BJR

https://doi.org/10.1259/bjr.20190610

Received: 12 July 2019 | Revised: 09 October 2019 | Accepted: 13 October 2019

Cite this article as:
Taylor-Phillips S, Stinton C. Double reading in breast cancer screening: considerations for policy-making. *Br J Radiol* 2020; **93**: 20190610.

REVIEW ARTICLE

Double reading in breast cancer screening: considerations for policy-making

SIAN TAYLOR-PHILLIPS, PhD and CHRIS STINTON

Warwick Medical School, University of Warwick, Coventry, CV4 7AL, England

Address correspondence to: Dr Sian Taylor-Phillips
E-mail: *s.taylor-phillips@warwick.ac.uk*

ABSTRACT

In this article, we explore the evidence around the relative benefits and harms of breast cancer screening using a single radiologist to examine each female's mammograms for signs of cancer (single reading), or two radiologists (double reading). First, we briefly explore the historical evidence using film-screen mammography, before providing an in-depth description of evidence using digital mammography. We classify studies according to which exact version of double reading they use, because the evidence suggests that effectiveness of double reading is contingent on whether the two radiologists are blinded to one another's decisions, and how the decisions of the two radiologists are integrated. Finally, we explore the implications for future mammography, including using artificial intelligence as the second reader, and applications to more complex three-dimensional imaging techniques such as tomosynthesis.

INTRODUCTION

Breast cancer is the leading cause of cancer deaths in females.[1] One pathway to reducing breast cancer mortality is early detection. To this end, many countries have implemented breast cancer screening programmes. The evidence for mortality benefits of breast cancer screening is mixed,[2] and there is considerable ongoing debate about the balance of benefits and harms of these programmes.[3,4] This debate extends into how to undertake mammography screening. In this article, we focus on the impact of a single reader (usually a breast radiologist) examining each female's mammograms, denoted henceforth as "single reading," in comparison to having two readers examine the womens mammograms, known as "double reading." Many European countries employ double reading, while in the United States a single reader plus computer-aided detection (CAD) is more common.[5–7] The European quality assurance guidelines for breast screening recommend double reading, with both readers blinded to the decision of the other whilst first assessing the female's mammograms, followed by resolving discordant results by consensus between the two readers or third reader arbitration.[7] They emphasize the greater importance of double reading when one of the readers is not yet specialized or examining sufficient volumes of mammograms. In this review, we explore the evidence comparing single and double reading, focusing on the impact of blinding and consensus methods for discordant results. We examine outcomes such as test accuracy and types of cancer detected, and consider how these may be related to the benefits and harms of screening. We also consider the implications for future iterations of breast screening, such as tomosynthesis and artificial intelligence (AI).

The past: double reading using film-screen mammography

The majority of research on the impact of double reading has come from studies using film mammography. In a systematic review and meta-analysis, Taylor and Potts reported that double reading increased recall rates relative to single reading.[8] Individual studies have suggested that 10–27% of the females who are recalled are done so by the second reader only.[9–11] The ultimate impact on recall rate for females attending screening will depend on whether females with discordant results (recalled by one reader and not the other) are recalled for further tests. Recall rates inevitably increase when recall occurs when it is indicated by either reader.[10] However, when consensus decisions or arbitration strategies are used, overall recall rates decrease.[12,13] Recall rate is an important marker for false-positive recalls, which is the most common harm from breast cancer screening, and should be minimized as far as possible. Females often experience anxiety in the wait for follow-up tests, which can persist even after receiving her final results that there was no cancer.[14] The mortality and morbidity benefits of screening are achieved through detecting cancer at an earlier stage when it is more amenable

to treatment, so cancer detection rate is linked to the benefits of breast cancer screening. However, some cancer types such as low grade ductal carcinoma *in situ* (DCIS) may be more associated with harm from overdiagnosis of clinically insignificant disease, so the type of cancers detected is an important measure of the potential benefits and harms of screening. In the aforementioned systematic review and meta-analysis, an increase in cancer detection rate was reported for double over single reading,[8] with individual studies reporting that 6–10% of cancers were not detected by the first reader.[9,10,12,15] Some studies have suggested that the additional tumours detected in double reader programmes are smaller (<15 mm),[16] and that the ratio of DCIS to invasive cancers is higher compared to single reader programmes.[17] Other studies have reported no statistically significant differences in the size or stage of cancers identified in single *vs* double reader strategies.[9,18] However, this might have been driven by a lack of power to detect differences in these small studies; there was a trend towards small cancers and a greater proportion of DCIS being detected by the second reader. Some studies indicate that double reading strategies require more resources than single reading strategies,[17,19,20] but this relationship is heavily dependent on the recall rates, because an assessment clinic takes much more radiologist time than the initial examination of the mammograms.

