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-SUPPRT@ec.europa.eu

ECIBC recommendation on:

Optimal number of readings for allowing mammography readers to work in mammography screening programmes

Healthcare question	Should an optimal number of readings vs. no specific number be used for allowing mammography readers to work in mammography screening programmes?
Population	Mammography readers working in mammography screening programmes
Intervention	Optimal number of readings
Comparison	no specific number
Main outcomes	False positive rate, sensitivity, breast cancer detection, recall rate
Setting	European Union
Perspective	Population (national health system)
Background	Breast cancer is the most common cancer in women world-wide. Early detection by screening is shown to reduce disease specific mortality. However, the service needs to be of high quality in order to balance the benefits and harms. In order to maintain and increase the quality in breast cancer services it seems important to establish educational schemes/recommendations/guidelines for the professionals dealing with screening mammography. One aspect is reader experience needed to continue screen reading to maintain a high quality level. The clinical questions initially formulated by the panel asked about the minimum number of readings to either start working or continuing as a mammography reader in a mammography screening programme. No evidence was identified to directly answer these questions. Research evidence did not distinguish mammography readers as those starting to work or already working in screening programmes. In addition, the results revealed that rather than a minimum number of readings, a range of optimal number of readings (minimum and maximum) seems to be the answer to these questions. In consequence three profiles have been developed to answer the following questions: - Should a minimum number of readings compared to no specific number be used to allow mammography readers to work in mammography screening programmes? - Should a maximum number of readings compared to no specific number be used to allow mammography readers to work in mammography screening programmes? - Should an optimal number of readings compared to no specific number be used to allow mammography readers to work in mammography screening programmes? Due to heterogeneity in interpretative volume strata and unavailability of essential data, it was not possible to pool the reported measures of effect (i.e. Odds Ratio) as

	planned in the protocol. Thus, the point estimates for false positive rate, recall rate, sensitivity and breast cancer detection were plotted. The overall slope, slope for each individual study and change in slope according to different cut-off points were calculated for each performance parameter when possible, otherwise the results were interpreted by the visual inspection of the scatter plots. This question refers to reading mammographies by each individual mammography reader, but double reading has been recommended in breast cancer screening programmes Three outcomes were used to inform these questions: false positive rate, sensitivity and breast cancer detection. When feasible, specificity was used to obtain false positive rate however, specificity was not plotted as few studies reported it. Due to paucity of data, it was not possible to get conclusions from recall rate.
Conflict of interest	Conflict of interest (CoI) for all GDG members were assessed and managed by the European Commission Joint Research Centre (JRC) following an established procedure in line with the institutional rules. GDG member participation in the development of the recommendations was restricted, according to CoI disclosure. Consequently, for this particular question, Edoardo Colzani, Axel Gräwingholt and Elsa Perez were recused from voting. Miranda Langendam, as external expert, was also not allowed to vote, according to the ECIBC rules of procedure. More information
Date of the systematic review	30/10/2018 (evidence of effects) 28/02/2019 (economic evidence)

Assessment by the ECIBC Guidelines Development Group (GDG)

Problem: Is the problem a priority?						
Judgement	No	Probably no	Probably yes	Yes	Varies	Don't know
Research evidence	Breast cancer is the most common cancer in women worldwide. Early detection by screening is shown to reduce disease specific mortality. However, the service needs to be of high quality in order to balance the benefits and harms. In order to maintain and increase the quality in breast cancer services it seems important to establish educational schemes/recommendations/guidelines for the professionals dealing with screening mammography. One aspect is reader experience needed to continue screen reading to maintain a high quality level.					
Additional considerations	This question was prioritised by the GDG. The panel noted that reading was done without computer-aided support and the recommendation applies to each individual reader's performance. Double reading, however, is recommended for further quality control (see					

	https://ecibc.jrc.ec.europa.eu/recommendations/details/Patient/doublereading/mammoreader).

Desirable effects: How substantial are the desirable anticipated effects?

