Radiol med (2012) 117:519–528
DOI 10.1007/s11547-011-0777-3

BREAST RADIOLOGY
SENOLOGIA

Interobserver agreement in breast radiological density attribution according to BI-RADS quantitative classification

Riproducibilità interlettore del giudizio relativo alla densità radiologica secondo la classificazione quantitativa BI-RADS

D. Bernardi[1] • M. Pellegrini[1] • S. Di Michele[1] • P. Tuttobene[1] • C. Fantò[1] • M. Valentini[1] M. Gentilini[2] • S. Ciatto[3]

[1]U.O. Senologia Clinica e Screening Mammografico, Dipartimento di Radiodiagnostica, APSS Trento I, Viale Verona Centro per i Servizi Sanitari, Palazzina C, Piano Terrazza, 38100 Trento, Italy
[2]Servizio Osservatorio Epidemiologico, Direzione Promozione ed Educazione alla Salute, APSS, Trento, Italy
[3]Centro Prevenzione Screening, ULSS 16, Padova, Italy
Correspondence to: D. Bernardi, Tel.: +39-0461-902371/+39-0461-902377, e-mail: Daniela.Bernardi@apss.tn.it

Received: 22 February 2011 / Accepted: 26 April 2011 / Published online: 7 January 2012

Abstract
Purpose. The authors sought to assess interobserver agreement in classifying mammography density according to quantitative Breast Imaging Reporting and Data System (BI-RADS) criteria.
***Materials and methods*.** Six expert mammography readers were tested on a set of 100 mammograms. Interobserver agreement was determined according to the kappa statistic, adjusting for chance agreement, on a four-category (D1 *vs*. D2 *vs*. D3 *vs*. D4) or two-category (D1–2 *vs*. D3–4) basis. Agreement with a panel of 12 readers who had been tested on the same set in a previous study was also assessed.
Results. The six readers showed good agreement when compared in pairs [agreement on a four-category basis was substantial (kappa=0.60–0.80) for 13 pairs and almost perfect (kappa>0.80) for two pairs); agreement on a two-category basis was substantial for 12 pairs and almost perfect for three pairs) or compared with the panel (on a four-category basis, agreement was substantial for five of six readers and almost perfect for one; on a two-category basis, agreement was substantial for all readers).
***Conclusions*.** In agreement with previous studies, visual classification of mammography density according to BI-RADS quantitative criteria was highly reproducible among readers; nevertheless, attribution to the "dense breast" (BI-RADS D3–4) category, which might be adopted as a determinant of different screening protocols (such as adjunct ultrasonography or yearly interval) varied

Riassunto
Obiettivo. *Scopo del presente lavoro è stato verificare la riproducibilità interosservatore nel classificare la densità mammografica in base ai criteri della classificazione quantitative Breast Imaging Reporting and Data System (BI-RADS).*
Materiali e metodi. *Sei lettori esperti di mammografia sono stati testati su un set di 100 mammografie. La concordanza interosservatore è stata valutata mediante la statistica kappa, che aggiusta per la concordanza casuale, rispetto a quattro (D1 vs. D2 vs. D3 vs. D4) o due categorie (D1-2 vs. D3-4). È stata verificata anche la concordanza con un panel di 12 lettori che avevano valutato lo stesso set di mammografie in uno studio precedente.*
Risultati*. I sei lettori hanno mostrato una buona concordanza quando confrontati in coppie (concordanza sulla base di quattro categorie sostanziale (kappa=0,60–0,80) per 13 coppie e quasi perfetta (kappa>0,80) per due coppie; concordanza sulla base di due categorie sostanziale per 12 coppie, quasi perfetta per tre) o rispetto al panel (concordanza sulla base di quattro categorie sostanziale per 5/6 lettori e quasi perfetta per un lettore; concordanza sulla base di due categorie sostanziale per tutti i lettori).*
Conclusioni. *In accordo con precedenti studi la classificazione visuale della densità mammografica secondo i criteri BI-RADS è risultata altamente*

among readers (range 6–15%). Controlled studies should be performed comparing visual with computer-density category attribution, the latter possibly being a better alternative due to its absolute reproducibility.

Keywords Breast · Diagnosis · Mammography · Density

riproducibile tra i diversi lettori: ciò nonostante l'attribuzione della categoria seno denso (BI-RADS D3-4), che potrebbe essere adottata come determinante di protocolli di screening differenziati (aggiunta dell'ecografia, frequenza annuale) varia tra i lettori (in questo studio dal 6% al 15%). Necessitano studi controllati che confrontino la classificazione visuale con quella computerizzata, potendo quest'ultima essere una valida alternativa per la sua riproducibilità assoluta.

Parole chiave *Mammella · Diagnosi · Mammografia · Densità*

Introduction

Breast radiological density is an important variable in diagnosing breast cancer. It has been associated with breast cancer risk on an individual [1–4] and familial [5–7] basis and is likely associated with mammography sensitivity, being a determinant of interval cancer risk [8–14].

