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

European Commission, Via Enrico Fermi 2749, I-21027 Ispra (Varese) - Italy. Telephone: (39)0332-78-9111. E-mail: JRC-CANCER-POLICY-SUPPORT@ec.europa.eu

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?

POPULATION:	Asymptomatic women
INTERVENTION:	double reading (with consensus or arbitration for discordant readings)
COMPARISON:	single reading
PURPOSE OF THE TEST:	
LINKED TREATMENTS:	
ANTICIPATED OUTCOMES:	Sensitivity, specificity, breast cancer detection, interval breast cancer, false positive screening result, recall for assessment, positive predictive value, breast cancer mortality, breast cancer stage, cumulative incidence of advanced cancers after a negative test.
SETTING:	European Union
PERSPECTIVE:	Population (national health system)
BACKGROUND:	Mammography screening is strongly recommended for women aged 50 to 69 because the benefits outweigh the harms. Many countries have organised programmes according to the EC recommendation 2003. Nevertheless, mammography sensitivity can be low, thus limiting the efficacy of screening. On the other hand, some of the undesirable effects of screening are due to the low specificity (false positive screening exams and invasive assessment) of the test. Practice varies with respect to image reading and diagnostic protocols. Optimising mammography sensitivity and specificity is therefore important to optimise the benefit-harm balance of screening. One of the methods that has been adopted to improve sensitivity of mammography screening is double reading, whereby the mammograms are read, generally independently, by two trained readers. If every mammogram that is read as positive by one or both readers is recalled for assessment, this method has necessarily a detrimental impact on specificity. To mitigate or avoid this problem, mammograms with discordant readings can be reviewed by a third reader (arbitration) or can be discussed by the two readers to reach consensus. Another possible scenario for doing consensus is when the two readers agree on a positive result. Published articles addressing this topic, however, are missing. The main objective of this question is whether a strategy in reading mammograms by double reading (independent or dependent, blinded or not) with a consensus conference (1st intervention) or double reading with arbitration (2nd intervention) is superior to single reading (comparison) with regards to the outcomes of breast cancer mortality, stage of breast cancer detected, interval cancer rate, advanced cancers in subsequent rounds, false positive and false negative results of screening mammograms, recall rates and breast cancer detection rate.
CONFLICT OF INTEREST:	Management of Conflicts of Interests (CoI): CoIs of all Guideline Development Group (GDG) members were assessed and managed by the Joint Research Centre (JRC) following an established procedure in line with European Commission rules. GDG member participation in the development of the recommendations was restricted, according to CoI disclosure. Consequently, for this particular question, the following GDG members were recused from voting: Axel Gräwingholt and Elsa Pérez Gómez for not providing additional information as requested. Miranda Langendam, as external expert, was also not allowed to vote, according to the ECIBC rules of procedure.

ASSESSMENT

Problem

Is the problem a priority?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ No ○ Probably no ○ Probably yes ● Yes ○ Varies ○ Don't know	Breast cancer is the second most common cancer in the world and, by far, the most frequent cancer among women with an estimated 1.67 million new cancer cases diagnosed in 2012—accounting for 25% of all cancers (1). Breast cancer ranks as the fifth leading cause of cancer death worldwide and the second leading cause of cancer-related death in developed regions (1). In the European Union, 367 090 women were diagnosed with breast cancer and 92 000 women died from the disease in 2012 (2). Breast cancer ranks fourth among the top five cancers with the highest disease burden (3). Annual incidence of breast cancer in the EU among women aged 50 to 69 is 2.7 per 1 000 and mortality is 0.5 per 1 000 (1).	This question was prioritised by the GDG

Test accuracy

How accurate is the test?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ Very inaccurate ○ Inaccurate ● Accurate ○ Very accurate ○ Varies ○ Don't know	**Test accuracy** Double reading (with consensus or arbitration) Sensitivity: 0.83 (95% CI: 0.67 to 0.94) Specificity: 0.96 (95% CI: 0.86 to 1.00) Single reading Sensitivity: 0.75 (95% CI: 0.63 to 0.86) Specificity: 0.95 (95% CI: 0.86 to 1.00)	**Breast cancer detection:** -(9) reported a higher cancer detection rate with double reading with or without consensus or arbitration-digital mammography than with single reading-film mammography (3.88 per 1000 vs 2.63 per 1000). **False positive screening result:** -(10) reported that double reading with or without consensus or arbitration mammograms conferred a higher risk of false positive results (OR = 2.06; 95% CI 2.00–2.13) than single reading. **Breast cancer invasiveness:** -Double reading with consensus detected 17.0% of DCIS and single reading detected 16.3% (5). -Double reading with arbitration detected 30.3% of DCIS and single reading detected 29.5% (6). -Double reading with consensus or arbitration detected 32% more small (<15 mm) invasive cancers than single reading, in prevalent screening. This detection increased to 73% more cancers in incident screening (11). The GDG notes that two additional subgroup analyses of the systematic review data were conducted. As there were

