MRI shows moderate intra-/inter-observer agreement, which can be improved by practice and experience. Automated observer-independent quantitative measurements of FGT are necessary to allow a standardized risk evaluation.

Key Points
- *Subjective FGT estimation with MRI shows moderate intra-/ inter-observer agreement in inexperienced readers.*
- *Inter-observer agreement can be improved by practice and experience.*
- *Automated observer-independent quantitative measurements can provide reliable and standardized assessment of FGT with MRI.*

✉ Katja Pinker
katja.pinker@meduniwien.ac.at

[1] Department of Biomedical Imaging and Image-guided Therapy, Division of Molecular and Gender Imaging, Medical University of Vienna/ Vienna General Hospital, Waehringer Guertel 18-20, 1090 Vienna, Austria

[2] Department of Biomedical Imaging and Image-guided Therapy, Computational Imaging Research Lab, Medical University of Vienna, Wien, Austria

[3] Department of Biomedical Imaging and Image-guided Therapy, Division of General and Pediatric Radiology, Medical University of Vienna, Wien, Austria

[4] Department of Scientific Computing in Medicine, State University of Florida, Tallahassee, FL, USA

[5] Department of Radiology, Molecular Imaging and Therapy Services, Memorial Sloan-Kettering Cancer Center, New York City, NY, USA

Introduction

The amount of fibroglandular breast tissue (FGT) is a recognized independent marker for breast cancer risk [1–5]. The American College of Radiology (ACR) [1] advises commenting on FGT, which is assessed by subjective visual estimation, when reporting mammography. However, it has been demonstrated that this assessment is prone to great intra- and inter-observer variability [6, 7]. The revised fifth edition of the ACR BI-RADS atlas has incorporated the recommendation to include such a subjective visual estimation of FGT with magnetic resonance imaging (MRI) [1]. This MRI feature has shown a high correlation to mammographic breast density assessment, and can be used to distinguish more clearly between density categories, which are closely related [8–10]. With regard to experience with mammography, it can be assumed that the assessment of FGT with MRI by subjective visual estimation would be prone to great intra- and inter-observer variability, which might limit its usefulness. However, no such data currently exists. The committee on BI-RADS recognizes that subjective estimates of FGT are imprecise [11, 12]. The investigation of three-dimensional, cross-sectional breast imaging modalities, such as MRI [11], in conjunction with observer-independent automated quantitative measurement systems [8, 13–17] for more reliable measures of the true proportion of FGT, and thus, breast cancer risk, are encouraged. The first automated observer-independent quantitative measurement approaches with MRI have been explored, with promising results [8, 13–16, 18]. Despite being aware of these limitations, for practical reasons, the committee on BI-RADS still recommends subjective visual estimation of FGT with MRI [19].

This study aimed to evaluate the inter- and intra-observer agreement of BI-RADS-based subjective visual estimation of FGT, and to investigate whether FGT assessment could benefit from an automated, observer-independent quantitative MRI measurement by comparing both approaches.

Materials and methods

Study design

Between February 2011 and June 2013, 90 women who were referred to our institution's Breast Health Care Centre for screening or diagnostic workup of abnormal imaging findings, and who ultimately had normal or benign imaging findings with mammography and ultrasound [BI-RADS 1 ($n = 71$) and 2 ($n = 9$)], were recruited for this institutional review board (IRB)-approved prospective study. The use of oral contraceptives, hormonal replacement therapy, and other types of anti-hormonal treatments, as well as known contraindications to MRI, were defined as exclusion criteria. All women gave

written, informed consent and underwent MRI for FGT assessment. In premenopausal women, all imaging studies were obtained between the 7th and 14th day of the menstrual cycle.

Imaging technique

MRI

All breast MRI examinations were performed in the prone position on a Siemens TimTrio MRI scanner at 3.0 T (Siemens, Erlangen, Germany) with a dedicated four-channel breast coil (In Vivo, Orlando, FL, USA). MR data for FGT quantification were acquired with the following sequence using the Dixon technique: TR/TE 6 ms/ 1.45 ms/ 2.67 ms; 256 slices; matrix 352 × 352; 1 mm isotropic; flip angle 6°; base resolution 352; phase resolution 100 %; bandwidth 440 Hz/Px; one average; 3 min 38 sec [17]. No contrast agent was applied.