The method most commonly used to classify mammography density is quantitative, proposed by the American College of Radiology in the Breast Imaging Reporting and Data System (BI-RADS) [15], which has replaced the less reproducible Wolfe's patterns [1, 3]. BI-RADS classification is commonly determined on a visual basis [3, 5, 9, 16]; however, being subjective and associated with suboptimal reproducibility [17–19], its replacement with absolutely reproducible computerised assessment has been suggested [6, 7, 19, 20].

In this study, a set of mammograms was classified according to BI-RADS quantitative density classification by a panel of radiologists involved in mammography reporting. The aim was to assess interobserver reproducibility of visual classification, its reliability in clinical use and the need for alternative methods such as computerised density assessment.

Materials and methods

The study was based on a set of 100 mammograms (digitalised mediolateral oblique and craniocaudal views of original film-screen mammograms) of women aged 50–69 years attending the Florence, Italy, screening programme: the same digitalised set had been used in a previous study of density classification reproducibility [21]. The set was made up of 69 screening tests consecutively reported as negative and 31 reported as negative and followed by interval cancers consecutively observed in the following 2 years. Six radiologists with experience in reading

Introduzione

La densità radiologica della mammella è indubbiamente una variabile importante in diagnostica senologica. Essa è stata correlata al rischio di carcinoma mammario su base individuale [1–4] o ereditaria [5–7], ed è verosimilmente associata alla sensibilità della mammografia essendo una determinante del rischio di carcinoma di intervallo [8–14].

La classificazione più comunemente usata per la definizione della densità è quella quantitativa percentuale, considerata più efficace dei patterns di Wolfe [1, 3] e proposta dall'American College of Radiology nel sistema Breast Imaging Reporting and Data System (BI-RADS) [15]. La classificazione della densità secondo BI-RADS è comunemente definita su base visuale [3, 5, 9, 16], anche se soffre di soggettività e implica problemi di riproducibilità [17–19], al punto che si è suggerita come alternativa la valutazione computerizzata [6, 7, 19, 20].

Lo scopo del presente studio, che è stato effettuato sottoponendo un set di mammografie ad un panel di radiologi comunemente addetti alla refertazione mammografica, è quello di valutare la riproducibilità interosservatore nella attribuzione della categoria di densità percentuale BI-RADS, al fine di definire se la valutazione visuale sia affidabile nella pratica clinica o debba essere abbandonata in favore a metodi alternativi più affidabili quale quello computer-assistito.

Materiali e metodi

Lo studio si è basato sulla lettura di un set di 100 mammografie digitali ottenute da immagini originali analogiche (due proiezioni: medio laterale obliqua e cranio caudale) di donne con età compresa tra 50 e i 69 anni partecipanti al programma di screening mammografico di Firenze; lo stesso set di immagini digitalizzate era stato impiegato in un precedente studio di riproducibilità della classificazione

clinical and screening mammography (at least 5,000 readings/year) were involved. They were not currently using the BI-RADS density classification, the criteria of which were briefly described before testing. Information was given on density reporting criteria as provided on the Web site by the American College of Radiology, with special mention to estimate the "volume" of the breast showing fibroglandular density by integrating the information of density area from the two standard mammography views [21]. The set was examined independently by each reader and classified according to four density categories: D1=0–25%, D2=26–50%, D3=51–75%, D4=76–100%. Interobserver agreement was assessed according to the kappa statistic, adjusting for chance agreement. Conventionally, kappa values of 0.00–0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80 and 0.81–1.00 indicate minimal, fair, moderate, substantial and almost perfect agreement, respectively [22].

In the absence of a reference standard for breast-density assessment, agreement in reporting was assessed in two different scenarios. Fifty cases had been classified by a panel of 12 radiologists in a previous reproducibility study [21]. The original set of 100 digitalised images was used, but the original reports by the panel of 12 radiologists were available only for the first half of the set. As this did not imply selection bias, the available report set was still large (600 readings), and the 12 radiologists were not specially trained in BI-RADS density reporting (unlike radiologists in this study and most radiologists currently involved in screening reading), we decided to use it as a side reference standard for the purposes of the study and for determining interobserver reproducibility between single readers. Agreement of the six radiologists in this study was thus assessed: (a) against the majority report of the panel (set of 50 cases) and (b) by comparing reports of the six radiologists in all possible pairs (15,000 cases; 15 pairs). In both scenarios, agreement was assessed on a four-category (D1 *vs*. D2 *vs*. D3 *vs*. D4) and two-category (D1–2 *vs*. D3–4) basis, the latter being more commonly used in defining breasts as either dense or not dense. When considering four density categories, the weighted kappa formula was used, which takes into account the degree of disagreement (one or two degrees) [22]. For such analysis, the SAS 9.1 statistical package was used.

Differences in the distribution of breast density category attribution between readers were checked by the chi-square test, with statistical significance being set at $p<0.05$.