Test result	Number of results per 1,000 women tested (95% CI)				№ of participants (studies)	Certainty of the evidence (GRADE)
	Prevalence 0.7%		Prevalence 0%			
	Double reading (with consensus or arbitration for discordant readings)	Single reading	Double reading (with consensus or arbitration for discordant readings)	Single reading		
True positives patients with breast cancer	6 (5 to 7)	5 (4 to 6)	0 (0 to 0)	0 (0 to 0)	252240 (3)[a]	⨁⨁⨁◯ MODERATE[1,2,3,b,c,d]
	1 more TP in double reading (with consensus or		0 fewer TP in double reading (with consensus or			

	arbitration for discordant readings)		arbitration for discordant readings)			
False negatives patients incorrectly classified as not having breast cancer	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)			
True negatives patients without breast cancer	953 (854 to 993)	943 (854 to 993)	960 (860 to 1000)	950 (860 to 1000)	252240 (3)[e]	⨁⨁◯◯ LOW[b,d,f]
	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)			
False positives patients incorrectly classified as having breast cancer	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)			

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.

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. 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

only 3 studies (5) (6) (4), a univariate random effects logistic regression model instead of a bivariate model was fitted. The model assumes a binomial distribution of the data and uses the maximum likelihood estimation, which is an alternative method to estimate between-study variance in situations with few studies or sparse data (12) (13).

Excluding (6) (due to arbitration not being performed in the same way in all disagreements):

Double reading with consensus or arbitration

Sensitivity: 0.76 (95% CI: 0.73 to 0.78) Specificity: 0.99 (95% CI: 0.99 to 0.99)

Single reading

Sensitivity: 0.69 (95% CI: 0.66 to 0.72) Specificity: 0.98 (95% CI: 0.98 to 0.98).

Excluding (4) (due to this study included prevalent screening only):

Double reading with consensus or arbitration

Sensitivity: 0.77 (95% CI: 0.75 to 0.79) Specificity: 0.95 (95% CI: 0.95 to 0.95)

Single reading

Sensitivity: 0.72 (95% CI: 0.69 to 0.74) Specificity: 0.96 (95% CI: 0.96 to 0.96).

The GDG interpreted that there is higher sensitivity with double reading with consensus or arbitration compared to single reading in mammography screening.

The GDG notes that the three studies reviewed did not use digital mammography (5) (6) (4).

The GDG notes that for other outcomes double vs single reading showed the following results per 100 000 screening mammograms: 37 more breast cancers detected; 48 fewer interval breast cancers; 482 more recalls; 443 more false positive results; 77 fewer true positives per 100 000 recalls; 4 more breast cancer in situ per 100.000 screening mammograms; breast cancer stage I 5 more; breast cancer stage II 2 more; breast cancer stage III 1 more; breast cancer stage IV no difference. No estimates were identified for breast cancer mortality.

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.

c. Duijim (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.
d. Unclear information about the used reference standard. Likely to be consistent with population 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.

Other outcomes

Outcomes	Anticipated absolute effects* (95% CI)		Relative effect (95% CI)	№ of mammograms (studies)	Certainty of the evidence (GRADE)
	Risk with single reading	Risk with double reading (with consensus or arbitration)			
Breast cancer detection	Low		**RR 1.08** (1.03 to 1.14)	593318 (7 cohort studies)[1,2,3,4,5,6,7]	⨁⨁⨁◯ MODERATE[a,b]
	460 per 100,000[c]	**497 per 100,000** (474 to 524)			
Interval breast cancer	Moderate		**RR 0.78** (0.64 to 0.95)	235708 (3 cohort studies)[3,6,7]	⨁⨁⨁◯ MODERATE[b,d]
	220 per 100,000[c]	**172 per 100,000** (141 to 209)			
Recall for assessment	Moderate		**RR 1.08** (0.99 to 1.18)	303526 (4 cohort studies)[1,2,6,7]	⨁⨁◯◯ LOW[b,e,f]
	6,030 per 100,000[c]	**6512 per 100,000** (5,970 to 7,115)			

The GDG discussed the increased rate of false positives and noted that the likelihood of recall and biopsy will vary based on the setting. There is possibly a higher rate of biopsies with double reading with consensus or arbitration in certain settings. In other settings, the GDG notes that many false positives will receive additional imaging, as the next step after double reading with consensus or arbitration and they will not immediately have a biopsy or will have no biopsy at all, since additional imaging was sufficient.