Image analysis

Subjective visual estimation of FGT with MRI

Four breast radiologists: two readers inexperienced in MRI FGT assessment (reader one – P.K., and reader two – R.W.) and two experienced readers (reader three – P.C., and reader four – G.W.) independently performed a subjective visual estimation of FGT with MRI.

To familiarize readers with the MRI assessment of FGT by subjective visual estimation, ten training cases for each of the four density categories, which were not included in the study population, were presented to all readers (Fig. 1). These cases were selected by an experienced reader from previous studies in healthy volunteers, where both MRI and mammography data was available. FGT with MRI was then assessed for each study using both of the acquired Dixon sequences (water-only and fat-only high-contrast images) and classified as: ACR **a** – almost entirely fatty; ACR **b** – scattered fibroglandular tissue; ACR **c** – heterogeneous FGT; or as ACR **d** – extreme FGT.

All readings were performed on a five-mega-pixel PACS workstation (IMPAX EE, Agfa HealthCare GmbH, Bonn, Germany). All MRI studies were independently arranged in random order for FGT assessment. After an interval of 2 months, all four readers reassessed all MRI studies. All examinations were again arranged in random order and the previously assigned FGT readings were withheld to avoid any bias.

Automated observer-independent quantitative measurement of FGT with MRI

Automated observer-independent quantitative measurements of FGT were obtained using a previously described MRI
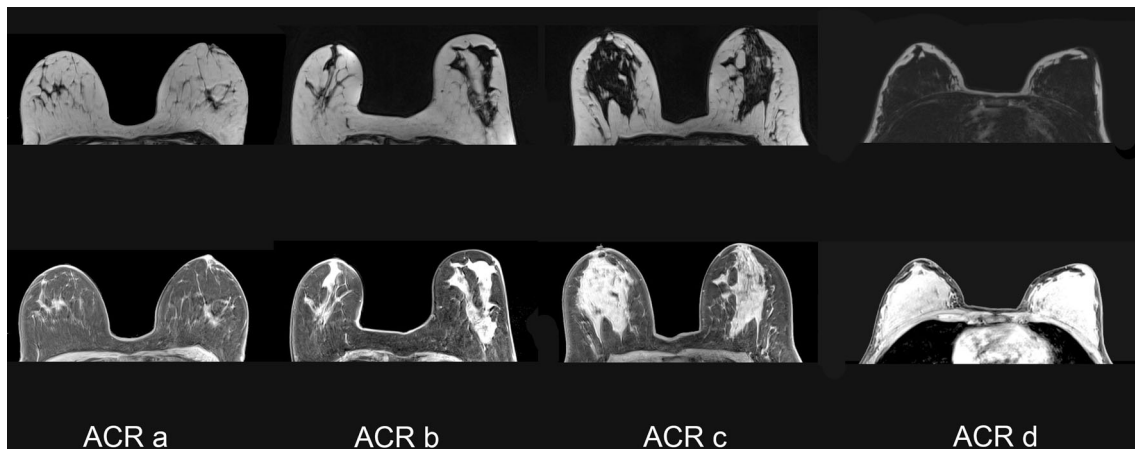
Fig. 1 Image examples for each of the four MRI FGT categories of the ACR BI-RADS atlas: ACR a – almost entirely fat, ACR b – scattered fibroglandular tissue, ARC c – heterogeneous fibroglandular tissue, and ACR d – extreme fibroglandular tissue, with the corresponding Dixon sequence-based images (upper row: fat only, lower row: glandular tissue only)

measurement system [17]. Percent fibroglandular volume (% FGV), as the ratio of the fibroglandular volume to the total breast volume, was calculated fully automatically in every woman. The calculated quantitative MRI FGT values were transformed into an MRI FGT grade analogous to the standard four ACR categories [17]. MRI FGT were scored from < 7.84 % (mean 5.67 %) as an MRI FGT grade **a**, from 7.84 to 25.88 % (mean 15.62 %) as MRI FGT grade **b**, from 26.25 to 44.15 % (mean 34.42 %) as MRI FGT grade **c**, and from 39.86 < (mean 49.74 %) as MRI FGT grade **d**.