Results

Table 1 shows the distribution of density categories attributed by the six radiologists in the set of 100 cases. Cate-

di densità mammografica [21]. Il set consiste di 69 esami di screening consecutivi diagnosticati come negativi e di 31 esami di screening diagnosticati come negativi e seguiti da carcinomi di intervallo consecutivi, osservati nei due anni successivi. Hanno partecipato allo studio attuale sei radiologi esperti nella lettura della mammografia clinica e di screening (almeno 5000 letture/anno); tali radiologi non utilizzavano correntemente la classificazione BI-RADS della densità mammografica, i cui principi sono stati illustrati sommariamente prima dello studio di classificazione. È stata fornita l'informazione disponibile sul sito web dell'American College of Radiology relativamente ai criteri di classificazione, in particolare per la stima del volume di mammella occupato da densità fibroghiandolare, integrando l'informazione derivata dalla stima dell'area di densità nelle due proiezioni standard [21].Ognuno dei radiologi ha visionato separatamente il test classificando ogni caso secondo una delle categorie di densità BI-RADS (D1=0%–25%, D2=26%–50%, D3=51%–75%, D4=76%–100% del volume mammario occupato da densità fibroghiandolare). La concordanza interlettore nella attribuzione della categoria di densità è stata valutata in base alla statistica kappa, che tiene conto della concordanza casuale. Convenzionalmente valori di kappa di 0,00–0,20, 0,21–0,40, 0,41–0,60, 0,61–0,80 e 0,81–1,00 vengono considerati rispettivamente indicativi di concordanza minima, scarsa, moderata, sostanziale e quasi perfetta [22].

In mancanza di uno standard di riferimento per la definizione della corretta classificazione, la concordanza è stata valutata in due diversi scenari. In un set di 50 casi era disponibile la diagnosi di maggioranza posta da un panel di 12 lettori che avevano partecipato ad un precedente studio [21] condotto con analoga finalità. Il set originale di 100 esami digitalizzati era disponibile all'uso, ma sfortunatamente le valutazioni originali da parte del panel di 12 lettori erano disponibili solo per la prima metà del set. Dal momento che questo non implicava alcun vizio di selezione, che il set era ancora sufficientemente ampio (600 letture) e che i radiologi del panel non erano particolarmente addestrati alla classificazione BI-RADS della densità (come non lo erano i radiologi coinvolti nel presente studio e la maggioranza dei radiologi correntemente addetta alla lettura di screening) abbiamo deciso di usare la valutazione del panel come standard di riferimento al fine dello studio, oltre alla valutazione della riproducibilità interoperatore tra i singoli lettori. La concordanza di ogni singolo lettore è stata quindi valutata rispetto alla diagnosi di maggioranza del panel (set di 50 casi) e successivamente nel set complessivo di 100 casi tra ogni possibile coppia dei sei lettori (15 coppie, 15.000 letture). In entrambi gli scenari la concordanza è stata valutata per le quattro categorie indipendenti (D1 vs. D2 vs. D3 vs. D4), e per le categorie inferiori e superiori accoppiate (D1–2 vs. D3–4), più comunemente

Table 1 Percent frequency of breast density categories as attributed according to BI-RADS classification (overall p=0.15)

Reader	Density categories (%)		p value (*vs*. others)
	D1–2	D3–4	
A	63	37	0.50
B	55	45	0.37
C	49	51	0.02
D	61	39	0.82
E	65	35	0.26
F	64	36	0.37

Tabella 1 *Frequenza percentuale di attribuzione delle categorie di densità BI-RADS (p complessiva: 0,15)*

Lettore	*Categorie di densità (%)*		*p (vs. altri)*
	D1-2	*D3-4*	
A	*63*	*37*	*0,50*
B	*55*	*45*	*0,37*
C	*49*	*51*	*0,02*
D	*61*	*39*	*0,82*
E	*65*	*35*	*0,26*
F	*64*	*36*	*0,37*

Table 2 Agreement (standard and weighted kappa) in attributing four (D1, D2, D3, D4) breast density BI–RADS categories by six readers compared with a reference panel (set of 50 cases)

Reader	Standard kappa	Standard kappa (CI 95%)	Weighted kappa	Weighted kappa (CI 95%)
A	0.44	0.27–0.61	0.63	0.50–0.75
B	0.52	0.35–0.69	0.68	0.56–0.80
C	0.68	0.52–0.84	0.81	0.70–0.91
D	0.52	0.35–0.69	0.68	0.56–0.81
E	0.58	0.41–0.74	0.72	0.61–0.84
F	0.52	0.3–0.69	0.69	0.56–0.81

Tabella 2 *Concordanza (kappa semplice e pesato) dei sei radiologi rispetto al panel di riferimento nella attribuzione delle quattro categorie di densità (D1 vs. D2 vs. D3 vs. D4) BI-RADS (set di 50 osservazioni)*

Lettore	*Kappa semplice*	*Kappa semplice (IC 95%)*	*Kappa pesato*	*Kappa pesato (IC 95%)*
A	*0,44*	*0,27–0,61*	*0,63*	*0,50–0,75*
B	*0,52*	*0,35–0,69*	*0,68*	*0,56–0,80*
C	*0,68*	*0,52–0,84*	*0,81*	*0,70–0,91*
D	*0,52*	*0,35–0,69*	*0,68*	*0,56–0,81*
E	*0,58*	*0,41–0,74*	*0,72*	*0,61–0,84*
F	*0,52*	*0,3–0,69*	*0,69*	*0,56–0,81*

gory distribution varied among all readers, although not to a statistically significant level (p=0.15). Only one reader showed a statistically significant difference in attributed density category distribution (excess proportion of dense breasts) when compared with the others. Table 2 shows the results of agreement assessment (standard and weighted kappa) on a four-category basis compared with the refer-

impiegate nella definizione di seno denso e non denso. Nella valutazione che considera quattro categorie di densità è stata impiegata la formula del kappa pesato che valuta diversamente il peso delle discordanze in funzione della loro entità (uno, due o tre ordini di misura) [22]. Per tale analisi è stato usato il programma di calcolo statistico SAS 9.1.