The GDG notes that double reading with consensus or arbitration has more benefits when readers are not highly experienced (7) whereas when both readers read 5 000 or more mammograms per year double reading with consensus or arbitration increases recalls and false positive results (5)(6)(8).

False positive screening result	Moderate		**RR 1.08** (0.98 to 1.19)	303526 (4 cohort studies) [1,2,6,7]	⨁⨁◯◯ LOW[b,e,f]
	5,540 per 100,000[c]	**5983 per 100,000** (5,429 to 6,593)			
Positive predictive value	Low		**RR 0.99** (0.95 to 1.03)	19722 (4 cohort studies) [1,2,6,7,g]	⨁⨁◯◯ LOW[b,e,f]
	7,700 per 100,000	**7623 per 100,000** (7,315 to 7,931)			
Breast cancer stage in situ	Low		**RR 1.10** (0.95 to 1.28)	152580 (2 cohort studies) [2,5]	⨁⨁◯◯ LOW[b,h,i]
	43 per 100,000	**47 per 100,000** (41 to 55)			
Breast cancer stage I	Low		**RR 1.04** (0.95 to 1.14)	152580 (2 cohort studies) [2,5]	⨁⨁◯◯ LOW[b,h,i]
	132 per 100,000	**137 per 100,000** (125 to 150)			
Breast cancer stage II	Low		**RR 1.03** (0.91 to 1.17)	152580 (2 cohort studies) [2,5]	⨁⨁◯◯ LOW[b,h,i]
	74 per 100,000	**76 per 100,000** (67 to 87)			
Breast cancer stage III	Low		**RR 1.06** (0.82 to 1.36)	152580 (2 cohort studies) [2,5]	⨁⨁◯◯ LOW[b,h,i]
	17 per 100,000	**18 per 100,000** (14 to 23)			
Breast cancer stage IV	Low		**RR 1.00** (0.61 to 1.63)	152580 (2 cohort studies) [2,5]	⨁◯◯◯ VERY LOW[b,h,i]
	5 per 100,000	**5 per 100,000** (3 to 8)			

Breast cancer mortality	Study population		not estimable	(studies)	-
	0 per 1,000	**0 per 1,000** (0 to 0)			
Cumulative incidence of advanced cancers after a negative test	Study population		not estimable	(studies)	-
	0 per 1,000	**0 per 1,000** (0 to 0)			

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.

a. 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).
b. Unclear information about the used reference standard. Likely to be consistent with population screening programs.
c. Basal risk is estimated from the median probability of the event detected by single reading mammography among the included studies.

	d. The intervention tested may differ from the intervention of interest: one study used only consensus (Duijm 2009), oney 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). e. 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). f. Effect estimate crosses the no effect threshold precluding affirming which intervention is more favourable. g. Total number of detected cancers (1350) as a proportion of women that were recalled (17434) h. 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. i. Small number of events (Leivo 1999, Posso 2016).	

Desirable Effects

How substantial are the desirable anticipated effects?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ Trivial ● Small ○ Moderate ○ Large ○ Varies ○ Don't know	See research evidence for test accuracy	The GDG notes that the test accuracy shows 1 more true positive detected with double reading with consensus or arbitration compared to single reading per 1 000 women screened. The GDG notes that other outcomes demonstrate an increase in breast cancer detection and decreased interval cancers detected on follow-up with double readings. The GDG notes that the effects identified are presented per mammography exam; therefore this increases the impact of the increased detection and decreased interval cancers per woman due to women having multiple mammograms. The GDG did not reach consensus and therefore voting

		was conducted. Among 20 GDG members eligible to vote, results were: 10 members voted 'small'; 9 members voted 'moderate'; 1 member abstained.

Undesirable Effects

How substantial are the undesirable anticipated effects?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ Large ○ Moderate ● Small ○ Trivial ○ Varies ○ Don't know	See research evidence for test accuracy	The GDG notes that the rate of additional false positives identified in double reading with consensus or arbitration compared to single reading were 443 per 100 000 screening mammograms. Nonetheless, the GDG felt that this number constitutes a small undesirable effect. The GDG agreed by consensus that the undesirable anticipated effects were small.

Certainty of the evidence of test accuracy

What is the overall certainty of the evidence of test accuracy?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ Very low ○ Low ● Moderate ○ High ○ No included studies	-	The GDG notes that the certainty of the evidence for true positives with breast cancer and false negatives was downgraded due to indirectness. This was due to the fact that screen film mammography was used rather than digital mammography in the studies assessed. The GDG also notes that downgrading was considered for risk of bias, as readings were independently screened in one study, not true blinding. The GDG notes that the direction of bias would like bias towards the null hypothesis. The GDG also notes that the definition of interval cancers were different in the studies as one study only looked at one screening round while the others looked at first and subsequent rounds of screening. The GDG agreed by consensus that the overall certainty of the evidence of test accuracy was moderate.

