

Confronto di due sistemi commerciali per l'analisi computerizzata di immagine (CAD) in ausilio alla interpretazione della mammografia di screening

Stefano CIATTO - Daniela AMBROGETTI
Rita BONARDI - Beniamino BRANCATO
Sandra CATARZI - Gabriella RISSO
Marco ROSSELLI DEL TURCO

Scopo. Confrontare l'accuratezza diagnostica di due sistemi CAD commerciali (CADx, R2) e il loro differente impatto quale ausilio alla lettura convenzionale di mammografie di screening.

Materiale e metodi. Il set di studio consta di 120 mammografie, 89 negative confermate tali e 31 con successivo carcinoma di intervallo [11 classificati "falsi negativi" (FN), 20 "segni minimi" (MS)]. Il set è stato digitalizzato e sottoposto ad elaborazione CAD con produzione di stampe cartacee delle mammografie con marcatura delle aree meritevoli di revisione. Sei radiologi esperti hanno visionato il set di mammografie tre volte consecutive, procedendo ad una lettura convenzionale e a due letture con ausilio rispettivamente di CADx e R2. I due sistemi CAD sono stati confrontati in termini di accuratezza diagnostica delle marcature e in base all'impatto della lettura CAD assistita rispetto alla lettura convenzionale e ad una simulazione di doppia lettura indipendente ottenuta combinando coppie di singole letture convenzionali.

Risultati. È evidente una maggiore marcatura di calcificazioni (218 vs 132, +65%) per R2 e di masse (208 vs 105, +98%) per CADx. CADx e R2 hanno marcato rispettivamente 15 e 17 su 31 carcinomi (sensibilità 48,3% vs 54,8%, $\chi^2=6,4$, $p=0,79$), 10 e 6 su 11 FN (90,9% vs 54,5%, $\chi^2=2,0$, $p=0,15$) e 5 e 11 su 20 MS (25,0% vs 55,0%, $\chi^2=2,6$, $p=0,10$). Quanto a specificità le marcature false positive di masse sono state in media (per caso) 1,60 per CADx e 0,75 per R2, quelle per calcificazioni 1,08 per CADx e 1,77 per R2 e quelle complessive 2,68 per CADx e 2,52 per R2. CADx e R2 hanno marcato rispettivamente 73 e 63 su 89 controlli negativi (specificità=0,18 vs 0,29, $\chi^2=2,52$, $p=0,11$). Tutti i radiologi hanno avuto una maggiore sensibilità alle letture CAD rispetto alla convenzionale. In media la sensibilità alla convenzionale è stata del 58,6% (109/186), a fronte di 70,9% (132/186) per CADx o R2 ($\chi^2=5,71$, $p=0,016$). La sensibilità per i casi FN è stata 71,2% (47/66) per la convenzionale, 84,8% (56/66) per CADx ($\chi^2=2,82$, $p=0,09$) e 80,3% (53/66) per R2 ($\chi^2=1,03$, $p=0,30$) (CADx vs R2, $\chi^2=0,21$, $p=0,64$). La sensibilità per casi MS è stata 51,6% (62/120) per la convenzionale, 63,3% (76/120) per CADx ($\chi^2=2,88$, $p=0,08$) e 65,8% (79/120) per R2 ($\chi^2=4,40$, $p=0,03$) (CADx vs. R2, $\chi^2=0,07$, $p=0,78$). I richiami ad approfondimento sono stati del 18,1% (97/534) per la convenzionale, 29,7% (159/534) per CADx ($\chi^2=5,72$, $p=0,01$) e 24,3% (130/534) per R2 ($\chi^2=10,11$, $p=10^{-5}$) (CADx vs R2, $\chi^2=3,71$, $p=0,05$). La doppia lettura è risultata significativamente più sensibile della lettura convenzionale ($\chi^2=29,6$, $p=10^{-6}$), CADx ($\chi^2=5,33$, $p=0,02$) e R2 ($\chi^2=5,33$, $p=0,02$). Il tasso di richiamo alla doppia lettura è risultato significativamente più elevato rispetto alla convenzionale ($\chi^2=21,5$, $p=10^{-6}$) mentre non è sta-

Comparison of two commercial systems for computer-assisted detection (CAD) as an aid to interpreting screening mammograms

Purpose. To compare the diagnostic accuracy of two commercial CAD systems (CADx and R2) and their impact as an aid to conventional reading of screening mammograms.

Materials and methods. The image set considered consisted of 120 mammograms, 89 confirmed negative and 31 with subsequent interval cancers (11 classified as false negatives (FN), 20 as "minimal signs" (MS)). The set was digitised and processed with CAD, and printouts obtained of the mammograms with indications of the areas warranting review. Six expert radiologists read the mammograms three times, once using conventional reading and twice using CAD reading with CADx and R2, respectively. The two CAD systems were compared in terms of diagnostic accuracy of the marks and the impact of CAD reading compared to conventional reading and to the use of independent second reading simulated by combining pairs of single conventional readings.