Le differenze di distribuzione della categorie di densità

Table 3 Agreement by pair (weighted kappa) among six radiologists classifying breast density according to four categories (D1, D2, D3, D4) using BI-RADS criteria (set of 100 cases): 95% confidence limits indicated in parentheses

Reader	A	B	C	D	E	F
A	–	0.72 (0.63–0.82)	0.62 (0.53–0.71)	0.69 (0.59–0.79)	0.67 (0.58–0.76)	0.73 (0.63–0.82)
B	0.72 (0.63–0.82)	–	0.67 (0.58–0.76)	0.73 (0.64–0.83)	0.65 (0.56–0.74)	0.71 (0.61–0.80)
C	0.62 (0.53–0.71)	0.67 (0.58–0.76)	–	0.61 (0.52–0.71)	0.61 (0.52–0.70)	0.63 (0.54–0.72)
D	0.69 (0.59–0.79)	0.73 (0.64–0.83)	0.61 (0.52–0.71)	–	0.72 (0.63–0.80)	0.87 (0.80–0.93)
E	0.67 (0.58–0.76)	0.65 (0.56–0.74)	0.61 (0.52–0.70)	0.72 (0.63–0.80)	–	0.80 (0.72–0.88)
F	0.73 (0.63–0.82)	0.71 (0.61–0.80)	0.63 (0.54–0.72)	0.87 (0.80–0.93)	0.80 (0.72–0.88)	–

Tabella 3 *Concordanza (kappa pesato) dei sei radiologi tra loro nella attribuzione delle quattro categorie di densità (D1 vs. D2 vs. D3 vs. D4) BI-RADS (set di 100 osservazioni): i limiti di confidenza al 95% sono indicati in parentesi*

Lettore	*A*	*B*	*C*	*D*	*E*	*F*
A	–	*0,72 (0,63–0,82)*	*0,62 (0,53–0,71)*	*0,69 (0,59–0,79)*	*0,67 (0,58–0,76)*	*0,73 (0,63–0,82)*
B	*0,72 (0,63–0,82)*	–	*0,67 (0,58–0,76)*	*0,73 (0,64–0,83)*	*0,65 (0,56–0,74)*	*0,71 (0,61–0,80)*
C	*0,62 (0,53–0,71)*	*0,67 (0,58–0,76)*	–	*0,61 (0,52–0,71)*	*0,61 (0,52–0,70)*	*0,63 (0,54–0,72)*
D	*0,69 (0,59–0,79)*	*0,73 (0,64–0,83)*	*0,61 (0,52–0,71)*	–	*0,72 (0,63–0,80)*	*0,87 (0,80–0,93)*
E	*0,67 (0,58–0,76)*	*0,65 (0,56–0,74)*	*0,61 (0,52–0,70)*	*0,72 (0,63–0,80)*	–	*0,80 (0,72–0,88)*
F	*0,73 (0,63–0,82)*	*0,71 (0,61–0,80)*	*0,63 (0,54–0,72)*	*0,87 (0,80–0,93)*	*0,80 (0,72–0,88)*	–

ence panel [21]. The standard kappa statistic showed moderate agreement for five of six readers and substantial agreement for one reader. The weighted kappa showed substantial agreement for five of six readers and almost perfect agreement for one reader. Table 3 shows results of agreement assessment (weighted kappa) on a four-category basis by pair. Agreement was substantial for 13 pairs and almost perfect for two pairs. Table 4 shows results of agreement assessment (standard kappa) on a two-category basis compared with the reference panel [21]. Agreement is substantial for all readers. Table 5 shows the results of agreement assessment (standard kappa) on a two-category basis by pair. Agreement was substantial for 12 pairs and almost perfect for three pairs.

Apart from statistical differences and depending on reader-to-reader coupling, 6–15 women in 100 would be allocated differently as to breast density (on a two-grade B1–2 *vs*. B3–4 scale) and might undergo a different screening intensity due to single-reader performance.

tra lettori sono state vagliate con il test chi-quadrato, ponendo la significatività statistica a un valore di p<0,05.