Certainty of the evidence of test's effects

What is the overall certainty of the evidence for any critical or important direct benefits, adverse effects or burden of the test?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ Very low ○ Low ○ Moderate ○ High ● No included studies	-	The GDG does not expect any difference in the direct benefits or harms from the test to women, whether it is double or single reading. The GDG judged that women may be reassured if they are aware that double reading on the mammogram will be performed. The GDG agreed that false positives were not considered a direct effect of the test. **Qualitative Evidence** (14) reported the results from a qualitative interview in 48 women from a randomly selected sample of women who were invited to attend organised breast cancer screening in 13 French departments between 2010 and 2011. 27 women chose the organised screening programme, which they considered to be trustworthy, as negative mammograms are double checked by a second radiologist. 21 women preferred individual screening, which they considered to be more reliable, less anonymous and providing them with more liberty to take control of their own health. The GDG also noted that the only harm may be that there is a time delay in women obtaining their results in double reading with consensus or arbitration compared to single reading. The GDG agreed by consensus there were no included studies.

Certainty of the evidence of management's effects

What is the overall certainty of the evidence of effects of the management that is guided by the test results?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ Very low ○ Low ● Moderate ○ High ○ No included studies	-	The GDG agreed by consensus that the overall certainty of the effects of management were moderate.

Certainty of the evidence of test result/management

How certain is the link between test results and management decisions?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ Very low ○ Low ○ Moderate ● High ○ No included studies	-	The GDG agreed by consensus that the overall certainty of the link between test results and management were high.

Certainty of effects

What is the overall certainty of the evidence of effects of the test?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ Very low ○ Low ● Moderate ○ High ○ No included studies	-	The GDG agreed by consensus that the overall certainty of effects were moderate, as the certainty of test accuracy results was moderate, there was moderate certainty in the treatment effects and high certainty with regards to link between the test results and the management decisions. The direct consequences of the test were not considered by the GDG to be a decisive element here.

Values

Is there important uncertainty about or variability in how much people value the main outcomes?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ Important uncertainty or variability ○ Possibly important uncertainty or variability ● Probably no important uncertainty or variability ○ No important uncertainty or variability ○ No known undesirable outcomes	-	The GDG agreed by consensus that there was probably no important uncertainty or variability in values by women.

Balance of effects

Does the balance between desirable and undesirable effects favor the intervention or the comparison?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ Favors the comparison ○ Probably favors the comparison ○ Does not favor either the intervention or the comparison ● Probably favors the intervention ○ Favors the intervention ○ Varies ○ Don't know	-	The GDG did not reach consensus and therefore voting was conducted. Among 20 GDG members eligible to vote, results were: 15 members voted 'probably favours intervention'; 2 members voted 'does not favour either the intervention or the comparison'; 2 members voted 'favours the intervention' and 1 member abstained.

Resources required

How large are the resource requirements (costs)?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ Large costs ○ Moderate costs ○ Negligible costs and savings ○ Moderate savings ○ Large savings ● Varies ○ Don't know	**Double reading vs. single reading (costs and resources used from the societal perspective).**	The GDG noted that due to the fact that two of the studies were from before the time of digital mammography (twenty years old), the resources evidence is very indirect. Therefore, the GDG did not consider the evidence from (15) and (16). The GDG considered only (17) for resource requirement evidence. The GDG notes that increased costs observed may be due to both additional costs of reading and for additional assessments required as a result of increased detection of false positives. The GDG also discussed that the (17) study includes early recall as a cost, which impacts the total costs of double reading. Cost of consensus and arbitration (radiologists' time and administrative costs) is included in double reading which is why the costs of double reading are higher than single reading. In (17) study, approximately 6% of examinations went to consensus or arbitration. GDG discussed that these figures are different, and slightly lower, in other European countries.

Study ID	Country	Year-value	Double reading with consensus (USD-PPP)	Single reading (USD-PPP)	Incremental cost (USD-PPP)	Quality
Reading cost per 1000 mammograms						
Brown1996	UK	1994	1036	0	1036	**Moderate**[a]
Leivo1999	Finland	1995	6178	0	6178	
Additional assessment cost per 1000 mammograms*						
Brown1996	UK	1994	2745	4510	-1765	**Low**[b]
Leivo1999	Finland	1995	51741	50163	1578	
Total screening cost per 1000 mammograms						
Brown1996	UK	1994	3781	4510	-729	**Low**[b]
Leivo1999	Finland	1995	57919	50163	7756	

[a] The quality is moderate due to indirectness. Both studies were conducted 20 years ago.
[b] The quality is low due to indirectness and imprecision. Both studies were conducted 20 years ago and show contradictory results.
*Costs from societal perspective includes working time and transportation.