Results. R2 highlighted more calcifications (218 vs 132, +65%) and CADx highlighted more masses (208 vs 105, +98%). CADx and R2 marked 15 and 17 out of 31 cancers, respectively (sensitivity 48.3% vs 54.8%, $\chi^2=6.4$, $p=0.79$), 10 and 6 out of 11 FN (90.9% vs 54.5%, $\chi^2=2.0$, $p=0.15$), respectively, and 5 and 11 out of 20 MS (25.0% vs 55.0%, $\chi^2=2.6$, $p=0.10$), respectively. As for specificity, the false positive markings for masses were on average (per case) 1.60 for CADx and 0.75 for R2, those for calcifications were 1.08 for CADx and 1.77 for R2 and the total false positive markings were 2.68 for CADx and 2.52 for R2. CADx and R2 marked 73 and 63 of 89 negative controls (specificity = 0.18 vs 0.29, $\chi^2=2.52$, $p=0.11$), respectively. All the radiologists showed greater sensitivity with CAD reading compared to conventional reading. On average, sensitivity with conventional reading was 58.6% (109/186), as against 70.9% (132/186) for CADx or R2 ($\chi^2=5.71$, $p=0.016$). Sensitivity for FN cases was 71.2% (47/66) with conventional reading, 84.8% (56/66) with CADx ($\chi^2=2.82$, $p=0.09$) and 80.3% (53/66) for R2 ($\chi^2=1.03$, $p=0.30$) (CADx vs R2, $\chi^2=0.21$, $p=0.64$). Sensitivity for MS cases was 51.6% (62/120) for conventional reading, 63.3% (76/120) for CADx ($\chi^2=2.88$, $p=0.08$) and 65.8% (79/120) for R2 ($\chi^2=4.40$, $p=0.03$) (CADx vs R2, $\chi^2=0.07$, $p=0.78$). The recall rates were 18.1% (97/534) for conventional reading, 29.7% (159/534) for CADx ($\chi^2=5.72$, $p=0.01$) and 24.3% (130/534) for R2 ($\chi^2=10.11$, $p=10^{-5}$) (CADx vs R2, $\chi^2=3.71$, $p=0.05$). Double reading was significantly more sensitive than conventional reading ($\chi^2=29.6$, $p=10^{-6}$), CADx ($\chi^2=5.33$, $p=0.02$) and R2 ($\chi^2=5.33$, $p=0.02$). The recall rate for double reading was significantly higher than for conventional reading ($\chi^2=21.5$, $p=10^{-6}$) whereas no significant

ta rilevata differenza significativa rispetto a CADx ($\chi^2=0,16$, $p=0,68$) o R2 ($\chi^2=3,4$, $p=0,06$).

Conclusioni. I due sistemi CAD testati impiegano algoritmi diversi ma hanno un'accuratezza diagnostica comparabile ed un impatto favorevole simile sulla sensibilità se impiegati in ausilio alla lettura convenzionale. La lettura singola con ausilio di CAD impiegando entrambi i sistemi risulta altrettanto specifica ma non altrettanto sensibile della doppia lettura indipendente: il suo uso in alternativa alla doppia lettura non può essere raccomandato e merita ulteriore validazione mediante studi prospettici controllati.

PAROLE CHIAVE: Mammografia - Diagnosi - Screening - CAD.

difference was detected when compared to CADx ($\chi^2=0.16$, $p=0.68$) or R2 ($\chi^2=3.4$, $p=0.06$).

Conclusions. Despite using different algorithms, the two CAD systems exhibit comparable levels of diagnostic accuracy and a similar positive impact on sensitivity when used as an aid to conventional reading. Single reading with either CAD system is as specific but not as sensitive to double independent reading: its use as an alternative to double reading cannot be recommended and should be investigated further by means of controlled prospective studies.

KEY WORDS: Mammography - Diagnosis - Screening - CAD.

Introduzione

L'analisi computerizzata di immagine in ausilio alla interpretazione della mammografia (CAD) è stata ampiamente testata e numerosi studi dimostrano che CAD consente un modesto aumento di sensibilità, peraltro bilanciato da una modesta diminuzione della specificità [1-5]. In particolare con la diffusione della mammografia digitale, che ne costituisce il contesto ideale (i segnali con cui CAD indica le aree di interesse sono riprodotti automaticamente sul monitor), CAD sarà verosimilmente sempre più impiegato nella routine. Ci sono diversi sistemi CAD disponibili in commercio, che impiegano algoritmi diversi e che sono stati validati separatamente in studi prospettici o retrospettivi. Non è facile confrontare le prestazioni dei diversi sistemi CAD: infatti, poiché ogni sistema è stato testato su casistiche di mammografia diverse, il confronto dei risultati potrebbe essere inficiato dalla diversa composizione di tali casistiche, che può determinare indipendentemente la prestazione di CAD. Un confronto corretto e non viziato dei diversi sistemi CAD richiederebbe che i sistemi siano testati sulla stessa casistica, ma pochi sono gli studi del genere in letteratura, per quanto ci è noto [6, 7]. Per questo motivo abbiamo condotto uno studio retrospettivo, testando sulla stessa casistica mammografica due sistemi CAD in commercio, per confrontare le loro prestazioni quale ausilio alla interpretazione convenzionale della mammografia.

Materiale e metodi

La casistica in esame costa di 120 esami mammografici originali selezionati dagli archivi del programma di screening della città di Firenze. Tutte le indagini erano state originariamente repertate come negative. La diagnosi di negatività risultava confermata al controllo biennale in 89 casi, mentre in 31 casi (carcinomi di intervallo), entro due anni dalla diagnosi di negatività, risultava essere stato diagnosticato un carcinoma mammario (lesione unica). Le precedenti mammografie negative dei 31 carcinomi di intervallo nella casistica in esame sono state selezionate dalla serie consecutiva di carcinomi di intervallo dal programma di screening di Firenze tra quelli classificati come "falsi negativi" (FN) o "segnali minimi" (MS) da due di noi (SC, GR), in base alle linee guida della Comunità Europea [8]. Gli 89 casi negativi di controllo sono stati selezionati a random dagli archivi di screening. Il protocollo di screening del programma di Firenze prevede l'esecuzione della mammografia in sola proiezione obliqua in

Introduction

Computer-aided image analysis as an adjunct to the interpretation of mammograms (CAD) has been widely tested, with several studies demonstrating that CAD allows a modest increase in sensitivity which is counterbalanced, however, by a modest decrease in specificity [1-5]. With the spread of digital mammography, the ideal setting for CAD (the signals with which CAD indicates the areas of interest are automatically displayed on-screen), CAD is likely to be increasingly employed in routine practice.

Several CAD systems are commercially available and each uses a different algorithm and has been validated separately in prospective or retrospective studies. Comparative studies of the performance of the different CAD systems are not easy to do since each system has been tested on different mammography case sets, so that comparisons could be biased by the different composition of the sets, which alone can influence the performance of CAD. Any correct and unbiased comparison of the different CAD systems would require that the systems be tested on the same case set, but only few studies of this kind have been reported in the literature to our knowledge [6, 7]. We therefore conducted a retrospective study in which two commercial CAD systems were tested on the same set of mammograms in order to compare their performance as an aid to conventional interpretation.