Risultati

La Tabella 1 mostra la distribuzione percentuale delle categorie di densità attribuite dai sei lettori nella classificazione del set complessivo di 100 casi. La distribuzione delle categorie varia tra i lettori ma non raggiunge la significatività statistica (p=0,15). Solo un lettore differisce dall'insieme degli altri a livello significativo per un eccesso di classificazione come seno denso. La Tabella 2 riporta i risultati della valutazione di concordanza (kappa semplice e pesato) dei sei lettori nella attribuzione delle quattro categorie BI-RADS di densità rispetto al panel di riferimento [21]. In base alla statistica kappa semplice la concordanza è almeno moderata per 5 su 6 lettori e sostanziale per uno, mentre in base al kappa pesato la concordanza è sostanziale per 5 su 6 lettori,

Table 4 Agreement (standard kappa) in attributing two (D1–2, D3–4) breast density BI-RADS categories by six readers compared with a reference panel (set of 50 cases)

Reader	Standard kappa	Standard kappa (CI 95%)
A	0.72	0.52–0.91
B	0.76	0.58–0.94
C	0.72	0.52–0.91
D	0.76	0.58–0.94
E	0.76	0.59–0.94
F	0.72	0.54–0.91

Tabella 4 *Concordanza (kappa semplice) dei sei radiologi rispetto al panel di riferimento nella attribuzione di due categorie di densità (D1-2 vs. D3-4) BI-RADS (set di 50 osservazioni) con indicazione dei limiti di confidenza al 95%*

Lettore	*Kappa semplice*	*Kappa semplice (IC 95%)*
A	*0,72*	*0,52–0,91*
B	*0,76*	*0,58–0,94*
C	*0,72*	*0,52–0,91*
D	*0,76*	*0,58–0,94*
E	*0,76*	*0,59–0,94*
F	*0,72*	*0,54–0,91*

Discussion

As intended by BI-RADS density classification principles, density value must provide an estimate of the masking effect of density and act as a predictor of the negative impact of density on sensitivity. This will enable the use of breast density categories as an indicator for tailored diagnostic approaches (such as adding ultrasonography or using shorter screening intervals in the presence of negative mammography and dense breast). It is evident that volumetric density must be estimated for such a purpose and not simply the area of density in each mammography view, which may also vary by view. For example, a volumetric density of one fourth of the breast (or 25%, e.g. corresponding to one quadrant) will translate into a 50% density area in both views, and a volumetric density of half a breast (or 50%, e.g. corresponding to both lower quadrants) will translate into a 100% area density in the craniocaudal view and a 50% area density in the lateral view. This aspect has been discussed and detailed in a previous study [21], and the assumption that volumetric density based on integration of area densities derived from the two mammographic views is needed is crucial to allow proper clinical use of breast-density assessments.

This study was based on a reference set and a panel of readers, both of which were sufficiently large to provide reliable assessment of interobserver reproducibility in categorising quantitative breast density according to

e quasi perfetta per uno. La Tabella 3 riporta i risultati della valutazione di concordanza (kappa pesato) nella attribuzione delle quattro categorie BI-RADS di densità dei sei lettori tra loro. Il kappa pesato risulta sostanziale per 13 coppie e quasi perfetto per 2 coppie. La Tabella 4 riporta i risultati della valutazione di concordanza (kappa semplice) dei sei lettori nella attribuzione di due categorie BI-RADS di densità (D1-2 vs. D3-4) rispetto al panel di riferimento [21]. La concordanza risulta sostanziale per tutti i lettori. La Tabella 5 riporta i risultati della valutazione di concordanza nella attribuzione di due categorie BI-RADS di densità (D1–2 vs. D3–4) dei sei lettori tra loro. La concordanza risulta sostanziale per 12 coppie, e quasi perfetta per 3 coppie.

A parte le considerazioni statistiche, in base alle diverse coppie di lettori da 6 a 15 soggetti su 100 possono essere classificati diversamente dal punto della densità (valutazione comparativa tra categorie D1–2 e D3–4) e potrebbero essere avviati a diversi regimi di sorveglianza in funzione del singolo lettore.

Discussione

Come suggeriscono i principi della classificazione BI-RADS, la valutazione della densità deve fornire una stima dell'effetto mascherante della densità stessa e funzionare come predittore di un effetto negativo sulla sensibilità. Ciò consentirà l'uso delle categorie di densità come indicatori per una scelta personalizzata del successivo approccio diagnostico (come ad esempio l'aggiunta dell'ecografia, o l'adozione di un intervallo di screening più breve in presenza di seno denso). È evidente che a tal fine la valutazione della densità deve essere volumetrica e non solo definire l'area di densità nelle due proiezioni, anche in considerazione del fatto che ci possono essere differenze sostanziali tra le proiezioni. Ad esempio una densità volumetrica di un quarto (25%, corrispondente ad esempio a un quadrante) si tradurrà in una area di densità del 50% in entrambe le proiezioni, mentre una densità volumetrica occupante metà della mammella (50%, ad esempio corrispondente ad entrambi i due quadranti inferiori) si tradurrà in un'area di densità del 100% nella proiezione assiale e del 50% nella proiezione laterale. Questo aspetto è stato discusso in dettaglio in un precedente studio [21] e la consapevolezza che la definizione di una densità volumetrica, basata sulle densità di area stimate nelle due proiezioni, è cruciale per consentire un uso clinico adeguato della densità.

Lo studio si basa su una casistica di riferimento e su un panel di lettori sufficiente a valutare la concordanza della attribuzione della densità radiologica del seno secondo la classificazione quantitativa BI-RADS, come dimostrato anche dai limiti di confidenza della statistica kappa che spaziano sempre solo in una categoria diagnostica adiacente a quella della stima puntuale.