Characteristics of the studies

Study	N of mammograms	Type of screening	Age	Screening-year	Type of mammography	Type of reader	Single reading unitary cost	Double reading unitary cost	Single reading recall rate	Double reading recall rate
Brown1996	33734	prevalent	50-64	1987-1991	screen-film	radiologist	Sc1	Sc1 + 1,04	6.9 %	4.2%
Leivo1999	95423	overall	50-59 (can be continued up to 64)	1990-1995	screen-film	radiologist	6	11	NR	NR

Brown J, Bryan S, Warren R (1996) Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. BMJ 312:809–812.
Leivo T, Salminen T, Sintonen H, et al (1999) Incremental cost-effectiveness of double-reading mammograms. Breast Cancer Res Treat 54:261–267.

References

1) (15) Brown J, et al. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. BMJ. 1996 Mar 30; 312(7034):809–12.

2) (16) Leivo T, et al. Incremental cost-effectiveness of double-reading mammograms. Breast Cancer Res Treat. 1999; 54(3): 261–7.

Double reading vs. single reading (costs and resources used from the Health System perspective).

The GDG notes that the reading and reporting time for one digital mammography in screen-reading setting is in average 33 to 48 seconds (18) (19) (20), which is lower than the time needed for consensus, arbitration or recall assessments.

The GDG agreed by consensus that the costs will vary according to the setting, but noted that they will always be greater with double reading than single reading. The GDG agreed that the actual cost of the reading of mammograms (whether it is double or single) is probably small with regards to the total cost of the screening programme. The GDG noted that the proportionate cost increase will vary and it may be negligible, moderate or large depending on the setting.

Study ID	Country	Year-value	Double reading with consensus (USD-PPP)	Single reading (USD-PPP)	Incremental cost (USD-PPP)	Quality
Reading cost per 1000 mammograms						
Brown1996	UK	1994	1036	0	1036	**Moderate**[a]
Leivo1999	Finland	1995	6178	0	6178	
Posso2016	Spain	2010	12283	4522	7761	
Additional assessment cost per 1000 mammograms						
Brown1996	UK	1994	2285	3754	-1469	**Low**[b]
Leivo1999	Finland	1995	NR	NR	1389	
Posso2016	Spain	2010	10020	5486	4534	
Total screening cost per 1000 mammograms						
Brown1996	UK	1994	3754	3781	-433	**Low**[b]
Leivo1999	Finland	1995	NR	NR	7568	
Posso2016	Spain	2010	22303	10295	12008	

[a] The quality is moderate due to indirectness. Two studies were conducted 20 years ago.
[b] The quality is low due to indirectness and imprecision. Two studies were conducted 20 years ago and one of them shows contradictory results.

Characteristics of the studies

Study	N of mammograms	Type of screening	Age	Screening-year	Type of mammography	Type of reader	Single reading unitary cost	Double reading unitary cost	Single reading recall rate	Double reading recall rate	Unitary cost of assessment (societal)
Brown1996	33734	prevalent	50-64	1987-1991	screen-film	radiologist	Sc1	Sc1 + 1,04	0,069	0,042	54
Leivo1999	95423	overall	50-59 (to 64)	1990-1995	screen-film	radiologist	6	11	NR	NR	NR
Posso2016	28636	overall	50-69	2009-2011	digital	radiologist	5	12	4,6%	4,8%	NA

Brown J, et al (1996) Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. BMJ 312:809–812.
Leivo T, et al (1999) Incremental cost-effectiveness of double-reading mammograms. Breast Cancer Res Treat 54:261–267.
Posso M, et al . Cost-Effectiveness of Double Reading versus Single Reading of Mammograms in a Breast Cancer Screening Programme. PLoS One. 2016 Jul 26;11(7):e0159806.

References

1) (15) Brown J, et al. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. BMJ. 1996 Mar 30; 312(7034):809–12.

2) (16) Leivo T, et al. Incremental cost-effectiveness of double-reading mammograms. Breast Cancer Res Treat. 1999; 54(3): 261–7.

3) (17) Posso M, et al. Cost-Effectiveness of Double Reading versus Single Reading of Mammograms in a Breast Cancer Screening Programme. PLoS ONE. Public Library of Science; 2016; 11(7):e0159806.

Certainty of evidence of required resources

What is the certainty of the evidence of resource requirements (costs)?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ Very low ● Low ○ Moderate ○ High ○ No included studies	The quality is probably low due to indirectness and imprecision. Two studies were conducted 20 years ago and one of them shows contradictory results. Only one study was performed based on digital mammography screening and it was conducted in Spain.	The GDG notes that this relates to the (17) study in Spain only.