Materials and methods

The set consisted of 120 original mammograms selected from the archives of the breast screening programme of Florence (Italy). All the examinations had originally been reported as negative. The negative diagnosis had been confirmed at the two-year follow-up in 89 cases, whereas 31 cases had been diagnosed with breast cancer (single lesion) within two years of the original negative diagnosis (interval cancers). The previous negative mammograms of the 31 interval cancers were selected from the consecutive series of interval cancers in the Florence screening programme among those classed as "false negative" (FN) or "minimal signs" (MS) by two of the authors (SC, GR), on the basis of the European Guidelines [8]. The 89 negative control cases were randomly selected from the screening archives. The Florence screening programme protocol envisages oblique-view mammography alone for repeat screening in subjects with fibroadipose breasts, and only oblique

TABELLA I. — Confronto di accuratezza dei sistemi CADx e R2. Valutazione di sensibilità sui casi e sul complesso delle alterazioni visibili nelle diverse proiezioni.

Sensibilità	CADx		R2		χ^2	p
	N.	%	N.	%		
— Soli casi MS	5/20	25	11/20	55	2,6	0,10
— Soli casi FN	10/11	90,9	6/11	54,5	2,0	0,15
— Casi totali	15/31	48,3	17/31	54,8	6,4	0,79
— Tutte le alterazioni visibili: masse	16/41	39	15/41	36,5	0	1
— Tutte le alterazioni visibili: microcalcificazioni	2/5	40	5/5	100	1,9	0,16
— Tutte le alterazioni visibili	18/46	39,1	20/46	43,4	0,04	0,83

TABLE I.—Comparison of the accuracy of the CADx and R2 systems. Evaluation of sensitivity on cases and on overall alterations visible in the different views.

Sensitivity	CADx		R2		χ^2	p
	No.	%	No.	%		
— MS cases alone	5/20	25	11/20	55	2.6	0.10
— FN cases alone	10/11	90.9	6/11	54.5	2.0	0.15
— Total cases	15/31	48.3	17/31	54.8	6.4	0.79
— All visible alterations: masses	16/41	39	15/41	36.5	0	1
— All visible alterations: microcalcifications	2/5	40	5/5	100	1.9	0.16
— All visible alterations	18/46	39.1	20/46	43.4	0.04	0.83

caso di screening ripetuto in soggetti con seno fibroadiposo. La sola proiezione obliqua era stata eseguita nella maggioranza dei carcinomi di intervallo (18 su 31) e dei controlli negativi (63 su 89) selezionati per lo studio.

Lo studio ha confrontato due sistemi CAD, sviluppati rispettivamente da CADx Systems Inc. (Beavercreek, Ohio, USA) e da R2 Technology Inc. (California, USA). Le mammografie originali sono state digitalizzate con metodologia descritta in precedenza [4, 5] e le immagini digitali ottenute sono state sottoposte ad analisi computerizzata, impiegando applicazioni iterative di sistemi intelligenti per l'identificazione di aree mammografiche meritevoli di una rivalutazione. Sono state prodotte stampe con indicazione della sede selezionata dal computer per la revisione (calcificazioni e masse sono state identificate sulla stampa con differenti marcature).

Una prima analisi ha valutato l'accuratezza dei due sistemi nell'identificare i carcinomi: CADx e R2 sono stati confrontati in termini di accuratezza diagnostica (sensibilità, specificità), in funzione del tipo di lesione (massa o calcificazioni), e in base al numero medio di marcature per singolo caso e per singola proiezione nella casistica complessiva.

Una successiva analisi ha valutato l'impatto di CAD sulla diagnosi radiologica. Sei radiologi esperti (<50.000 mammografie lette) correntemente addetti alla refertazione di mammografie di screening hanno valutato la casistica, esposta su un visore rotante. Anzitutto è stata eseguita una lettura con-

views had been obtained in the majority of interval cancers (18 of 31) and negative controls (63 of 89) selected for this study.

The study compared two CAD systems, one developed by CADx Systems Inc. (Beavercreek, Ohio, USA) and the other by R2 Technology Inc. (California, USA). The original mammograms were digitised as previously described [4, 5] and the images subjected to computer analysis using iterative applications of embedded intelligent systems to identify and highlight areas deserving review. Printouts were then produced with an indication of these areas (calcifications and masses are indicated by different marks on the print-out).

The first analysis assessed the accuracy of the two systems in detecting cancers: CADx and R2 were compared in terms of diagnostic accuracy (sensitivity, specificity) relative to the type of lesion (mass or calcification), and on the basis of the average number of marks per case and per view in the total case set.

The second analysis assessed the impact of CAD on the radiological diagnosis. Six expert radiologists (<50,000 mammograms interpreted) currently involved in screening mammography reporting assessed the cases on a rotating viewer. First, they performed conventional reading followed by another two reading sessions carried out with the aid of the printouts produced by the two CAD systems. The CAD

TABELLA II. — Confronto di accuratezza dei sistemi CADx e R2. Valutazione di specificità (le marcature delle lesioni carcinomatose non sono considerate).

	CADx	R2	χ^2	p
Marcature per proiezione: masse	0,59 (192/324)	0,27 (90/324)	64	10 ⁻⁶
Marcature per proiezione: microcalcificazioni	0,40 (130/324)	0,65 (213/324)	41,6	10 ⁻⁶
Marcature per proiezione	0,99 (322/324)	0,93 (303/324)	14,6	10 ⁻⁴
Marcature per caso: masse	1,60 (192/120)	0,75 (90/120)		
Marcature per caso: microcalcificazioni	1,08 (130/120)	1,77 (213/120)		
Marcature per caso	2,68 (322/120)	2,52 (303/120)		
Specificità (su 89 controlli negativi)	0,18 (16/89)	0,29 (26/89)	2,52	0,11

TABLE II.—Comparison of the accuracy of the CADx and R2 systems. Evaluation of specificity (markings of carcinomatous lesions are not considered).