Judgement	Trivial	Small	Moderate	Large	Varies	Don't know

Research evidence

Outcomes	Anticipated absolute effects* (95% CI)		Relative effect (95% CI)	№ of participants (studies)	Certainty of the evidence (GRADE)
	Risk with no specific number	Risk with an optimal number of readings			
False positive rate assessed with: defined as the proportion of positive screening examinations among all screening exams without a breast cancer diagnosis within the follow-up period.	The increasing number of readings is related with an overall decreasing of false positive rate considering all data from the 7 studies together. (linear R2=0.4951; p<0.001). A cut-off ≤ 3500 annual readings is associated with a statistically significant reduction of an absolute 1.35% false positives every 1000 readings increase of experience (p=0.01) whereas a cut-off ≥ 3500 annual readings is associated with a non-significant decreasing of an absolute 0.13%		-	(7 observational studies)[1,2,3,4,5,6,7,ab]	⨁◯◯◯ VERY LOW[c,d,e]

	false positives every 1000 readings increase of experience (p=0.126)			
Sensitivity assessed with: Proportion of screening examinations interpreted as positive (BIRADS 0, 4, 5 or referred for assessment) among all screening exams who were given a diagnosis of BC within the follow-up period	The visual inspection of the scatter plot reveals that the increasing number of readings is related with an overall increasing of sensitivity up to approximately 7000 annual readings (between 0.893 and 0.939). Over this number of readings it appears to be a decrease of sensitivity.	-	(6 observational studies)[2,3,4,6,7,8,a,b]	⊕◯◯◯ VERY LOW[e,f,g]
Breast cancer detection assessed with: defined as the number of cancers detected per 1000 screening exams. Cancer included all breast cancer (ductal carcinoma in situ or invasive)	The visual inspection of the scatter plot reveals that the increasing number of readings is related with an overall increasing of breast cancer detection up to approximately 7000 annual readings (between 5.05 and 8.3 per 1000 screening exams). Over this number of readings, it appears to be a decrease of the breast cancer detection.	-	(4 observational studies)[2,6,7,8,a,b]	⊕◯◯◯ VERY LOW[e,f,g]

Recall rate - not measured	The included studies did not assess the outcome of recall rate.-

1. Alberdi RZ, Llanes AB Ortega RA Expósito RR Collado JM Verdes TQ Ramos CN Sanz ME Trejo DS Oliveres XC, group., CFPR (Cumulative False Positive Risk). Effect of radiologist experience on the risk of false-positive results in breast cancer screening programs. Eur Radiol; 2011.
2. Hoff SR, Myklebust TÅ,Lee CI,Hofvind S.. Influence of Mammography Volume on Radiologists' Performance: Results from BreastScreen Norway.. Radiology; 2019.
3. Elmore JG, Jackson SL,Abraham L,Miglioretti DL,Carney PA,Geller BM,Yankaskas BC,Kerlikowske K,Onega T,Rosenberg RD,Sickles EA,Buist DS.. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy.. Radiology; 2009.
4. Duncan KA, Scott NW. Is film-reading performance related to the number of films read? The Scottish experience. Clinical Radiology; 2011.
5. Théberge I, Hébert-Croteau N,Langlois A,Major D,Brisson J. Volume of screening mammography and performance in the Quebec population-based Breast Cancer Screening Program. CMAJ; 2005.
6. Théberge I, Chang SL,Vandal N,Daigle JM,Guertin MH,Pelletier E,Brisson J.. Radiologist interpretive volume and breast cancer screening accuracy in a Canadian organized screening program. J Natl Cancer Inst; 2014.
7. Buist DS, Anderson ML,Haneuse SJ,Sickles EA,Smith RA,Carney PA,Taplin SH,Rosenberg RD,Geller BM,Onega TL,Monsees BS,Bassett LW,Yankaskas BC,Elmore JG,Kerlikowske K,Miglioretti DL.. Influence of annual interpretive volume on screening mammography performance in the United States. Radiology; 2011.
8. Cornford, E. J., Turnbull, A. E., James, J. J., Tsang, R., Akram, T., Burrell, H. C., Hamilton, L. J., Tennant, S. L., Bagnall, M. J., Puri, S., Ball, G. R., Chen, Y., Jones, V.. Accuracy of GE digital breast tomosynthesis vs supplementary mammographic views for diagnosis of screen-detected soft-tissue breast lesions. Br J Radiol; 2016.