A limitation of the evidence-base is that in clinical practice film mammography has now been replaced by digital mammography.[21,22] Randomized controlled trials have indicated that cancer detection rates are similar between the two approaches,[23] or slightly higher for digital mammography than film mammography.[24] Subgroup analyses has shown that compared to film mammography, diagnostic accuracy of digital mammography is higher for females who are younger than 50, with dense breasts or pre-/perimenopausal.[25] Further, there is some evidence that the types of cancers missed by digital *vs* film mammography differ systematically.[26] On this basis, the benefits and harms of double reading for film-screen mammography may not be generalizable to digital mammography.

The present: double reading using digital mammography

Within the last 3 years, five studies (one each from Denmark, Germany, Netherlands, Spain, and the UK) have compared single *vs* double reading using digital mammography within breast cancer screening programmes.[27–31] There is considerable variability in the approaches and characteristics of the studies. For example, four studies were retrospective,[27–29,31] and one was prospective.[30] Data were collected over 1 year in two studies[27,29] and 2 years in three studies.[28,30,31] Study samples ranged from 25,579[31] to 805,206[27] females, who were between the ages of 50–69,[28,29,31] 50–75,[30] or 47–73.[27] Recall was conducted following a mixture of consensus,[31] consensus with arbitration,[28,29] when either reader recommended it,[30] or a mix of third reader arbitration, two reader consensus, and larger group consensus.[27] In each of the studies, data were collected from breast cancer screening programmes, with mammographic interpretation conducted by trained readers (predominantly radiologists) who read at least 5000[27–29,31] or at least 10,000[30] mammograms per year.

The effect that the number of readers has on recall has been reported in four studies, with recall rates ranged from 3.0–4.8% for single reader to 3.1–6.2% for double reader strategies.[27–30] Within individual studies there is variation in which strategy resulted in higher recall, with some studies suggesting that the proportion of females who are recalled for further testing is greater when using double reader strategies,[28,30] while others have suggested that recall rates are higher when using single reader strategies,[29] or that the results are mixed.[27] The variability can be explained by the range of different policies for managing discordant reader interpretations of mammograms. When recall was conducted as indicated by either one of the readers, double reading produced higher recall rates than single reading: 3.6% *vs* 3.0%[30] and 6.2% *vs* 4.8%.[27] Logically, this method of implementation will increase recall rates, as the second reader can only add to the tally of recalls. In studies that used consensus with/or arbitration, recall was generally (though not always)[28] lower for double reader than single reader strategies.[27,29] These strategies allow for further consideration of whether to recall, either by the same readers or others. The mechanism by which these systems impact on recall rates may be complex, explaining the heterogeneity. Individual readers may change their reading behaviour dependent on the strategy for discordant cases. For example, they may be more willing to recall more cases in a bid to increase cancer detection rates knowing they have a safety net of arbitration to remove inappropriate recalls. In general, it appears that double reading can reduce recall rates.

Cancer detection has been examined in five studies.[27–31] In each of these, cancer detection rate was somewhat higher for double (5.2–8.8 per 1000 screens) than single reading (4.8–8.0 per 1000 screens). Within-study differences in the cancer detection rates of single and double reading ranging from 0.10[31] to 0.83[30] per 1000 screens. There is no consistent pattern to indicate that any one approach to recall (consensus with/or arbitration, when either reader recommends it) leads to greater cancer detection. Similarly, sensitivity has been reported to be somewhat higher for double reading (72.0–94.8%) than single reading (65.5–87.8%),[28,29,31] and higher for double reading with either reader recalling (94.8%)[28] than double reading with consensus (72.0–79.9%).[29,31] In general, interval cancers appear to be less common with double reading (0.6–3.0 per 1000 screens) than single reading (0.9–6.1 per 1000 screens).[27–29,31]

