



EUROPEAN COMMISSION
JOINT RESEARCH CENTRE

Directorate F - Health, Consumers & Reference Materials (Ispra)
Health in Society

European Commission Initiative on Breast Cancer (ECIBC): European guidelines on breast cancer screening and diagnosis Evidence profile

Healthcare question	Should double reading (with consensus or arbitration for discordant readings) vs. single reading be used to screen mammograms for early detection of breast cancer in mammography screening programmes?
Date	April 2017
Abbreviations	CI: Confidence interval RR: Risk ratio

Table 1/2: Test Accuracy

Double reading (with consensus or arbitration for discordant readings)		Single reading	
Sensitivity	0.83 (95% CI: 0.67 to 0.94)	Sensitivity	0.75 (95% CI: 0.63 to 0.86)
Specificity	0.96 (95% CI: 0.86 to 1.00)	Specificity	0.95 (95% CI: 0.86 to 1.00)

Prevalence	0.7%	0%
-------------------	------	----

Outcome	Nº of studies (Nº of participants)	Study design	Factors that may decrease certainty of evidence					Effect per 1,000 women tested				Test accuracy CoE	
								pre-test probability of 0.7%		pre-test probability of 0%			
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	Double reading (with consensus or arbitration for discordant readings)	Single reading	Double reading (with consensus or arbitration for discordant readings)	Single reading		
True positives (TP)	3 studies 252240 participants ^a	cross-sectional (cohort type accuracy study) ^{1,2,3}	not serious ^b	serious ^c	not serious ^d	not serious	none	6 (5 to 7)	5 (4 to 6)	0 (0 to 0)	0 (0 to 0)	⊕⊕⊕○ MODERATE	
								1 more TP in double reading (with consensus or arbitration for discordant readings)		0 fewer TP in double reading (with consensus or arbitration for discordant readings)			
								1 (0 to 2)	2 (1 to 3)	0 (0 to 0)	0 (0 to 0)		
								1 fewer FN in double reading (with consensus or arbitration for discordant readings)		0 fewer FN in double reading (with consensus or arbitration for discordant readings)			
False negatives (FN)													
	True negatives (TN)	3 studies 252240 participants ^e	cross-sectional (cohort type accuracy study) ^{1,2,3}	not serious ^b	serious ^c	not serious	serious ^f	none	953 (854 to 993)	943 (854 to 993)	960 (860 to 1000)	950 (860 to 1000)	⊕⊕○○ LOW
									10 more TN in double reading (with consensus or arbitration for discordant readings)		10 more TN in double reading (with consensus or arbitration for discordant readings)		
									40 (0 to 139)	50 (0 to 139)	40 (0 to 140)	50 (0 to 140)	
10 fewer FP in double reading (with consensus or arbitration for discordant readings)									10 fewer FP in double reading (with consensus or arbitration for discordant readings)				
False positives (FP)													

Explanations

- a. Pooled detection rate ‰ (overall): Double reading with consensus or arbitration: 4.7‰ (95%CI 3.4 to 6.1‰). Single reading: 4.2‰ (95%CI 3.0 to 5.5‰) (Duijm 2009, Gromet 2008, Warren 1995).
- b. Unclear information about the used reference standard. Likely to be consistent with population screening programs.
- c. The quality of the evidence was downgraded due to indirectness. First, the follow-up for interval cancers was different between studies and therefore it affects the estimated sensitivity. Warren 1995 assessed data from first screening round (3-year follow up for interval cancers). Gromet 2008 included one-year interval cancers. Duijm 2009 included two-year interval cancers. Second, all studies were performed based on data from screen-film mammography, which is an old technique that has been replaced by digital mammography in most of the European programmes.
- d. Duijm (2009) showed a lower sensitivity compared to the other studies (2-year follow up for interval cancers). Warren (1995) showed the highest sensitivity (included only first screening round). These results are compatible with data from breast cancer screening programs.
- e. Pooled false positive rate ‰ (overall): Double reading with consensus or arbitration: 46.1‰ (95%CI 28.6‰ to 67.4‰). Single reading: 47.0‰ (95%CI 29.4‰ to 68.6‰) (Duijm 2009, Gromet 2008, Warren 1995).
- f. Wide confidence intervals for false positive results might imply different consequences and decisions for stakeholders.