Table 5 Agreement by pair (weighted kappa) among six radiologists classifying breast density according to two categories (D1–2 *vs*. D3–4) using BI-RADS criteria (set of 100 cases): 95% confidence limits indicated in parentheses

Reader	A	B	C	D	E	F
A	–	0.79 (0.68–0.91)	0.72 (0.59–0.85)	0.74 (0.61–0.88)	0.78 (0.66–0.91)	0.76 (0.63–0.89)
B	0.79 (0.68–0.91)	–	0.76 (0.63–0.89)	0.75 (0.62–0.88)	0.79 (0.67–0.91)	0.77 (0.65–0.90)
C	0.72 (0.59–0.85)	0.76 (0.63–0.89)	–	0.64 (0.50–0.79)	0.64 (0.50–0.78)	0.66 (0.52–0.80)
D	0.74 (0.61–0.88)	0.75 (0.62–0.88)	0.64 (0.50–0.79)	–	0.87 (0.77–0.97)	0.94 (0.86–1)
E	0.78 (0.66–0.91)	0.79 (0.67–0.91)	0.64 (0.50–0.78)	0.87 (0.77–0.97)	–	0.89 (0.80)–0.98)
F	0.76 (0.63–0.89)	0.77 (0.65–0.90)	0.66 (0.52–0.80)	0.94 (0.86–1)	0.89 (0.80)–0.98)	–

Tabella 5 *Concordanza (kappa semplice) dei sei radiologi tra loro nella attribuzione di due categorie di densità (D1-2 vs. D3-4) BI-RADS (set di 100 osservazioni): I limiti di confidenza al 95% sono indicati in parentesi*

Lettore	*A*	*B*	*C*	*D*	*E*	*F*
A	–	*0,79 (0,68–0,91)*	*0,72 (0,59–0,85)*	*0,74 (0,61–0,88)*	*0,78 (0,66–0,91)*	*0,76 (0,63–0,89)*
B	*0,79 (0,68–0,91)*	–	*0,76 (0,63–0,89)*	*0,75 (0,62–0,88)*	*0,79 (0,67–0,91)*	*0,77 (0,65–0,90)*
C	*0,72 (0,59–0,85)*	*0,76 (0,63–0,89)*	–	*0,64 (0,50–0,79)*	*0,64 (0,50–0,78)*	*0,66 (0,52–0,80)*
D	*0,74 (0,61–0,88)*	*0,75 (0,62–0,88)*	*0,64 (0,50–0,79)*	–	*0,87 (0,77–0,97)*	*0,94 (0,86–1)*
E	*0,78 (0,66–0,91)*	*0,79 (0,67–0,91)*	*0,64 (0,50–0,78)*	*0,87 (0,77–0,97)*	–	*0,89 (0,80)–0,98)*
F	*0,76 (0,63–0,89)*	*0,77 (0,65–0,90)*	*0,66 (0,52–0,80)*	*0,94 (0,86–1)*	*0,89 (0,80)–0,98)*	–

BI-RADS, as shown by kappa statistic confidence limits, which range only in one category adjacent to the correct estimate.

A methodological bias of the study might be the use of digitalised screen-film mammograms: breast density might be different at screen film compared with digital mammography, the latter being considered to have a higher resolution and sensitivity in dense breasts [23–24]. Even assuming differences in breast density as displayed at screen-film and digital mammography (breasts with intermediate density at screen-film mammography might be attributed to a less dense category at digital mammography), the study addresses judgement reproducibility of density, not the prevalence of density categories. Thus, using digitalised original screen-film mammograms should not represent a real bias in consideration of the study purpose.

This study confirms that interobserver agreement in reporting breast density according to BI-RADS quantitative criteria is satisfactory, which confirms the findings of a previous study on the same mammography set [21]. Some

Un appunto metodologico allo studio potrebbe riguardare il fatto che la casistica campione consiste di mammografie analogiche successivamente digitalizzate, nelle quali la densità mammaria potrebbe risultare diversa in rapporto ad un maggiore potere di risoluzione della mammografia digitale con una conseguente maggiore sensibilità nei seni densi [23, 24]. Pur ipotizzando che la definizione dell'immagine in mammografia digitale sia diversa da quella analogica (seni di densità intermedia alla analogica possono risultare meno densi alla digitale), il presente studio si riferisce alla riproducibilità del giudizio relativo alla densità e non alla prevalenza delle diverse densità; ne consegue che l'uso di casistica originalmente acquisita con metodo analogico non dovrebbe costituire un vizio di rilievo per lo studio.

Lo studio conferma che la concordanza inter-operatore nella classificazione quantitativa, analogamente ad un precedente studio condotto in Italia sulla stessa casistica campione [19, 21]. Un valore aggiunto potrebbe essere rappresentato dal fatto che nello studio precedente erano coinvolti radiologi selezionati su base volontaria (non necessariamente rappresentativi del lettore medio quanto a

added value may be accounted for, as in the previous study, selected volunteer radiologists (i.e. not necessarily representative of the average reader as to individual skill and experience) were employed, whereas in this study, an entire unselected team of dedicated radiologists in a screening programme were involved. Agreement was particularly high on a two-category basis (D1–2 *vs*. D3–4), which is mostly used to define dense breasts and is a possible candidate to indicate alternative personalised screening protocols, such as the adoption of a yearly screening interval or the adjunct of ultrasonography. The latter has been reported to improve sensitivity and might become a routine reality in the future [25–27].