	CADx	R2	χ^2	p
Markings per view: masses	0.59 (192/324)	0.27 (90/324)	64	10 ⁻⁶
Markings per view: microcalcifications	0.40 (130/324)	0.65 (213/324)	41.6	10 ⁻⁶
Markings per view	0.99 (322/324)	0.93 (303/324)	14.6	10 ⁻⁴
Markings per case: masses	1.60 (192/120)	0.75 (90/120)		
Markings per case: microcalcifications	1.08 (130/120)	1.77 (213/120)		
Markings per case	2.68 (322/120)	2.52 (303/120)		
Specificity (out of 89 negative controls)	0.18 (16/89)	0.29 (26/89)	2.52	0.11

venzionale ed in seguito sono state eseguite due altre sessioni di lettura con l'ausilio delle stampe prodotte dai due sistemi CAD in esame. Le letture CAD sono state eseguite separatamente, a distanza di una/due settimane dalla convenzionale, all'oscuro dei referti forniti nelle precedenti letture. I lettori sono stati invitati ad indicare le anomalie mammografiche (lato e sede) per le quali avrebbero richiesto un approfondimento diagnostico indicando la lesione su uno schema grafico della mammografia. I radiologi non sono stati informati dei risultati delle letture fino al completamento dello studio. I risultati sono stati valutati in base alla sensibilità e tasso di richiamo ad approfondimento diagnostico (determinato sui soli controlli negativi=1-specificità). La sensibilità è stata valutata su tutti i 31 carcinomi e separatamente sui casi FN e MS. L'analisi statistica delle differenze osservate si è basata sul test χ^2 (limite di significatività= $p<0,05$).

Complessivamente, il confronto tra lettura convenzionale e CAD si è basato su 186 (31×6) "letture" di carcinoma e 534 (89×6) letture di controlli negativi per ognuna delle due modalità CAD in valutazione. Si è stabilito che alla lettura CAD il giudizio posto alla convenzionale potesse essere modificato solo nel senso di un richiamo aggiuntivo ad accertamento diagnostico. Il confronto tra lettura CAD e convenzionale ha preso in considerazione quali modifiche del giudizio posto dalla convenzionale solo i casi nei quali

readings were performed separately one to two weeks after the conventional reading and blinded to the findings of the previous readings. The readers were asked to indicate any mammographic abnormalities (side and site) for which they would request further investigation by marking the lesion on a diagram of the mammogram. The radiologists were blinded to the results throughout the study. The results were assessed in terms of sensitivity and recall rate (determined on the negative controls alone=1-specificity). Sensitivity was assessed on all of the 31 cancers and separately on the FN and MS cases. Statistical analysis of the differences was performed with the χ^2 test (significance was inferred for values of= $p<0.05$).

Overall, the comparison between conventional reading and CAD was based on 186 (31×6) "readings" of carcinoma and 534 (89×6) readings of negative controls for each of the two CAD modalities. It was established that CAD reading could only change the judgement made at conventional reading by adding a patient recall. In comparing CAD and conventional reading, we regarded as changes to the judgement at conventional reading only those cases in which a) a case diagnosed as negative at conventional reading had been b) subsequently diagnosed as positive at CAD, in the presence of a CAD mark. The cases positive at conventional reading were automatically considered positive at CAD

TABELLA III. — Confronto di lettura convenzionale (CONV), CADx and R2. Referti positivi dei singoli radiologi in casi di carcinoma (veri positivi) e in controlli negativi (falsi positivi).

Lettore	Veri positivi: FN (6 letture×11=66)			Veri positivi: MS (6 letture×20=120)			Veri positivi: totale (6 letture×31=186)			Falsi positivi: totale (6 letture×89=534)		
	CONV	CADx	R2	CONV	CADx	R2	CONV	CADx	R2	CONV	CADx	R2
A	7	8	8	2	4	7	9	12	15	9	15	14
B	8	8	9	7	10	12	15	18	21	20	29	24
C	7	10	8	16	18	16	23	28	24	14	34	21
D	9	10	10	10	13	12	19	23	22	14	19	18
E	8	10	9	9	12	13	17	22	22	12	30	29
F	8	10	9	18	19	19	26	29	28	28	32	29
Totale su 6 letture	47	56	53	62	76	79	109	132	132	97	159	130

TABLE III.—Comparison of conventional reading (CONV), CADx and R2. Positive findings of the single radiologists in carcinoma cases (true positives) and negative controls (false positives).

Reader	True positives: FN (6 readings×11=66)			True positives: MS (6 readings×20=120)			True positives: total (6 readings×31=186)			False positives: total (6 readings×89=534)		
	CONV	CADx	R2	CONV	CADx	R2	CONV	CADx	R2	CONV	CADx	R2
A	7	8	8	2	4	7	9	12	15	9	15	14
B	8	8	9	7	10	12	15	18	21	20	29	24
C	7	10	8	16	18	16	23	28	24	14	34	21
D	9	10	10	10	13	12	19	23	22	14	19	18
E	8	10	9	9	12	13	17	22	22	12	30	29
F	8	10	9	18	19	19	26	29	28	28	32	29
Total of 6 readings	47	56	53	62	76	79	109	132	132	97	159	130

a) un caso diagnosticato come negativo alla convenzionale era stato b) successivamente diagnosticato come positivo a CAD, in presenza di una marcatura CAD. I casi positivi alla convenzionale sono stati automaticamente considerati come positivi anche a CAD. Questa modalità deriva dall'intenzione di evitare sottovalutazioni (falsi negativi) conseguenti all'impiego di CAD: abbiamo infatti voluto specificamente evitare che la lettura CAD, in assenza di marcatura, potesse far rientrare una precedente decisione di approfondimento diagnostico, posta in base alla lettura convenzionale.