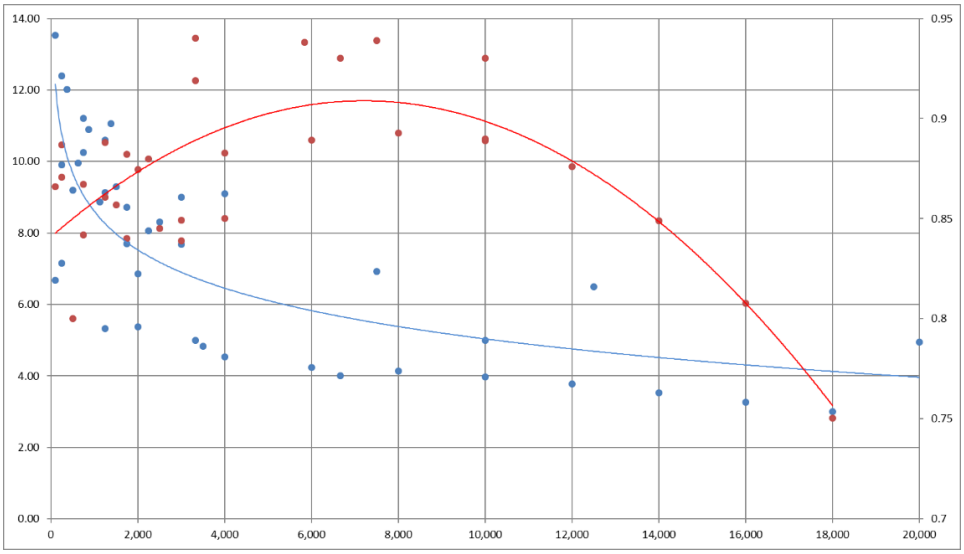
a. The research evidence considered refers to organised screening programs.
b. Hoff et al. 2019 has used data from the Cancer Registry of Norway. The interpretation and reporting of these data are the sole responsibility of the authors, and no endorsement by the Cancer Registry of Norway is intended nor should be inferred.
c. 2 out of 7 studies did not show a statistically significant relationship (trend) between false positive rate and number of readings according to their analysis. According to our slope analysis 4 out of 7 studies do not show a statistically significant relationship (trend).
d. The available data is non-weighted and with no estimation of the 95%CI. The overall slope for false positives and number of readings is statistically significant. A cut-off was selected when the slope above this point was no longer significant (flat slope or relationship).
e. All studies were retrospective and these outcomes were not adjusted for

relevant confounding variables. Number of readings was mostly self-reported or not accounting for professionals' experience in other settings.

f. According to visual inspection of the scatter plot only 2 studies were compatible with the estimation of a maximum number of readings (Cornford 2011 and Hofvind - unpublished).

g. The summary of findings represent a visual interpretation of the scatter plot and no statistical inference was taken. The available data is non-weighted and with no estimation of the 95%CI.

Figure: Combined scatter plot for false positive rate and sensitivity



- X (horizontal) axis: number of readings
- Y (vertical) left axis: false positive rate (%)
- Y (vertical) right axis: sensitivity (proportion)For clearer visualisation the false positive rate scale rages from 0 to 14% and sensitivity ranges from 0.7 to 0.95, where all points are grouped.
- Red dots and curve: sensitivity
- Blue dots and curve: false positive rate

We extrapolated the point estimate for sensitivity at 3,500 annual readings (the estimated minimum according to false positive analysis) which was 0.89 and then we extrapolated the number of readings over 7,000 (when sensitivity peaked) associated to 0.89 sensitivity which as 11,000 annual readings.

The effect of differences in breast cancer incidence rate across countries in breast cancer detection was not ignored. The GDG considered that breast cancer incidence rate varies across countries and thus breast cancer detection was considered a less reliable indicator.