Statistical analysis

Statistical analyses were performed using statistical software (IBM SPSS Statistics Version 22.0). Inter- and intra-observer agreement of FGT assessment with MRI by subjective visual estimation and agreement with automated observer-independent quantitative measurements of FGT with MRI were analyzed using a Cohen's kappa coefficient.

To express the differences between subjective evaluation of FGT with MRI for each individual reader and reading, the Wilcoxon signed rank test and Mantel Haenszel statistics were used.

The strength of agreement was expressed in k values: with almost perfect agreement for values from 0.81 to 0.99; substantial agreement for values from 0.61 to 0.80; values from 0.41 to 0.60 indicated moderate agreement; fair agreement was given for values from 0.21 to 0.40; slight agreement for values from 0.01 to 0.20, and values less than or equal to zero represented less than chance agreement [20].

Results

Results for BI-RADS-based subjective visual estimation of FGT with MRI and the respective MRI density grades derived by automated observer-independent quantitative measurements of FGT with MRI are summarized in Table 1.

Subjective visual estimation of FGT with MRI

Inter-observer agreement

Inter-observer agreement of subjective visual estimation of MRI FGT for all four readers is summarized in Table 2. In the first reading round, inter-observer agreement for subjective visual estimation of FGT with MRI in inexperienced readers was moderate (R1mr1 – R2mr1; k = 0.435). A substantial agreement (R3mr1 – R4mr1; k = 0.798) was observed in experienced readers.

In the second reading, inter-observer agreement for both the inexperienced and experienced readers improved substantially (range; k = 0.727 to k = 0.830).

Intra-observer agreement

Intra-observer agreement is summarized in Table 3. Experienced readers achieved better results than inexperienced readers. Intra-observer (range: k = 0.679 to k = 0.594) agreement for the inexperienced readers was moderate. Intra-observer agreement for the experienced readers was almost perfect (range: k = 0.882 to k = 0.847).

Automated observer-independent quantitative measurement of FGT with MRI

Automated observer-independent quantitative measurements of FGT with MRI were successfully performed for every examination using the previously described technique [17]. Automated observer-independent quantitative measurements of FGT with MRI ranged from 1.3 % to 76.1 % (mean 20.6 %). The translation of the calculated percentages to one

Table 1 BI-RADS-based subjective visual estimation of FGT with MRI for each reader and the respective MRI density grades derived from automated observer-independent quantitative measurements of FGT with MRI

	ACR BI-RADS Categories				
	A	B	C	D	Total
R1_1	21 (26.25 %)	18 (22.50 %)	23 (28.75 %)	18 (22.50 %)	80 (100 %)
R1_2	16 (20.00 %)	23 (28.75 %)	23 (28.75 %)	17 (21.25 %)	80 (100 %)
R2_1	8 (10.00 %)	21 (26.25 %)	24 (30.00 %)	27 (33.75 %)	80 (100 %)
R2_2	15 (18.75 %)	20 (25.00 %)	27 (33.75 %)	18 (22.50 %)	80 (100 %)
R3_1	17 (21.25 %)	20 (25.00 %)	24 (30.00 %)	19 (23.75 %)	80 (100 %)
R3_2	16 (20.00 %)	23 (28.75 %)	23 (28.75 %)	18 (22.5 %)	80 (100 %)
R4_1	15 (18.75 %)	26 (32.50 %)	24 (30.00 %)	15 (18.75 %)	80 (100 %)
R4_2	17 (21.25 %)	22 (27.50 %)	27 (33.75 %)	14 (17.50 %)	80 (100 %)
qMR	27 (33.75 %)	26 (32.50 %)	16 (20.00 %)	11 (13.75 %)	80 (100 %)

ACR American College of Radiology; *BI-RADS* Breast Imaging Reporting and Data System; *FGT* Fibroglandular tissue; *R1_1* Reader one, first reading; *R1_2* Reader one, second reading; *R2_1* Reader two, first reading; *R2_2* Reader two, second reading; *R3_1* Reader three, first reading; *R3_2* Reader three, second reading; *R4_1* Reader four, first reading; *R4_2* Reader four, second reading

of the four MRI density grade categories is summarized in Table 1.

classified fewer breasts as dense ($n = 27$, categories C and D) than non-dense ($n = 53$, categories A and B) (Table 1).