Three recent studies have examined the characteristics of the cancers detected by different readers.[27,28,30] Posso and colleagues retrospectively analyzed data from 28,636 females (aged 50–59 years) who had participated in a biennial population-based breast cancer screening programme in Barcelona, Spain, and were followed up for 2 years.[28] Comparing the additional cancers detected only by the second reader to those detected by a single reader, there was little difference in the stage of cancers. However, the additional cancers detected only by the second reader had a greater frequency of DCIS (54.5% *vs* 20.5%), and more small (<10 mm) cancers (54.5% *vs* 37.7%), and fewer with lymph node involvement (9.1% *vs* 21.6%) than those detected by a single reader. In a prospective study of 99,013 consecutive females (aged 50–75 years) who were taking part in the Dutch

nationwide biennial breast cancer screening programme, and who were followed-up for 2 years,[30] there were few statistically significant differences between the characteristics of cancers detected by the first reader and the additional cancers detected by second readers (*i.e.* mammographic features, histology of invasive cancers, lymph node status, oestrogen receptor status, progesterone receptor status, or Her2/Neu receptor status). The only statistically significant difference was that of the cancers detected only by the second reader a greater proportion were of a lower tumour grade (59.0% *vs* 39.8%). These lower grade cancers are associated with a better prognosis. Like Posso et al,[28] a somewhat greater proportion of the additional cancers detected by second reader than the first reader were DCIS; double reading resulting in an increase in the proportion of DCIS by 19% compared to 12% for invasive cancers. Further, compared to the first reader the second reader tended to detect more low grade DCIS (23.8% *vs* 15.7% of all cancers detected), with a greater proportion of the invasive cancers being small (<20 mm, 86.9 *vs* 79.5% of all cancers detected). The trend towards second readers identifying smaller cancers and a higher ratio of DCIS to invasive cancers is supported in a large retrospective study using data from 805,206 females who had taken part in the English breast cancer screening programme over 1 year.[27] The authors found that cancers detected only by a second reader were more likely to be DCIS (30.5% *vs* 22.0%), and that they were of a lower grade (17.0% *vs* 8.9%). Further, the second readers identified invasive cancers that were significantly smaller (mean size 14.2 *vs* 16.7 mm) and involved few nodes (1–2 nodes: 12.6% *vs* 17.8%) compared to those identified by the first reader. The mammographic appearance of DCIS is commonly microcalcifications, so the second reader appears to be systematically picking up smaller mammographic features and microcalcifications, which may either have been missed in the first readers search, or dismissed as low risk by the first reader. Smaller low grade tumours and low grade DCIS have been associated with overdiagnosis (the detection of disease that would not have caused harm during a person's lifetime).[32–34] The balance of benefits and harms of these extra cancers is currently unclear, including what proportion represent very early detection resulting in morbidity and mortality benefit, and what proportion are simply harmed by overdiagnosis of disease which would not have become symptomatic within the female's lifetime. This is dependent on the characteristics of the cancer detected, and the age and health of the female screened.

A small number of recent studies have considered factors that might influence recall and cancer detection. The first of these factors is mammographic breast density. Breasts are made up of a mixture of fibrous and glandular tissue and fatty tissue. Dense breast tissue (with a high ratio of fibrous and glandular tissue to fatty tissue) can mask cancers on mammograms.[35] There is evidence to suggest that greater mammographic breast density is associated with lower programme sensitivity, and an increased risk of breast cancer (including interval cancers).[36,37] Von Euler-Chelpin et al reported a non-statistically significant decrease in sensitivity with each increasing category of breast density, and that sensitivity was somewhat lower following single reading than double reading at all levels of mammographic density. For single reading, sensitivity was 71.3% at the lowest density category (<25% fibroglandular tissue) and 40% at the highest density category (>75% fibroglandular tissue), compared to 76.9% in the lowest density category and 44% in the highest density category for double reading. A similar pattern was observed for false-positive recalls and interval cancers.[29] Whilst one might assume that double reading is of more benefit in females with dense breasts, due to the increased difficulty of the mammography reading task, there is little evidence of such a systematic pattern.