Additional considerations

Higher rates on breast cancer detection and sensitivity were considered desirable effects by the GDG.

The GDG notes that the intervention in this question is reading within a specific threshold and the alternative is reading without a specific suggested threshold, thus, there is benefit on average but the magnitude of those benefits varies depending on where the starting/comparator point is and how it compares to the intervention.

	The GDG agreed by consensus that the desirable anticipated effects of reading mammograms within a threshold vary.

Undesirable effects: How substantial are the undesirable anticipated effects?						
Judgement	Large	Moderate	Small	Trivial	Varies	Don't know
Research evidence	See research evidence for desirable effects					
Additional considerations	The GDG notes that increase in number of false positives would be an undesirable effect There is harm on average if the reader of mammograms is reading outside the suggested range but the undesirable effects, similar to what occurred with the desirable effects, vary depending on where the starting/comparator point is and how it compares to the intervention. The GDG therefore agreed by consensus that the undesirable anticipated effects varied.					

Certainty of evidence: What is the overall certainty of the evidence of effects?					
Judgement	Very low	Low	Moderate	High	No included studies
Research evidence	-				
Additional considerations	The GDG agreed that the overall certainty of the evidence is very low, due to the limitations in study design and analysis, inconsistency across included studies and imprecision. There were fewer data points in the higher range for mammography readings than in the lower range; therefore, the panel judged they were less certain for the upper threshold (11,000).				

Values: Is there important uncertainty about or variability in how much people value the main outcomes?					
Judgement	Important uncertainty or variability	Possibly important uncertainty or variability	Probably no important uncertainty or variability	No important uncertainty or variability	No known undesirable outcomes
Research evidence	No systematic review was carried out.				

Additional considerations	The GDG judged that there was probably no important uncertainty in how much women value the main outcomes

Balance of effects: Does the balance between desirable and undesirable effects favor the intervention or the comparison?							
Judgement	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
Research evidence	-						
Additional considerations	The range for reading mammograms was selected because the balance is most favourable there. Although the GDG was uncertain how large the benefit was, they agreed there was a net benefit. Therefore, the GDG judged that the balance probably favours the intervention						

Resources required: How large are the resource requirements (costs)?							
Judgement	Large costs	Moderate costs	Negligible costs and savings	Moderate savings	Large savings	Varies	Don't know

Research evidence

Reported costs of false positive results

Study	Country	Year value	Resources (1 year post index mammogram)	Resources cost	False-Positive Cost	Quality
Lee 2016	USA	2012 USD	Biopsy	1131	1131 (**only biopsy**)	Very low[1]
Ong 2015	USA	2012 USD	Add diagnostic mammography	166	766 (**sum all additional diagnostic tests**)	Very low[1]
			Other imaging*	150		
			Biopsy	344		
			Pathology	21		
			Outpatient consultation	17		
			Other	67		
Chubak 2010	USA	2007 USD	Breast imaging	338	338 (**median of all additional diagnostic tests**)	Very low[1]
			Surgery consult	628		
			Fine needle aspiration	642		
			Breast biopsy	2192		
			Mastectomy	6226		

USD= US Dollars.

[1]The study was conducted in the USA. The costs, resources used may not be applicable to European settings. Observational data was included. Screening produces false positive results, which may affect quality of life. The actual effect on quality of life varies in its extent and duration.

*Ultrasound, magnetic resonance imaging. Biopsy was performed in 29.3 % of the women, and 0.4 % of those women (301) underwent prophylactic mastectomy.

No studies identified for sensitivity or cancer detection rate

Additional considerations

In one study, false positive results occurred in 9.9% of women; most false positives (87.3%) were followed by breast imaging only. The study provided the median cost of all additional diagnostic tests as a consequence of a false positive which has been considered more accurate than the sum or the costs of biopsy only. In this study, a false positive was associated to a total median cost of 338 USD (1).

The GDG noted that there are some savings related to the decrease in false positives if readers remain within the proposed range, but these cannot be quantified. Since double reading is recommended for screening, the true costs related to the intervention are unknown because the modelling was done on single readings.