Cost effectiveness

Does the cost-effectiveness of the intervention favor the intervention or the comparison?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ Favors the comparison ○ Probably favors the comparison ○ Does not favor either the intervention or the comparison ○ Probably favors the intervention ○ Favors the intervention ● Varies ○ No included studies	**Cost-effectiveness per detected cancer (double vs. single reading)**	The GDG suggests consideration of local cost effectiveness data for application to different settings. Research evidence was only identified for Spain (17). The GDG noted that in Europe a common fixed threshold for cost-effectiveness is not used. The GDG therefore agreed that the cost-effectiveness varies. In some settings it may not be cost-effective.

Study ID	Country	Year-value	Incremental cost (USD-PPP)	Icremental effect (cancers detected)	ICER	Quality
ICER per cancer detected (societal perspective)						
Brown1996	UK	1994	-7287	30	Dominant	**Low**[a]
Leivo1999	Finland	1995	740160	29	25523	
ICER per cancer detected (health system perspective)						
Brown1996	UK	1994	-4327	30	Dominant	**Low**[a]
Leivo1999	Finland	1995	722130	29	24901	
Posso2016	Spain	2010	252788	11	22981	

[a] The quality is low due to indirectness and imprecision. Two studies were conducted 20 years ago and one of them shows contradictory results.

Characteristics of the studies

Study	N of mammograms	Type of screening	Age	Screening-year	Type of mammography	Type of reader	Single reading unitary cost	Double reading unitary cost	Single reading recall rate	Double reading recall rate	Unitary cost of assessment (societal)
Brown1996	33734	prevalent	50-64	1987-1991	screen-film	radiologist	Sc1	Sc1 + 1,04	0,069	0,042	54
Leivo1999	95423	overall	50-59 (to 64)	1990-1995	screen-film	radiologist	6	11	NR	NR	NR
Posso2016	28636	overall	50-69	2009-2011	digital	radiologist	5	12	4,6%	4,8%	NA

Brown J, et al (1996) Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. BMJ 312:809–812.
Leivo T, et al (1999) Incremental cost-effectiveness of double-reading mammograms. Breast Cancer Res Treat 54:261–267.
Posso M, et al . Cost-Effectiveness of Double Reading versus Single Reading of Mammograms in a Breast Cancer Screening Programme. PLoS One. 2016 Jul 26;11(7):e0159806.

References

1) (15) Brown J, et al. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. BMJ. 1996 Mar 30; 312(7034):809–12.

2) (16) Leivo T, et al. Incremental cost-effectiveness of double-reading mammograms. Breast Cancer Res Treat. 1999; 54(3): 261–7.

3) (17) Posso M, et al. Cost-Effectiveness of Double Reading versus Single Reading of Mammograms in a Breast Cancer Screening Programme. PLoS ONE. Public Library of Science; 2016; 11(7):e0159806.

Equity

What would be the impact on health equity?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ Reduced ○ Probably reduced ● Probably no impact ○ Probably increased ○ Increased ○ Varies ○ Don't know	-	The GDG agreed by consensus that there would probably be no impact on health equity.

Acceptability

Is the intervention acceptable to key stakeholders?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ No ○ Probably no ○ Probably yes ● Yes ○ Varies ○ Don't know	-	The GDG judged that patients would likely find this intervention acceptable. The GDG judged that certain radiologists and clinicians may not find double reading with consensus or arbitration acceptable. The GDG judged that policy-makers would likely find this acceptable as evidenced by its widespread use in current practice.

Feasibility

Is the intervention feasible to implement?

JUDGEMENT	RESEARCH EVIDENCE	ADDITIONAL CONSIDERATIONS
○ No ○ Probably no ○ Probably yes ● Yes ○ Varies ○ Don't know	-	The GDG judged by consensus that it would likely be feasible to implement. The GDG notes that in some settings, capacity (human resources of mammography readers) may make the feasibility of performing double reading with consensus or arbitration more challenging.