Inoltre, abbiamo simulato una doppia lettura indipendente [9] combinando in coppie le sei letture convenzionali secondo tutte le possibili combinazioni. In tal modo abbiamo potuto confrontare 6 letture CAD con 15 possibili combinazioni di doppia lettura simulata. La simulazione della doppia lettura è stata condotta come doppia lettura "indipendente", nel senso che un caso è stato considerato positivo se refertato come positivo anche da uno solo dei due lettori. Anche se esistono altre possibili modalità di "arbitrare" le doppie letture discordanti, abbiamo scelto questa modalità in quanto è quella in uso presso il programma di screening di Firenze.

as well. The rationale for this decision was that we wanted to avoid under-estimation (false negatives) resulting from the use of CAD: more specifically, we wanted to prevent negative CAD readings, i.e. without markings, cancelling previous recall decisions based on the conventional reading.

Furthermore, we simulated independent double reading [9] by pairing up the six single-radiologist conventional readings in all possible combinations. This way, we were able to compare 6 CAD readings CAD with 15 possible combinations of simulated double reading. Simulation of double reading was conducted as an "independent" double reading in the sense that cases were assumed to be positive even if they were considered positive even by only one of the two readers. Although there are other ways to solve discordant double readings, we opted for this method because it is the one used in the Florence screening programme.

Results

The case set considered consisted of 120 cases, or 324 views, 31 cancers (20 classified as MS, 11 as FN) and 89

TABELLA IV. — Confronto di accuratezza diagnostica della doppia lettura indipendente (simulata) rispetto alla lettura convenzionale o assistita da CAD.

	Sensibilità %: FN	Sensibilità %: MS	Sensibilità %: totale	Tasso di richiamo % (1-specificità)
Convenzionale	71,2	51,6	58,6	18,1
CADx	84,8	63,3	70,9	29,7
R2	80,3	65,8	70,9	24,3
Doppia lettura	86,6	76,0	79,7	28,6

TABLE IV.—Comparison of diagnostic accuracy of independent double reading (simulated) compared to conventional and CAD reading

	Sensitivity %: FN	Sensitivity %: MS	Sensitivity %: total	Recall rate % (1-specificity)
Conventional	71.2	51.6	58.6	18.1
CADx	84.8	63.3	70.9	29.7
R2	80.3	65.8	70.9	24.3
Double reading	86.6	76.0	79.7	28.6

Risultati

La casistica mammografica in esame consiste di 120 casi, o 324 proiezioni, 31 carcinomi (20 classificati come MS, 11 come FN) e 89 controlli negativi. Le aree marcate da CAD per una revisione sono state 340 per CADx (media=2,8 per caso o 1,06 per proiezione, 132 microcalcificazioni, 208 masse) e 323 per R2 (media=2,7 per caso o 0,99 per proiezione, 218 microcalcificazioni, 105 masse). Non sono state rilevate differenze significative tra CADx e R2 per quanto riguarda la frequenza complessiva delle marcature, mentre è risultata evidente una maggiore marcatura di calcificazioni (218 vs 132, +65%) per R2 e di masse (208 vs 105, +98%) per CADx.

La tabella I mostra i risultati del confronto tra CADx e R2 quanto a sensibilità. La sede del carcinoma è stata marcata almeno in una proiezione in 15 su 31 casi da CADx (sensibilità in base ai casi=48,3%) e in 17 su 31 casi da R2 (sensibilità in base ai casi=54,8%, $\chi^2=6,4$, $p=0,79$). Sempre in base ai casi, CADx e R2 hanno marcato rispettivamente 10 e 6 su 11 FN (90,9% vs 54,5%, $\chi^2=2,0$, $p=0,15$) e 5 e 11 su 20 MS (25,0% vs 55,0%, $\chi^2=2,6$, $p=0,10$). Alla revisione dei radiogrammi sono state evidenziate 46 aree di alterazione mammografia (41 masse e 5 gruppi di calcificazioni) nella sede dei carcinomi e considerando le diverse proiezioni. CADx e R2 hanno marcato rispettivamente 16 e 15 su 41 masse (39,0% vs 36,5%, $\chi^2=0$, $p=1$), e 2 e 5 su 5 calcificazioni (40,0% vs 100%, $\chi^2=1,9$, $p=0,16$). Complessivamente CADx e R2 hanno marcato rispettivamente 18 e 20 su 46 alterazioni (39,1% vs 43,4%, $\chi^2=0,04$, $p=0,83$).

La specificità è stata valutata senza considerare le marcature in corrispondenza di lesioni carcinomatose. I dati relativi alla specificità sono indicati nella tabella II. Su un totale di 324 proiezioni, CADx e R2 hanno posto rispettivamente 192 e 90 marcature (0,59 vs 0,27, $\chi^2=64,0$, $p=10^{-6}$). Le marcature corrispondenti a calcificazioni sono state rispettiva-

negative controls. CADx marked a total of 340 areas for review (mean=2.8 per case or 1.06 per view, 132 microcalcifications, 208 masses) and R2 a total of 323 (mean=2.7 per case or 0.99 per view, 218 microcalcifications, 105 masses). There were no significant differences between CADx and R2 as regards overall marks, but R2 marked more calcifications (218 vs 132, +65%) and CADx marked more masses (208 vs 105, +98%).

Table I shows the results of the comparison between CADx and R2 in terms of sensitivity. The site of the cancer was marked in at least one view in 15 of 31 cases by CADx (sensitivity per case=48.3%) and in 17 of 31 cases by R2 (sensitivity per case=54.8%, $\chi^2=6.4$, $p=0.79$). Still on a case basis, CADx and R2 marked 10 and 6 of 11 FN (90.9% vs 54.5%, $\chi^2=2.0$, $p=0.15$) and 5 and 11 of 20 MS (25.0% vs 55.0%, $\chi^2=2.6$, $p=0.10$), respectively. On reviewing the mammograms, 46 areas of abnormality (41 masses and 5 calcification clusters) were found at the cancer site and considering the different views. CADx and R2 marked 16 and 15 of 41 masses (39.0% vs 36.5%, $\chi^2=0$, $p=1$), and 2 and 5 out of 5 calcifications (40.0% vs 100%, $\chi^2=1.9$, $p=0.16$), respectively. Overall, CADx and R2 marked 18 and 20 out of 46 alterations (39.1% vs 43.4%, $\chi^2=0.04$, $p=0.83$), respectively.