References

- 1. Duijm LEM, et al. Inter-observer variability in mammography screening and effect of type and number of readers on screening outcome. Br J Cancer; 2009.
- 2. Gromet M, . Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. AJR Am J Roentgenol; 2008.
- 3. Warren RM, Duffy SW, Bashir S. The value of the second view in screening mammography. Br J Radiol; 1996 .

Table 2/2: Other outcomes

Certainty assessment							Nº of participants		Effect		Certainty	Importance
Nº of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Double reading (with consensus or arbitration for discordant readings)	Single reading	Relative (95% CI)	Absolute (95% CI)		
Breast cancer detection												
7	randomised trials 1,2,3,4,5,6,7	not serious ^a	not serious	serious ^b	not serious	none	2931/593318 (0.5%)	0.5% ^c	RR 1.08 (1.03 to 1.14)	37 more per 100,000 (from 14 more to 64 more)	⊕⊕⊕○ MODERATE	CRITICAL
Interval breast cancer												
3	randomised trials ^{3,6,7}	not serious ^a	not serious	serious ^d	not serious	none	339/235708 (0.1%)	0.2% ^c	RR 0.78 (0.64 to 0.95)	48 fewer per 100,000 (from 79 fewer to 11 fewer)	⊕⊕⊕○ MODERATE	CRITICAL
Recall for assessment												
4	randomised trials ^{1,2,6,7}	not serious ^a	not serious	serious ^e	serious ^f	none	19722/303526 (6.5%)	6.0% ^c	RR 1.08 (0.99 to 1.18)	482 more per 100,000 (from 60 fewer to 1,085 more)	⊕⊕○○ LOW	CRITICAL
False positive screening result												
4	randomised trials ^{1,2,6,7}	not serious ^a	not serious	serious ^e	serious ^f	none	18265/303526 (6.0%)	5.5% ^c	RR 1.08 (0.98 to 1.19)	443 more per 100,000 (from 111 fewer to 1,053 more)	⊕⊕○○ LOW	CRITICAL

Certainty assessment							Nº of participants		Effect		Certainty	Importance
Nº of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Double reading (with consensus or arbitration for discordant readings)	Single reading	Relative (95% CI)	Absolute (95% CI)		
Positive predictive value												
4	randomised trials ^{1,2,6,7}	not serious ^a	not serious	serious ^e	serious ^f	none	1457/19722 (7.4%)	7.7%	RR 0.99 (0.95 to 1.03)	77 fewer per 100,000 (from 385 fewer to 231 more)	⊕⊕○○ LOW	CRITICAL
Breast cancer stage in situ												
2	randomised trials ^{2,5}	not serious ^a	not serious	serious ^g	serious ^h	none	81/152580 (0.1%)	0.0%	RR 1.10 (0.95 to 1.28)	4 more per 100,000 (from 2 fewer to 12 more)	⊕⊕○○ LOW	CRITICAL
Breast cancer stage I												
2	randomised trials ^{2,5}	not serious ^a	not serious	serious ^g	serious ^h	none	219/152580 (0.1%)	0.1%	RR 1.04 (0.95 to 1.14)	5 more per 100,000 (from 7 fewer to 18 more)	⊕⊕○○ LOW	CRITICAL
Breast cancer stage II												
2	randomised trials ^{2,5}	not serious ^a	not serious	serious ^g	serious ^h	none	121/152580 (0.1%)	0.1%	RR 1.03 (0.91 to 1.17)	2 more per 100,000 (from 7 fewer to 13 more)	⊕⊕○○ LOW	CRITICAL
Breast cancer stage III												
2	randomised trials ^{2,5}	not serious ^a	not serious	serious ^g	serious ^h	none	29/152580 (0.0%)	0.0%	RR 1.06 (0.82 to 1.36)	1 more per 100,000 (from 3 fewer to 6 more)	⊕⊕○○ LOW	CRITICAL