SUMMARY OF JUDGEMENTS

	JUDGEMENT						
PROBLEM	No	Probably no	Probably yes	**Yes**		Varies	Don't know
TEST ACCURACY	Very inaccurate	Inaccurate	**Accurate**	Very accurate		Varies	Don't know
DESIRABLE EFFECTS	Trivial	**Small**	Moderate	Large		Varies	Don't know
UNDESIRABLE EFFECTS	Large	Moderate	**Small**	Trivial		Varies	Don't know
CERTAINTY OF THE EVIDENCE OF TEST ACCURACY	Very low	Low	**Moderate**	High			No included studies
CERTAINTY OF THE EVIDENCE OF TEST'S EFFECTS	Very low	Low	Moderate	High			**No included studies**
CERTAINTY OF THE EVIDENCE OF MANAGEMENT'S EFFECTS	Very low	Low	**Moderate**	High			No included studies
CERTAINTY OF THE EVIDENCE OF TEST RESULT/MANAGEMENT	Very low	Low	Moderate	**High**			No included studies
CERTAINTY OF EFFECTS	Very low	Low	**Moderate**	High			No included studies
VALUES	Important uncertainty or variability	Possibly important uncertainty or variability	**Probably no important uncertainty or variability**	No important uncertainty or variability			No known undesirable outcomes
BALANCE OF EFFECTS	Favors the comparison	Probably favors the comparison	Does not favor either the intervention or the comparison	**Probably favors the intervention**	Favors the intervention	Varies	Don't know
RESOURCES REQUIRED	Large costs	Moderate costs	Negligible costs and savings	Moderate savings	Large savings	**Varies**	Don't know
CERTAINTY OF EVIDENCE OF REQUIRED RESOURCES	Very low	**Low**	Moderate	High			No included studies
COST EFFECTIVENESS	Favors the comparison	Probably favors the comparison	Does not favor either the intervention or the comparison	Probably favors the intervention	Favors the intervention	**Varies**	No included studies
EQUITY	Reduced	Probably reduced	**Probably no impact**	Probably increased	Increased	Varies	Don't know
ACCEPTABILITY	No	Probably no	Probably yes	**Yes**		Varies	Don't know

	JUDGEMENT						
FEASIBILITY	No	Probably no	Probably yes	**Yes**		Varies	Don't know

TYPE OF RECOMMENDATION

Strong recommendation against the intervention	Conditional recommendation against the intervention	Conditional recommendation for either the intervention or the comparison	**Conditional recommendation for the intervention**	Strong recommendation for the intervention
○	○	○	●	○

CONCLUSIONS

Recommendation

The ECIBC's Guidelines Development Group suggests using double reading (with consensus or arbitration for discordant readings) over single reading to screen mammograms for early detection of breast cancer in mammography screening programmes (conditional recommendation, moderate certainty of the evidence).

Justification

Overall justification

The GDG suggests by consensus that double reading (with consensus or arbitration) over single reading be used to diagnose breast cancer in mammography screening.

Only one study with digital mammography was included in the evidence, which limited the GDG to be able to issue a strong recommendation.

Detailed justification

Test accuracy

The GDG notes that most studies reviewed did not use digital mammography. The GDG notes that there is a higher sensitivity of breast cancer detection with double reading with consensus or abitration in mammography screening.

Desirable Effects

The GDG judged that the desirable anticipated effects, including additional breast cancers detected were small.

Undesirable Effects

The GDG judged that the undesirable anticipated effects, including additional false positive screening results identified, were small.

Certainty of the evidence of test accuracy

The GDG notes that only one study of digital mammography was identified. The other identified studies were published based on screen-film mammography since double reading with consensus or arbitration has been

standard practice in many settings for a number of years. The GDG notes that indirectness of the evidence was a concern as the studies included did not use digital mammography as is used in current practice.

Resources required
The GDG judged that the resources required will vary, but noted that they will always be greater with double reading with consensus or arbitration than single reading. The GDG noted that the proportionate cost increase will vary and it may be negligible, moderate or large depending on the setting. The GDG notes that increased costs observed may be due to both additional costs of reading and for additional assessments required as a result of increased detection or false positives.

Subgroup considerations

1. The GDG notes that in the context of double reading with consensus or arbitration, no differences were observed in accuracy when arbitration or consensus or both were used to reconcile differences in interpretation between mammography readers.

Implementation considerations

1. In settings with many low-volume mammography readers, the balance of benefits and harms may be even greater. The GDG refers readers to the PICO Question 7: 'What is the optimal annual interpretive volume for radiologists reading screening mammograms?' in the CCIB report, addressed by the QASDG regarding the experience level of mammography readers.

2. In some settings, capacity (human resources of mammography readers) should be scaled up to implement double readings. In settings where double readings are already in practice, the GDG suggests continued use of double readings with consensus or arbitration.

3. The GDG notes that a consideration that can favour double reading with consensus or arbitration is in those settings with many low volume mammography readers; the desirable effects of double reading with consensus or arbitration were found to be greater with less undesirable effects in these settings as compared to high volume mammography reader settings.

Monitoring and evaluation

1. The GDG suggests reporting the proportion of double reading with consensus or arbitration of mammograms that occur in practice. The GDG refers this suggestion to the QASDG for consideration.