Specificity was assessed without considering markings in correspondence with carcinomatous lesions. The specificity data are reported in table II. Out of a total of 324 views, CADx and R2 made a total of 192 and 90 marks, respectively (0.59 vs 0.27, $\chi^2=64.0$, $p=10^{-6}$). The calcification marks were 130 for CADx and 213 for R2 (0.40 vs 0.65, $\chi^2=41.6$, $p=10^{-6}$). Overall, CADx made a total of 322 marks and R2 a total of 303 (mean marks per view=0.99 vs 0.93, $\chi^2=14.6$, $p=10^{-4}$, respectively). Of a total of 120 cases, CADx made 1.60 marks for masses and R2 made 0.75, on average (per case). The marks for calcifications

mente 130 e 213 per CADx e R2 (0,40 vs 0,65, $\chi^2=41,6$, $p=10^{-6}$). Complessivamente, le marcature sono state rispettivamente 322 e 303 per CADx e R2 (media di marcature per proiezione=0,99 vs 0,93, $\chi^2=14,6$, $p=10^{-4}$). Su 120 casi totali le marcature di masse sono state in media (per caso) 1,60 per CADx e 0,75 per R2. Le marcature per calcificazioni sono state invece in media (per caso) 1,08 per CADx e 1,77 per R2. Complessivamente, le marcature per caso sono state in media 2,68 per CADx e 2,52 per R2. CADx e R2 hanno marcato rispettivamente 73 e 63 su 89 controlli negativi (specificità=0,18 vs 0,29, $\chi^2=2,52$, $p=0,11$).

La tabella III mostra i risultati delle letture convenzionali, CADx and R2 da parte dei 6 radiologi. Tutti i radiologi hanno avuto una maggiore sensibilità alle letture CAD rispetto alla convenzionale. In media la sensibilità alla convenzionale è stata del 58,6% (109/186), a fronte di 70,9% (132/186) per CADx o R2 ($\chi^2=5,71$, $p=0,016$). La sensibilità per i casi FN è stata 71,2% (47/66) per la convenzionale, 84,8% (56/66) per CADx ($\chi^2=2,82$, $p=0,09$) e 80,3% (53/66) per R2 ($\chi^2=1,03$, $p=0,30$) (CADx vs R2, $\chi^2=0,21$, $p=0,64$). La sensibilità per casi MS è stata 51,6% (62/120) per la convenzionale, 63,3% (76/120) per CADx ($\chi^2=2,88$, $p=0,08$) e 65,8% (79/120) per R2 ($\chi^2=4,40$, $p=0,03$) (CADx vs R2, $\chi^2=0,07$, $p=0,78$). I richiami ad approfondimento sono stati del 18,1% (97/534) per la convenzionale, 29,7% (159/534) per CADx ($\chi^2=5,72$, $p=0,01$) e 24,3% (130/534) per R2 ($\chi^2=10,11$, $p=10^{-5}$) (CADx vs R2, $\chi^2=3,71$, $p=0,05$).

La tabella IV mostra i risultati medi delle letture convenzionali, CADx e R2 dei 6 radiologi rispetto alla doppia lettura indipendente simulata (15 possibili combinazioni, 165 letture FN, 300 letture MS, 465 letture di cancro, 1335 letture di controlli negativi). La doppia lettura è risultata più sensibile di qualsiasi altra forma di lettura singola, convenzionale o CAD, anche se la differenza è risultata significativa a) rispetto alla convenzionale nei casi FN ($\chi^2=6,69$, $p=0,009$), nei casi MS ($\chi^2=22,6$, $p=10^{-6}$) e nei casi totali ($\chi^2=29,6$, $p=10^{-6}$), b) rispetto a CADx nei casi MS ($\chi^2=6,26$, $p=0,01$) e nei casi totali ($\chi^2=5,33$, $p=0,02$) e c) rispetto a R2 nei casi MS ($\chi^2=4,00$, $p=0,04$) e nei casi totali ($\chi^2=5,33$, $p=0,02$). Il tasso di richiamo alla doppia lettura è risultato significativamente più elevato rispetto alla convenzionale ($\chi^2=21,5$, $p=10^{-6}$) mentre non è stata rilevata differenza significativa rispetto a CADx ($\chi^2=0,16$, $p=0,68$) o R2 ($\chi^2=3,4$, $p=0,06$).

Discussione

Questo studio si basa su una casistica sufficientemente ampia da consentire un valido confronto dell'efficacia di due sistemi CAD. L'analisi della accuratezza diagnostica dei due sistemi mostra che essi impiegano algoritmi decisamente diversi. CADx è più sensibile per le masse e R2 per le calcificazioni. Ogni tentativo di migliorare la sensibilità dello screening, compreso l'impiego di CAD, mira di fatto a ridurre gli errori diagnostici dello screening, e quindi i carcinomi di intervallo. Poiché la maggioranza di questi ultimi appare in mammografia sotto forma di massa, non di calcificazioni, l'algoritmo alla base di CAD dovrebbe idealmente mirare alla identificazione preferenziale delle masse, il che non sembra al momento avvenire per i sistemi CAD noti [3, 10, 11]. Peraltro, anche se la differenza nell'algoritmo impie-

were 1.08 for CADx and 1.77 for R2 on average (per case). Overall, the marks per case were on average 2.68 for CADx and 2.52 for R2. CADx and R2 marked 73 and 63 out of 89 negative controls, respectively (specificity=0.18 vs 0.29, $\chi^2=2.52$, $p=0.11$).