Comparison of subjective visual estimation and automated observer-independent quantitative measurements of FGT with MRI

Results for the agreement between subjective visual estimation and the quantitative measurements of FGT with MRI are summarized in Table 4.

There was only fair to moderate agreement between subjective visual estimation and the quantitative measurements of FGT with MRI, ranging from k = 0.209 to 0.497.

Compared to subjective visual estimation, automated observer-independent quantitative measurement of FGT

Discussion

Our results demonstrate that inter- and intra-observer agreement of subjective visual estimation of FGT was moderate in inexperienced readers. Experienced readers achieved better results, with a substantial inter-observer agreement and a perfect intra-observer agreement, which implies that practice and experience can reduce observer-dependency. Thus, an automated observer-independent quantitative system, which allows reproducible measurements, seems to be better suited for a reliable and standardized assessment of FGT with MRI.

The results of our study show that, analogous to FGT assessment with mammography by subjective visual estimation, MRI is observer-dependent. There was only moderate inter- and intra-observer agreement in inexperienced readers. In the second reading round, inexperienced readers achieved better results,

Table 2 Inter-observer agreement between the first and second reading for the subjective visual estimation of FGT with MRI

First Reading Round		Second Reading Round	
Observers	Kappa	Observers	Kappa
R1_1 – R2_1	**0.435**	**R1_2 – R2_2**	**0.727**
R1_1 – R3_1	0.782	R1_2 – R3_2	0.761
R1_1 – R4_1	0.650	R1_2 – R4_2	0.676
R2_1 – R3_1	0.479	R2_2 – R3_2	0.865
R2_1 – R4_1	0.445	R2_2 – R4_2	0.831
R3_1 – R4_1	**0.798**	**R3_2 – R4_2**	**0.830**

Kappa Kappa value; *MRI* Magnetic resonance imaging; *R1_1* Reader one, first reading; *R1_2* Reader one, second reading; *R2_1* Reader two, first reading; *R2_2* Reader two, second reading; *R3_1* Reader three, first reading; *R3_2* Reader three, second reading; *R4_1* Reader four, first reading; R4_2 Reader four, second reading

The bolded Reader combinations are the inexperienced readers (R1 and R2) and the expert readers (R3 and R4) fits excellent

Table 3 Intra-observer agreement between the first and second reading of FGT with MRI by subjective visual estimation

Intra-Observer Agreement	Kappa
R1_1 – R1_2	0.679
R2_1 – R2_2	0.594
R3_1 – R3_2	0.882
R4_1 – R4_2	0.847

R1_1 Reader one, first reading; *R1_2* Reader one, second reading; *R2_1* Reader two, first reading; *R2_2* Reader two, second reading; *R3_1* Reader three, first reading; *R3_2* Reader three, second reading; *R4_1* Reader four, first reading

R4_2; Reader four, second reading

Table 4 Agreement between the first and second reading of FGT with MRI by subjective visual estimation and automated observer-independent quantitative measurements of FGT with MRI

Subjective vs. quantitative MRI FGT	Kappa
R1_1 – qMR	0.400
R1_2 – qMR	0.394
R2_1 – qMR	0.209
R2_2 – qMR	0.373
R3_1 – qMR	0.369
R3_2 – qMR	0.400
R4_1 – qMR	0.497
R4_2 - qMR	0.481

qMR quantitative MRI FGT measurements; *R1_1* Reader one, first reading; *R1_2* Reader one, second reading; *R2_1* Reader two, first reading; *R2_2* Reader two, second reading; *R3_1* Reader three, first reading; *R3_2* Reader three, second reading; *R4_1* Reader four, first reading; *R4_2* Reader four, second reading