The idea of automatic computer assessment of density is definitely appealing, as it reduces radiologists' workload and is supposed to have absolute reproducibility. Nevertheless, according to this study, this does not seem to be extremely important, as visual classification is quite reproducible. It is worth noting that the study was performed by readers who were not particularly trained in using the BI-RADS density classification and did not use it in daily practice. It might be suggested that proper training might obtain even higher agreement compared with that observed in this study.

Nevertheless it is evident that visual classification is associated with some degree of individual disagreement, even when only two density classes (D1–2 *vs*. D3–4) are attributed, being more evident for borderline D2–D3 cases. In this study, despite the high judgement reproducibility on statistical grounds, patients with dense breast (D3–4), a category that might be adopted for special diagnostic protocols, differed by 6–15% depending on the reader. Double reading to assess breast density would not solve the problem of reproducibility, as 6–15% of contrasting reports (on a two-grade D1–2 *vs*. D3–4 scale) would need further arbitration by a third reader. Such a workload seems unjustified, and computerised density assessment, which offers absolute reproducibility, should be explored instead.

Although computerised density assessment may be appealing, few reports on the subject suggest that volumetric breast density values as determined by computer are systematically lower compared with visual classification [28, 29]. Considering that all we know about breast density value as a risk factor or as a determinant of mammography sensitivity is based on visual classification, breast density values obtained by computer should be adapted to match commonly used visual classification categories. Thus, controlled studies are needed to compare visual and computerised density assessment and define criteria to adapt latter to former classification values. This will enable the use of software programmes for automatic density assessment that are under development and will be soon

competenza ed esperienza), mentre nel presente studio viene valutato un insieme complessivo, non selezionato, di radiologi dedicati alla lettura in un programma di screening. Lo studio mostra risultati soddisfacenti, particolarmente quando si considerino due sole classi di densità (D1–2 vs. D3–4), che potrebbero essere impiegate per definire protocolli di screening differenziati nei seni densi, quali l'adozione di un intervallo annuale o l'esecuzione di ecografia complementare. In particolare quest'ultima opzione pare avere risvolti diagnostici positivi e potrebbe in futuro entrare nella routine [25–27].

L'idea di una valutazione automatica della densità mediante computer è indubbiamente allettante, sia perché solleva il radiologo da un carico di lavoro, sia perché consente una riproducibilità assoluta. Tuttavia, alla luce del presente studio, questo secondo vantaggio non sembra così importante, in virtù della buona riproducibilità di giudizio interlettore. È bene notare che il presente studio è stato condotto con operatori non particolarmente addestrati alla classificazione BI-RADS di densità che non era infatti impiegata nella routine di lavoro: è legittimo pensare che un congruo addestramento potrebbe consentire concordanze anche superiori a quella naturale osservata.

È comunque innegabile che il giudizio visuale sia associato ad una certa discordanza, anche nella attribuzione di due sole classi di densità (D1–2 vs. D3–4), più evidente nei casi borderline tra D2 e D3. Nel presente studio, pur in presenza di una forte concordanza di giudizio sul piano statistico, è un fatto che l'attribuzione di densità potenzialmente implicante diversi protocolli diagnostici, varia in questo studio dal 6% al 15% in funzione dell'operatore. Una doppia lettura non risolverebbe il problema della riproducibilità proprio per la frequenza di valutazioni discordanti (6%–15% per le categorie D1–2 vs. D3–4) e sarebbe necessario un arbitrato per giungere ad una decisione. Un simile impegno non pare giustificato e sarebbe meglio esplorare l'affidabilità della determinazione computerizzata della densità, che ha una riproducibilità pressoché assoluta.

Se quindi una stima computerizzata della densità è comunque gradita, dalle limitate evidenze di letteratura risulta che il valore percentuale di densità (su base volumetrica) fornito dal computer sia sistematicamente inferiore a quello attribuito dall'uomo su base visuale [28, 29]. Poiché tutto quello che conosciamo sul valore della densità mammografia, sia come fattore di rischio sia come determinante della sensibilità della mammografia, si basa su categorie determinate con classificazione visuale, è evidente che i valori ottenibili con la valutazione computerizzata debbano essere tradotti nell'equivalente visuale di comune uso. Per tale motivo è auspicabile che vengano condotti studi controllati che confrontino la valutazione visuale e quella computerizzata, in modo da consentire la traduzione della stima computerizzata in valori comparabili alla stima visuale. Questo faciliterà l'uso corrente dei software per la valutazione automatica

implemented as a tool in digital mammography workstations.

Overall, limitations are evident for all methods of evaluating density as a ratio of fibroglandular density to absolute breast volume starting from more or less precise and reproducible values of 2D-assessed density area values. A more precise assessment might be possible in the future by breast density evaluation using tomosynthesis imaging, which might best fit breast density use as a covariate in breast cancer risk modelling.

della densità che già cominciano a comparire in commercio come optional nelle stazioni di lettura mammografica.