A second factor that might influence recall and cancer detection in breast cancer screening programmes is reader pairings. Brennan and colleagues examined sensitivity, specificity, and correct interpretation of mammograms between exhaustive theoretical-pairings of readers using test sets.[38] In this study, 12 board certified radiologists interpreted three sets of 60 images with known outcomes (40 normal/benign cases, and 20 cancers). Performance was calculated for individual readers and every possible pair of readers, with single reader performance being compared against the average performance of pairs of readers. There was some variation in performance between the best, worst, and random pairs of readers. For both the best- and the random pairs, sensitivity, specificity, and the proportion of correct interpretations were significantly better that the cohort's average single reader performance (and somewhat better for the best pairs compared to random pairs). In contrast, there was no significant difference in sensitivity between the worst pairs and the average single reader performance. This suggests that double reading might not always be better than single reading, and that it might be beneficial to randomize reader pairs. Optimization of pairings could be feasible; unfortunately, no data were available on which to determine these "best" pairs in this study. Caution is warranted in extrapolating from experimental studies to clinical practice, as some prior research has suggested little-to-no correlation between performance on test sets and clinical outcomes such as cancer detection and cancers being missed.[39,40] Some breast screening centres already strategically choose pairings, or more junior and senior staff, and of high and low recall readers. Telemedicine provides increasing opportunities for this.

A third factor is blinding of second reader to the decisions of the first reader. Klompenhouwer et al compared outcomes in a prospective series of 87,487 mammograms from a biennial breast cancer-screening programme in the Netherlands.[41,42] In this study blinded and non-blinded double reading strategies were alternated on a monthly basis, with females recalled if either reader recommended it. While no differences were observed between the strategies in terms of cancer detection rates or the positive-predictive value of recall, blinded reading led to an increased programme sensitivity (83.1% *vs* 75.5%) and a lower interval cancer rate (1.5% *vs* 3.1%) compared to non-blinded reading. However, this came at a cost of more females being recalled (3.3% *vs* 2.9%), a higher false-positive rate (25.8% *vs* 22.1%), more biopsies being carried out (17.4 *vs* 14.3 per 1000 screens), and a higher benign biopsy rate (10.1 *vs* 7.7 per 1000 screens). This finding is specific to systems where recall is if either reader recommends. Blinding leads to more independent decision-making, so more disagreements between readers. In this

system all disagreements result in recall, so blinding increases recall. In a system where there is arbitration of discordant results blinding may reduce recall, as it prevents an unblinded reader copying a decision to recall, which would effectively bypass the arbitration process.

In the 1960s, Smith introduced the term "alliterative errors" to describe the influence that readers can have on each other's decision making.[43] He posited that if a first reader misses an anomaly or places undue significance to a finding this might increase the likelihood of a subsequent reader drawing the same conclusion. There is a wealth of data suggesting that social influences play an important role in decision making, and that our beliefs and behaviours can be altered when we are faced with conflicting information from others (conformity).[44] However, there are few studies of conformity in relation to medical practice. Kaba and colleagues conducted a series of experimental studies that used patient simulation models. They found that medical and nursing students reported diagnoses that corresponded to incorrect vital signs given by confederates rather than the correct vital signs,[45,46] and that medical students were more likely to insert needles into an incorrect location on the knee when the skin had pre-existing insertion holes compared to unmarked skin.[47] Schöbel and colleagues conducted a hypothetical medical decision-making study in which psychology students were asked to diagnose which of two diseases was present based on symptoms. They found a lower proportion of the participants decisions reflected their own beliefs when a prior diagnosis was given by a (hypothetical) medical director (who had a higher hierarchical rank) compared to when a prior diagnosis was given by a (hypothetical) physician assistant (who was of equal hierarchical rank). Direct data on the effect of social influences on performance in breast cancer screening studies are sparse. In the study by Klompenhouwer et al, discrepancies between readers in the interpretation of mammograms were more common during blinded reading than non-blinded reading (57% *vs* 30%).[41] In our unpublished analysis of the CO-OPS trial, we have found similar patterns suggesting when unblinded the second reader may copy the first. Overall, there is some evidence that the full benefits of double reading are only realized with blinding.