improving to a substantial inter-observer agreement. Experienced readers, in general, achieved better results, with a substantial inter-observer agreement and an almost-perfect intra-observer agreement. However, even experienced readers improved their inter-observer agreement within their already substantial agreement from 0.798 to 0.830. These results indicate that it is necessary to familiarize readers with this new MRI BI-RADS feature, and that further practice, especially for inexperienced readers, is warranted to keep inter- and intra-observer agreement to a minimum. Nevertheless, it remains doubtful whether such subjective visual estimation of FGT, because it is so dependent on practice and experience, should be used for risk evaluation, management, and the assessment of preventive breast cancer measures in women.

The limitations of subjective visual estimates of FGT have also been recognized by the committee on BI-RADS and the investigation of observer-independent automated quantitative measurement systems [8, 13–17] for more reliable measures of the true proportion of FGT have been encouraged. Automated observer-independent quantitative measurement approaches have been developed and tested for both mammography and MRI [8, 13–16, 18, 21].

Wengert et al. introduced and validated a measurement system for MRI, which was used in this study. This measurement system for MRI allows an automated, observer-independent, robust, reproducible, volumetric, quantitative FGT assessment through different levels of breast composition [17].

To our knowledge, there is currently no study that has compared subjective visual estimation of MRI FGT to an automated observer-independent quantitative MRI measurement system. The results of our study demonstrate that there are distinct differences in subjective visual and automated

observer-independent quantitative MRI FGT estimation, with only fair to moderate agreement (k = 0.209–0.497).

Compared to subjective visual estimation, automated observer-independent quantitative measurements of FGT with MRI classify fewer breasts as dense (categories C and D) than non-dense (categories A and B). These findings are in good agreement with previously published results, which compared automated observer-independent quantitative measurements of FGT with MRI to subjective mammographic FGT estimation. Khazen et al. found a twofold overestimation of mammography breast density assessment compared to MRI [13]. Based on a twofold error between mammography and MRI breast density assessment in patients with dense breast, Lee et al. [22] concluded that mammography has a limited capacity for breast density estimation due to the two-dimensional character of the modality. Thompson et al. showed that breast density assessment with interactive tissue segmentation on precontrast T1-weighted MRI revealed consequently lower results than semi-automated quantitative breast density assessment with mammography [15]. It can be expected that automated observer-independent quantitative measurements of FGT with MRI will provide the necessary standardization for a reliable measurement, as well as tracking of alterations in FGT over time. Together with clinical parameters and risk factors, this potentially might facilitate a more accurate individual breast cancer risk stratification, management, and an assessment of preventive breast cancer measures in women.

A limitation of the current study is the relatively small number of participants, as well as the number of volunteers in whom the used software was initially validated. However, this is the first study to address inter- and intra-observer agreement of subjective visual estimation of FGT with MRI, as recommended by BI-RADS, and the findings have been corroborated by previous experience with mammography and initial results for automated quantitative MRI measurements of FGT [10, 17, 23]. Another limitation is that it is uncertain as to which quantitative measurement approach of FGT comes closest in reflecting the histopathological composition of breast tissue, and therefore, a direct correlation of the amount of FGT in MRI as well as mammography with histopathology is difficult in the clinical setting.

In conclusion, subjective visual estimation of FGT with MRI shows moderate intra- and inter-observer agreement, which can be improved by practice and experience. Therefore, automated observer-independent quantitative measurements of FGT with MRI seem to be more appropriate to enable a standardized risk evaluation, management, and the assessment of preventive breast cancer measures in women.

Institutional Review Board approval was obtained. Written informed consent was obtained from all subjects (patients) in this study. Methodology: prospective, cross-sectional study, performed at one institution.