Nel complesso è evidente che ci sono limiti per qualsiasi metodo voglia valutare la densità radiologica su base volumetrica partendo da misurazioni più o meno precise e riproducibili della densità di area su base bidimensionale. Una valutazione più affidabile potrebbe essere consentita in futuro usando la tomosintesi, che potrebbe consentire una integrazione più affidabile della densità radiologica come covariata in modelli per il calcolo del rischio individuale.

Conflict of interest None

References/*Bibliografia*

1. Wolfe JN (1976) Risk for breast cancer development determined by mammographic parenchimal pattern. Cancer 37:2486–2492
2. Ciatto S, Zappa M (1993) A prospective study of the value of mammographic patterns as indicators of breast cancer risk in a screening experience. Eur J Radiol 17:122–125
3. Brisson J, Diorio C, Masse B, (2003) Wolfe's parenchymal pattern and percentage of the breast with mammographic densities: redundant or complementary classification? Cancer Epidemiol Biomarkers Prev 12:728–732
4. Vachon CM, Kuni CC, Anderson K et al (2000) Association of mammographic defined percent breast density with epidemiologic risk factors for breast cancer (United States). Cancer Causes Control 11:653–662
5. Boyd NF, Dite GS, Stone J et al (2002) Heritability of mammographic density, a risk factor for breast cancer. New Engl J Med 347:886–894
6. Ursin G, Ma H, Wu H et al (2003) Mammographic density and breast cancer in three ethnic groups. Cancer Epidemiol Biomarkers Prev 12:332–338
7. Chen Z, Wuy AH, Gauderman WJ et al (2004) Does mammographic density reflect ethnic differences in breast cancer incidence rates? Am J Epidemiol 159:140–147
8. Egan RL, Mosteller RC (1977) Breast cancer mammography patterns. Cancer 40:2087–2090
9. Ciatto S, Visioli C, Paci E, Zappa M (2004) Breast density as a determinant of interval cancer at mammographic screening. Br J Cancer 90:393–396
10. Peeters PH, Verbeek AL, Hendriks JH et al (1989) The occurrence of interval cancers in the Nijmegen screening programme. Br J Cancer 59:929–932
11. Young K, Wallis M, Blank R, Moss S (1997) Influence of number of views and mammographic film density on the detection of invasive cancers: results from the NHS Breast Screening Programme. Br J Radiol 70:482–488
12. Buist DSM, Porter PL, Lehman C et al (2004) Factors contributing to mammography failure in women aged 40-49 years. J Natl Cancer Inst 96:1432–1440
13. Caumo F, Vecchiato F, Pellegrini M et al (2009) Analysis of interval cancers observed in an Italian mammography screening programme (2000-2006). Radiol Med 114:907–914
14. Caumo F, Brunelli S, Zorzi M et al (2011) Benefits of double reading of screening mammograms: retrospective study on a consecutive series. Radiol Med 116:575–583
15. American College of Radiology (2003) The ACR Breast Imaging Reporting and Data System (BI-RADS®). American College of Radiology, Reston
16. Ciatto S, Bonardi R, Zappa M (2001) Impact of replacement hormone therapy in menopause on breast radiologic density and possible complications of mammography in the assessment of breast masses. Radiol Med 101:39–43
17. Boyd NF, Wolfson C, Moskowitz M et al (1986) Observer variation in the classification pf mammographic parenchymal patterns. J Chroinic Dis 39:465–472
18. Berg WA, Campassi C, Langenberg P, Sexton MJ (2000) Breast imaging reporting and data system: inter- and intraobserver variability in feature analysis and final assessment. AJR Am J Roentgenol 174:1769–1777
19. Vachon CM, Sellers TA, Vierkant RA et al (2002) Case control study of increased mammographic breast density response to hormone replacement therapy. Cancer Epidemiol Biomarkers Prev 11:1382–1388
20. Zhou C, Chan HP, Petrick N et al (2001) Computerized image analysis estimation of breast density on mammograms. Med Phys 28:1056–1069
21. 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
22. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
23. Rosselli Del Turco M, Mantellini P, Ciatto S et al (2007) Full-field digital versus screen-film mammography: comparative accuracy in concurrent screening cohorts. AJR Am J Roetgenol 189:860–866
24. Pisano ED, Gatsonis C, Hendrick E et al (2005) Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med 353:1773–1783

25. Berg WA, Blume JD, Cormack JB et al (2008) Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. J Am Med Ass 299:2151–2163
26. Corsetti V, Houssami N, Ferrari A et al (2008) Breast screening with ultrasound in women with mammography-negative dense breasts: evidence on incremental cancer detection and false positives, and associated cost. Eur J Cancer 44:539–544
27. Corsetti V, Houssami N, Ghirardi M et al (2011) Evidence of the effect of adjunct ultrasound screening in women with mammography-negative dense breasts: Interval breast cancers at 1year follow-up. Eur J Cancer 47:1021–1026
28. Rafferty E, Smith A, Niklason L (2009) Comparison of three methods of estimating breast density: BI-RADS density scores using full field digital mammography, breast tomosynthesis, and volumetric breast density. Proffered paper at Rad Soc North Am, Chicago, USA: ssM01
29. Tuncbilek N, Sezer A, Uğur U et al (2009) Qualitative and quantitative analysis of fibroglandular tissue in the digital environment. Proffered paper at 10th National Congress of Breast Diseases, Izmir, Turkey