There may be increased costs if there are mammography readers who are reading above the range, as the centre would have to hire additional readers in order to remain within the proposed range. On the other hand, a higher number of mammograms will

	yield a higher income The GDG agreed by consensus that the resources required are unknown.

Certainty of evidence of required resources: What is the certainty of the evidence of resource requirements (costs)?					
Judgement	Very low	Low	Moderate	High	No included studies
Research evidence	-				
Additional considerations	The GDG agreed that the certainty of the resources evidence is very low, due to the quality of the studies included.				

Cost effectiveness: Does the cost-effectiveness of the intervention favor the intervention or the comparison?							
Judgement	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
Research evidence	No studies identified to directly answer this question						
Additional considerations	No studies were included.						

Equity: What would be the impact on health equity?							
Judgement	Reduced	Probably reduced	Probably no impact	Probably increased	Increased	Varies	Don't know
Research evidence	No systematic review was carried out.						
Additional considerations	For women the GDG agreed by consensus that health equity probably increases because of less variability in the quality of the readings and the health outcomes (random inequality may be reduced).						

Acceptability: Is the intervention acceptable to key stakeholders?						
Judgement	No	Probably no	Probably yes	Yes	Varies	Don't know
Research evidence	No systematic review was carried out.					
Additional considerations	The GDG judged that the acceptability for key stakeholders would be as follows: Healthcare professionals: the optimal range (including a maximum) might have limitations in acceptability among readers that work in several programmes as the total amount of annual readings may be above the maximum suggested in the range. For the lower threshold, this range should be acceptable because it is lower than the previously recommended threshold (5,000) (European Guidelines 4th Edition). The higher threshold is new and there were questions about the acceptability, which led the GDG to say "don't know". Also related to programs with lower reading volumes. Women: there would be no concerns. Policy-makers: The change in policy from prior threshold may limit the acceptability because of challenges with implementing new recommendations. Overall, the GDG agreed by consensus that the intervention would probably be acceptable to key stakeholders.					

Feasibility: Is the intervention feasible to implement?						
Judgement	No	Probably no	Probably yes	Yes	Varies	Don't know
Research evidence	No systematic review was carried out.					
Additional considerations	The GDG noted that there may be some barriers to implementation in some stakeholders: Policy-makers: It is feasible but the option might be more difficult to implement in small breast cancer screening programmes where the readers' annual number of readings is lower. Also, it might take time to change the number of readings to reach a higher or lower level. Overall, the GDG agreed by consensus that the intervention would be feasible					

Summary of judgements

	Judgements						
Problem	No	Probably no	Probably yes	Yes		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 evidence	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
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 (GDG) suggests that mammography readers read between 3,500 and 11,000 mammograms annually in organised mammography screening programmes (conditional recommendation, very low certainty of the evidence).**
Justification	The right balance between sensitivity (true positives) and specificity (false positives) was found within the selected range of readings per reader.
Subgroup considerations	None were considered by the GDG.
Implementation considerations	1. This optimal range should be implemented as part of screening programmes that use double reading (which is a conditional recommendation previously issued: https://ecibc.jrc.ec.europa.eu/recommendations/details/Patient/doublereading/mammoreader), as double reading of mammograms will improve the overall quality of the reading. 2. Efforts to provide support with centralised reading may help implementing this recommendation. 3. The number of readings should be averages over longer periods and use appropriate sampling for this measurement. 4. The GDG recognised that local circumstances of the individual's performance will affect the quality of readings as much as the number of readings. The suggested range applies to mammography screening, not tomosynthesis.
Monitoring and evaluation	Programmes should continue to monitor the number of readings by readers and the quality indicators of the reading process
Research priorities	1. More published evidence is required which could also come from good monitoring data. 2. Further research on when to read during the day and how many mammograms to read per day, time needed and variation of time for readings between readers, relationship between breast cancer detection and false positives, comparison between those reading only screening mammograms and those doing screening and clinical/diagnostic mammography too, as this may influence results. 3. More research on types of training for mammography readers 4. Research into the role of artificial Intelligence in dual reading.