References

1. D'Orsi CJSE, Mendelson EB, Morris EA et al (2013) ACR BI-RADS® atlas, breast imaging reporting and data system. American College of Radiology, Reston
2. Boyd NF, Lockwood GA, Byng JW, Tritchler DL, Yaffe MJ (1998) Mammographic densities and breast cancer risk. Cancer Epidemiol Biomarkers Prev 7:1133–1144
3. Boyd NF, Martin LJ, Bronskill M, Yaffe MJ, Duric N, Minkin S (2010) Breast tissue composition and susceptibility to breast cancer. J Natl Cancer Inst 102:1224–1237
4. McCormack VA, dos Santos Silva I (2006) Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiol Biomarkers Prev 15:1159–1169
5. Huo CW, Chew GL, Britt KL et al (2014) Mammographic density-a review on the current understanding of its association with breast cancer. Breast Cancer Res Treat 144:479–502
6. Ciatto S, Houssami N, Apruzzese A et al (2005) Categorizing breast mammographic density: intra- and interobserver reproducibility of BI-RADS density categories. Breast 14:269–275
7. Zhou C, Chan HP, Petrick N et al (2001) Computerized image analysis: estimation of breast density on mammograms. Med Phys 28:1056–1069
8. Nie K, Chang D, Chen JH, Hsu CC, Nalcioglu O, Su MY (2010) Quantitative analysis of breast parenchymal patterns using 3D fibroglandular tissues segmented based on MRI. Med Phys 37: 217–226
9. Wei J, Chan HP, Helvie MA et al (2004) Correlation between mammographic density and volumetric fibroglandular tissue estimated on breast MR images. Med Phys 31:933–942
10. Nie K, Chen JH, Chan S et al (2008) Development of a quantitative method for analysis of breast density based on three-dimensional breast MRI. Med Phys 35:5253–5262
11. Kopans DB (2008) Basic physics and doubts about relationship between mammographically determined tissue density and breast cancer risk. Radiology 246:348–353
12. Harvey JA, Gard CC, Miglioretti DL et al (2013) Reported mammographic density: film-screen versus digital acquisition. Radiology 266:752–758
13. Khazen M, Warren RM, Boggis CR et al (2008) A pilot study of compositional analysis of the breast and estimation of breast mammographic density using three-dimensional T1-weighted magnetic resonance imaging. Cancer Epidemiol Biomarkers Prev 17:2268–2274
14. Tagliafico A, Bignotti B, Tagliafico G et al (2014) Breast density assessment using a 3T MRI system: comparison among different sequences. PLoS One 9, e99027
15. Thompson DJ, Leach MO, Kwan-Lim G et al (2009) Assessing the usefulness of a novel MRI-based breast density estimation algorithm in a cohort of women at high genetic risk of breast cancer: the UK MARIBS study. Breast Cancer Res 11:R80
16. Wang J, Azziz A, Fan B et al (2013) Agreement of mammographic measures of volumetric breast density to MRI. PLoS One 8, e81653
17. Wengert GJ, Helbich TH, Vogl WD et al (2015) Introduction of an automated user-independent quantitative volumetric magnetic resonance imaging breast density measurement system using the Dixon sequence: comparison with mammographic breast density assessment. Invest Radiol 50:73–80
18. Tagliafico A, Tagliafico G, Astengo D, Airaldi S, Calabrese M, Houssami N (2013) Comparative estimation of percentage breast tissue density for digital mammography, digital breast tomosynthesis, and magnetic resonance imaging. Breast Cancer Res Treat 138:311–317
19. D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA et al (2013) ACR BI-RADS® atlas, breast imaging reporting and data system, 5th edn. American College of Radiology, Reston
20. Viera AJ, Garrett JM (2005) Understanding interobserver agreement: the kappa statistic. Fam Med 37:360–363
21. Morrish OW, Tucker L, Black R, Willsher P, Duffy SW, Gilbert FJ (2015) Mammographic breast density: comparison of methods for quantitative evaluation. Radiology. doi:10.1148/radiol.14141508:141508
22. Lee NA, Rusinek H, Weinreb J et al (1997) Fatty and fibroglandular tissue volumes in the breasts of women 20-83 years old: comparison of X-ray mammography and computer-assisted MR imaging. AJR Am J Roentgenol 168:501–506
23. Klifa C, Carballido-Gamio J, Wilmes L et al (2010) Magnetic resonance imaging for secondary assessment of breast density in a high-risk cohort. Magn Reson Imaging 28:8–15