Table III shows the results of the conventional, CADx and R2 readings by the six radiologists. All radiologists displayed greater sensitivity in the CAD reading sessions compared to conventional reading. On average sensitivity at conventional reading was 58.6% (109/186), as against 70.9% (132/186) for CADx or R2 ($\chi^2=5.71$, $p=0.016$). Sensitivity for FN cases was 71.2% (47/66) for conventional, 84.8% (56/66) for CADx ($\chi^2=2.82$, $p=0.09$) and 80.3% (53/66) for R2 ($\chi^2=1.03$, $p=0.30$) (CADx vs R2, $\chi^2=0.21$, $p=0.64$). Sensitivity for MS cases was 51.6% (62/120) for conventional reading, 63.3% (76/120) for CADx ($\chi^2=2.88$, $p=0.08$) and 65.8% (79/120) for R2 ($\chi^2=4.40$, $p=0.03$) (CADx vs R2, $\chi^2=0.07$, $p=0.78$). The recall rate was 18.1% (97/534) for conventional reading, 29.7% (159/534) for CADx ($\chi^2=5.72$, $p=0.01$) and 24.3% (130/534) for R2 ($\chi^2=10.11$, $p=10^{-5}$) (CADx vs R2, $\chi^2=3.71$, $p=0.05$).

Table IV shows the average results of conventional, CADx and R2 readings of the six radiologists in comparison to the simulated independent double reading (15 possible combinations, 165 FN readings, 300 MS readings, 465 cancer readings, 1335 negative control readings). Double reading was more sensitive than any other form of reading, whether conventional or CAD, although the difference was significant a) with respect to conventional reading in FN cases ($\chi^2=6.69$, $p=0.009$), in MS cases ($\chi^2=22.6$, $p=10^{-6}$) and in total cases ($\chi^2=29.6$, $p=10^{-6}$), b) with respect to CADx in MS cases ($\chi^2=6.26$, $p=0.01$) and in total cases ($\chi^2=5.33$, $p=0.02$) and c) with respect to R2 in MS cases ($\chi^2=4.00$, $p=0.04$) and in total cases ($\chi^2=5.33$, $p=0.02$). The recall rate for double reading was significantly higher than that for conventional reading ($\chi^2=21.5$, $p=10^{-6}$) whereas there was no significant difference with CADx ($\chi^2=0.16$, $p=0.68$) or R2 ($\chi^2=3.4$, $p=0.06$).

Discussion

The present study is based on a sufficiently large series to enable a reliable comparison of the efficacy of the two CAD systems. The analysis of diagnostic accuracy of the two systems shows that they use very different algorithms. CADx is more sensitive for masses and R2 is more sensitive for calcifications. Any attempt to improve the sensitivity of screening, including the use of CAD, aims at reducing diagnostic errors, and therefore interval cancers. Because the majority of interval cancers show up on mammograms as masses rather than calcifications the CAD algorithm should ideally aim at the preferential detection of masses, which does not seem to be the case for the current CAD systems [3, 10, 11]. In addition, although there is a clear difference between the CADx and R2 algorithms, this difference did not translate into any significant difference in sensitivity, whatever the assessment. Specificity

gato da CADx e R2 è evidente, non si è tradotta in una differenza significativa in sensibilità, comunque valutata. La specificità è risultata bassa per entrambi i sistemi, confermando che questo è al momento il maggior limite di CAD: la bassa specificità si traduce in un eccesso di marcature che comporta un rilevante "rumore di fondo" che può finire con l'infastidire il lettore e questi può finire per sottovalutare le marcature di CAD. R2 è risultato lievemente più specifico di CADx ma anche in questo caso la differenza non è risultata significativa. È risultata evidente, invece, la differenza per entrambi i sistemi nel marcare i casi FN e MS: questi ultimi sono caratterizzati da alterazioni morfologiche al limite della percettibilità, evidentemente difficili da percepire anche all'analisi computerizzata: l'osservazione conferma la validità di una simile classificazione dei carcinomi di intervallo consigliata dalla Comunità Europea [8].

L'analisi dell'impatto di CAD quale ausilio alla lettura convenzionale si è basata su un classico studio retrospettivo, già impiegato in precedenza [4, 5], i cui possibili difetti sono già stati discussi e non dovrebbero rappresentare un problema importante per quanto riguarda il confronto tra i due sistemi CAD. L'impatto dell'uso di CAD era in parte prevedibile in base all'analisi della accuratezza di CAD, e lo studio conferma che la lettura CAD è superiore a quella convenzionale in quanto significativamente più sensibile, al costo di una riduzione di specificità accettabile. La sensibilità per i casi FN e MS mostra differenze inverse per CADx e R2 ma la sensibilità complessiva dei due sistemi risulta sostanzialmente identica. R2 risulta lievemente più specifico di CADx (75,7 vs 70,3%), ma anche in questo caso non a livelli significativi.

La lettura con l'ausilio di CAD è stata proposta come possibile alternativa alla doppia lettura in quanto studi precedenti hanno suggerito una accuratezza diagnostica comparabile [4, 5] con costi evidentemente minori per la lettura CAD. Questa possibilità risulta di particolare interesse considerando le difficoltà esistenti in alcuni paesi europei, a corto di radiologi addestrati, nel rispettare le linee guida della Comunità Europea che raccomandano la doppia lettura di routine [8]. I risultati di questo studio, peraltro, non sembrano supportare una simile ipotesi, in quanto la doppia lettura è risultata significativamente più sensibile (e altrettanto specifica) rispetto alla lettura singola con l'ausilio di entrambi i sistemi CAD. Una spiegazione di un simile riscontro può risiedere nel fatto che la doppia lettura implica la combinazione di due diversi criteri diagnostici individuali, il che può correggere certi limiti sistematici di interpretazione (ad esempio la tendenza a sottovalutare le piccole densità asimmetriche), che affliggono uno solo dei due lettori. Nella lettura con CAD, invece, gli eventuali limiti sistematici di interpretazione del lettore restano invariati, e la segnalazione da parte di CAD di alterazioni che sarebbero state misconosciute alla lettura convenzionale può essere sottovalutata dal lettore anche quale riflesso all'eccesso di marcature da parte del sistema. È invece verosimile che CAD aumenti la sensibilità per i casi FN, che presumibilmente non vengono diagnosticati per caduta di attenzione o per stanchezza, e sono relativamente facili da percepire sia da un secondo lettore che dal primo, allertato dalla marcatura CAD: ciò sembra confermato dall'osservazione, nel presente studio, di una sensibilità per i casi FN sostanzialmente simile tra doppia lettura (86,6%) e lettura CAD (CADx=84,8, R2=80,3%).

was low for both systems, which confirms that this is currently the major limitation of CAD: low specificity translates into excessive marking, or a substantial "background noise" that may lead the reader to underestimate CAD markings. R2 was slightly more specific than CADx but again the difference was not significant. Instead, there was a clear difference for both systems in marking the FN and MS cases: MS cases are characterised by morphological alterations that are so subtle as to constitute a challenge even for computer-based image analysis. This observation confirms the validity of a classification of interval cancers such as that recommended by the European Commission [8].