Research priorities

1. The GDG suggests further research examining the cost-effectiveness of double vs single reading of digital mammography in different settings. Cost-effectiveness data was only identified for Spain.

2. The GDG suggests new research using observational studies comparing double reading with consensus or arbitration with single reading in the context of digital mammography. Additional research could also be performed to assess accuracy within the context of double readings assessing a single reader vs with the addition of a second reader, which is performed in practice.

3. The GDG suggests the use of formal radiologist blinding in research to improve the quality of evidence on double vs single readings.

4. The GDG notes that newer screening strategies such as digital breast tomosynthesis (DBT) or automatic computer assisted detection (CAD) was excluded from the analysis of this question with double vs single mammography. Future research could assess the impact of double reading using CAD and/or DBT systems.

REFERENCES SUMMARY

1. Ferlay, J, Soerjomataram, I , Ervik, M, Dikshit, R , Eser, S , Mathers, C, Rebelo, M, Parkin, DM, Forman, D, Bray, F. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide. 2013.
2. Ferlay J, Steliarova-Foucher E,Lortet-Tieulent J,Rosso S,Coebergh JW,Comber H,Forman D,Bray F.. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. Eur J Cancer; 2013.
3. Tsilidis KK, Papadimitriou N,Capothanassi D,Bamia C,Benetou V,Jenab M,Freisling H,Kee F,Nelen A,O'Doherty MG,Scott A,Soerjomataram I,Tjonneland A,May AM,Ramon Quiros J,Pettersson-Kymmer U,Brenner H,Schottker B,Ordonez-Mena JM,Karina Dieffenbach A,Eriksson S,Bogeberg Mathiesen E,Njolstad I,Siganos G,Wilsgaard T,Boffetta P,Trichopoulos D,Trichopoulou A.. Burden of Cancer in a Large Consortium of Prospective Cohorts in Europe. J Natl Cancer Inst; 2016.
4. Warren RM, Duffy SW,Bashir S. The value of the second view in screening mammography. Br J Radiol; 1996 .
5. 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.
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. Tonita JM, Hillis JP,Lim CH.. Medical radiologic technologist review: effects on a population-based breast cancer screening program. Radiology; 1999.
8. 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.
9. Blanch J, Sala M,Román M,Ederra M,Salas D,Zubizarreta R,Sanchez M,Rué M,Castells X, group., CFPR. Cumulative risk of cancer detection in breast cancer screening by protocol strategy.. Breast Cancer Res Treat; 2013.
10. Román R, Sala M,Salas D,Ascunce N,Zubizarreta R,Castells X, Group., Cumulative,False,Positive,Risk. Effect of protocol-related variables and women's characteristics on the cumulative false-positive risk in breast cancer screening.. Ann Oncol; 2012.
11. Blanks RG, Wallis MG,Moss SM.. A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: results from the UK National Health Service breast screening programme. J Med Screen; 1998.
12. Nyaga VN. Metaprop: a Stata command to perform meta-analysis of binomial data. Arch Public Health; 2014.
13. Takwoingi Y. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. Stat Meth Med Res; 2017.
14. Kalecinski J, Régnier-Denois V,Ouédraogo S,Dabakuyo-Yonli TS,Dumas A,Arveux P,Chauvin F.. [Organized or individual breast cancer screening: what motivates women?]. Sante Publique; 2015.
15. Brown J, Bryan S,Warren R. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. BMJ; 1996.
16. 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.
17. Posso M, Carles M,Rué M,Puig T,Bonfill X.. Cost-Effectiveness of Double Reading versus Single Reading of Mammograms in a Breast Cancer Screening Programme.. PloS One; 2016.
18. Bernardi D, Ciatto S,Pellegrini M,Anesi V,Burlon S,Cauli E,Depaoli M,Larentis L,Malesani V,Targa L,Baldo P,Houssami N.. Application of breast tomosynthesis in screening: incremental effect on mammography acquisition and reading time.. Br J Radiol; 2012.
19. Gilbert FJ, Tucker L,Gillan MG,Willsher P,Cooke J,Duncan KA,Michell MJ,Dobson HM,Lim YY,Purushothaman H,Strudley C,Astley SM,Morrish O,Young KC,Duffy SW.. The TOMMY trial: a comparison of TOMosynthesis with digital MammographY in the UK NHS Breast Screening Programme--a multicentre retrospective reading study comparing the diagnostic performance of digital breast tomosynthesis and digital mammography with digital mammography alone.. Health Technol Assess; 2015.
20. Skaane P, Bandos AI,Gullien R,et al.. Prospective trial comparing full-field digital mammography (FFDM) vs. combined FFDM and tomosynthesis in a population-based screening programme using independent double reading with arbitration.. Eur Radiol; 2013.