Certainty assessment							Nº of participants		Effect		Certainty	Importance
Nº of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Double reading (with consensus or arbitration for discordant readings)	Single reading	Relative (95% CI)	Absolute (95% CI)		
Breast cancer stage IV												
2	randomised trials ^{2,5}	not serious ^a	not serious	serious ^g	very serious ^h	none	8/152580 (0.0%)	0.0%	RR 1.00 (0.61 to 1.63)	0 fewer per 100,000 (from 2 fewer to 3 more)	⊕○○○ VERY LOW	CRITICAL
Breast cancer mortality												
									not estimable		-	
Cumulative incidence of advanced cancers after a negative test												
									not estimable		-	

Explanations

- Unclear information about the used reference standard. Likely to be consistent with population screening programs.
- The intervention tested may differ from the intervention of interest: two studies used only consensus (Duijm 2009, Leivo 1999), three studies only arbitration (Gromet 2008, Liston 2003, Tonita 1999), and two studies a combination of consensus and arbitration (Pauli 1996, Posso 2016). Readers were radiologists in five studies (Duijm 2009, Gromet 2008, Leivo 1999, Liston 2003, Posso 2016), radiographers in one study (Pauli 1996) and a combination of them in one study (Tonita 1999). Three studies included experienced readers (>5000 mammograms/year) (Duijm 2009, Gromet 2008, Posso 2016), three studies included less experienced readers (Liston 2003, Pauli 1996, Tonita 1999), and one study does not report the experience (Leivo 1999).
- Basal risk is estimated from the median probability of the event detected by single reading mammography among the included studies.
- The intervention tested may differ from the intervention of interest: one study used only consensus (Duijm 2009), one study only arbitration (Gromet 2008), and one study a combination of consensus and arbitration (Pauli 1996). Readers were radiologists in two studies (Duijm 2009, Gromet 2008), and radiographers in one study (Pauli 1996). Two studies included experienced readers (>5000 mammograms/year) (Duijm 2009, Gromet 2008), and one study included less experienced readers (Pauli 1996). The outcome is reported after different follow up times: 12 months in one study (Gromet 2008), 18 months in one study (Pauli 1996), and 24 months in one study (Duijm 2009).
- The intervention tested may differ from the intervention of interest: one study used only consensus (Duijm 2009), two studies only arbitration (Gromet 2008, Tonita 1999), and one study a combination of consensus and arbitration (Posso 2016). Readers were radiologists in three studies (Duijm 2009, Gromet 2008, Posso 2016), and a combination of radiologists and radiographers in one study (Tonita 1999). Three studies included experienced readers (>5000 mammograms/year) (Duijm 2009, Gromet 2008, Posso 2016), and one study included less experienced readers (Tonita 1999).
- Effect estimate crosses the no effect threshold precluding affirming which intervention is more favourable.

- g. The intervention tested may differ from the intervention of interest: one study used only consensus (Leivo 1999), and one study a combination of consensus and arbitration (Posso 2016). Readers were radiologists in both studies (Leivo 1999, Posso 2016). One study included experienced readers (>5000 mammograms/year) (Posso 2016), and one study does not report the experience.
- h. Small number of events (Leivo 1999, Posso 2016).

References

1. Tonita JM, Hillis JP, Lim CH.. Medical radiologic technologist review: effects on a population-based breast cancer screening program. Radiology; 1999.
2. Posso MC, Puig T, Quintana MJ, Solà-Roca J, Bonfill X.. Double versus single reading of mammograms in a breast cancer screening programme: a cost-consequence analysis. Eur Radiol; 2016.
3. Pauli R, Hammond S, Cooke J, Ansell J.. Comparison of radiographer/radiologist double film reading with single reading in breast cancer screening. J Med Screen; 1996.
4. Liston JC, Dall BJG. Can the NHS Breast Screening Programme afford not to double read screening mammograms?. Clin Radiol; 2003.
5. Leivo T, Salminen T, Sintonen H, Tuominen R, Auerma K, Partanen K, Saari U, Hakama M, Heinonen OP.. Incremental cost-effectiveness of double-reading mammograms. Breast Cancer Res Treat; 1999.
6. Gromet M. Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. AJR Am J Roentgenol; 2008.
7. Duijm LEM, et al. Inter-observer variability in mammography screening and effect of type and number of readers on screening outcome. Br J Cancer; 2009.