The analysis of the impact of CAD as an aid to conventional reading was based on a classic retrospective study. This type of study has been used in the past [4, 5] and its possible defects have already been discussed and should not constitute a major problem for comparing the two CAD systems. The impact of the use of CAD was in part expected on the basis of the analysis of its accuracy; the study confirms that CAD reading is superior to conventional reading as it is significantly more sensitive, even though it entails an acceptable reduction in specificity. The sensitivity for FN and MS cases shows inverse differences for CADx and R2 but the overall sensitivity of the two systems is practically identical. R2 is slightly more specific than CADx (75.7 vs 70.3%), but here again the difference is not significant.

CAD reading has been proposed as a possible alternative to double reading since the two methods have been reported to have comparable diagnostic accuracy [4, 5] with clearly lower costs for CAD reading. This possibility is particularly interesting if we consider that some European countries have difficulties complying with the European Commission's recommendations for routine double reading due to the shortage of trained radiologists [8]. The results of this study, however, do not appear to support such a hypothesis, in that double reading was found to be significantly more sensitive (and of equal specificity) compared to single reading with the aid of either of the two CAD systems. This finding may be explained by the fact that double reading involves combining the diagnostic criteria of two individuals, which may correct certain systematic reading limitations (e.g., the tendency to underestimate small asymmetrical densities) affecting only one of the readers. In CAD reading, instead, possible systematic reading limitations are unchanged, and the marking by CAD of alterations that would have been missed at conventional reading may be underestimated by the reader even as a reaction to the system's tendency to overmark. CAD is, on the other hand, likely to increase sensitivity for FN cases, which presumably go undetected due to a fall in attention or fatigue, and are relatively easy to perceive by both the second and first reader, alerted by the CAD marking: this seems to be confirmed in our study by the finding of similar sensitivity rates for FN between double reading (86.6%) and CAD reading (CADx=84.8, R2=80.3%).

Conclusioni

In conclusione, il confronto dei due sistemi CAD mostra che, nonostante evidenti differenze degli algoritmi impiegati, essi hanno un'accuratezza diagnostica comparabile ed un impatto favorevole simile sulla sensibilità se impiegati in ausilio alla lettura convenzionale. R2 risulta lievemente, ma non significativamente più specifico di CADx. La lettura singola con ausilio di CAD impiegando entrambi i sistemi risulta altrettanto specifica ma non altrettanto sensibile della doppia lettura indipendente (ancorché simulata): il suo uso nella routine come un'alternativa alla doppia lettura non può essere raccomandato e merita ulteriore validazione mediante studi prospettici controllati.

Conclusions

In conclusion, the comparison between the two CAD systems shows that, despite the clear differences in the algorithms used, they have a comparable diagnostic accuracy and a favourable impact on sensitivity when used as a aid to conventional reading. R2 is slightly, though not significantly, more specific than CADx. Single reading with the aid of CAD with either of the two systems is as specific but not as sensitive as independent double reading (albeit simulated in this study): its routine use as an alternative to double reading cannot be recommended and deserves further validation by means of controlled prospective studies.

Bibliografia/References

- 1) Warren Burhenne LJ, Wood SA, D'Orsi CJ: Potential contribution of computer aided detection to the sensitivity of screening mammography. *Radiology* 215: 554-562, 2000.
- 2) Freer TW, Ulissey MJ: Screening mammography with computer aided detection: prospective study of 12,860 patients in a community breast center. *Radiology* 220: 781-786, 2001.
- 3) Malich A, Marx C, Facius M *et al*: Tumor detection rate of a new commercially available computer-aided detection system. *Eur Radiol* 11: 2454-2459, 2001.
- 4) Ciatto S, Rosselli Del Turco M, Risso G *et al*: Comparison of standard reading and computer aided detection (CAD) on a national proficiency test of screening mammography. *Eur J Radiol* 45: 135-138, 2003.
- 5) Ciatto S, Brancato B, Rosselli Del Turco M *et al*: Comparison of standard reading and computer aided diagnosis (CAD) on a proficiency test of screening mammography. *Radiol Med* (in press).
- 6) Lechner M, Nelson M, Elvecrog R: Comparison of two commercially available computer-aided detection (CAD) systems. *Applied Radiology* 31: 31-35, 2002.
- 7) Shile PE, Guingrich JA: Detecting "missed" breast cancers: a comparison of CAD systems. *Applied Radiology* 31(Suppl):1-3, 2002.
- 8) European Guidelines for Quality Assurance in Mammographic Screening. Perry N, Broeders M, de Wolf C, Törnberg S (eds.), 3rd edition, European Commission, Luxembourg, pp.155-158, 2001.
- 9) Brancato B, Ciatto S, Bricolo D *et al*: Valutazione dell'impatto della doppia lettura da parte di lettori esperti in un'indagine mammografia di massa. *Radiol Med* 100: 21-23, 2000.
- 10) Brem RF, Schoonjans JM, Hoffmeister J *et al*: Evaluation of breast cancer with a computer-aided detection system by mammographic appearance, histology and lesion size. *Radiology* 217: 400, 2000.
- 11) Bazzocchi M, Facecchia I, Zuiani C *et al*: Application of a computer-aided detection (CAD) system to digitalized mammograms for identifying microcalcifications. *Radiol Med* 101: 334-430, 2001.

*Dott. S. Ciatto
Centro per lo Studio
e la Prevenzione Oncologica
Viale A. Volta, 171
I-50131 Firenze FI
Tel. 055/5012214
Fax 055/5001623
E-mail: s.ciatto@cspo.it*