FUTURE RESEARCH OPPORTUNITIES

One of the most anticipated evolutions of breast cancer screening is the potential for AI to replace the reader in examining the mammograms. There are many potential roles for AI, but we will only examine the potential for AI to replace the second reader here. The impact that this will have on overall accuracy is not straightforward, because of the complex interplay of factors we have already described in determining the accuracy of double reading. Measurement of the accuracy of the AI system itself is relatively straightforward using test sets with known cancer status. The impact of implementation of AI as second reader will depend on several factors. Firstly, blinding the reader to the decision of the AI system. If unblinded the reader may place too heavy reliance on the AI system, and align their decisions with those of AI. This removes some of the beneficial effects of double reading, so blinding the readers initial decision should be implemented. Secondly, if there is arbitration of discordant decisions between the first reader and AI. The accuracy of arbitration could be reduced if the arbitrating reader(s) have either too much or too little confidence in the accuracy of AI, in a similar manner to the alliterative errors and conformity described for double reading. The impact of tests is dependent on how clinicians interpret their results, and how they influence downstream decision-making.[48] Thirdly, if there is recall if either reader suggests, then there will be a requirement for very high specificity of the AI system to prevent the recall rate becoming unmanageably high. Fourthly, the variability between readers. In systems where very high recall readers have been systematically paired with low recall readers to control overall recall rates, replacing the second reader with AI will remove that safety net, which may result in increased recall rates. In fact, the introduction of an AI system with high test–retest reliability may in fact increase variability in the service given to females attending screening. To assess the impact of AI as a second reader requires a test–treat trial in which females are randomized to receive either the current double reading pathway, or the proposed double reading with AI pathway. This will allow measurement of reader behaviour reading mammograms, arbitrating discordant cases, and at the assessment clinic in the presence of AI in the pathway, and the impact on overall accuracy.

A second theme in breast screening research is a move towards three-dimensional imaging at the initial screen, either digital tomosynthesis or an abbreviated or fast version of MRI. These three-dimensional images take longer to examine, and so the cost of double reading is increased. There is some evidence that double reading is still beneficial when using digital tomosynthesis.[49] The lessons learned about the benefits of blinding, and the potential importance of reader pairing should be generalizable to three-dimensional imaging. Similarly, the finding that the extra cancers detected by the second reader are more likely to be smaller and of better prognosis should also apply overall in three-dimensional imaging, although more research on this is required.

One of the most interesting and promising avenues for future research lies in reconsidering the fundamentals of breast cancer screening. How often should females be screened, at what ages, with what recall threshold? Some of the most interesting questions consider trade-offs in allocating the funding for breast screening. What system overall maximises benefit and minimizes harm? This paper has considered changes to a single screening round. Double reading provides a change in accuracy and potentially recall threshold, with an increase in initial costs due to extra reader time. Similarly, tomosynthesis, and AI may affect accuracy and costs at each screening round. Increased investment in each screening round can increase cancer detection particularly of early stage disease and DCIS through increasing number of readers or technological advances. How might investment in this strategy compare to reducing the time between screens (round length), or extending the ages of eligibility? Decreasing the screening round length provides enhanced opportunities to detect faster growing cancers, which may otherwise be missed due to length time bias. Mammography is an ineffective test in younger females, and mammography in older females is more likely to be associated with overdiagnosis. There is a complex

interplay between all of these screening variables and the benefits and harms of breast screening. Future research should aim to make broad comparisons between strategies for implementing breast screening. Further, research which considers only a single screening round, such as research into double reading should maintain an emphasis on describing the characteristics of cancers detected and interval cancers, to allow careful extrapolation to the potential benefits and harms of breast cancer screening.

CONCLUSION

Double reading in breast cancer screening has the potential to reduce false-positive recall rates and increase cancer detection rates. However, this is dependent on both readers independently examining the females's mammograms for signs of cancer without knowledge of each others decisions, and an effective system of arbitration of discordant results. Replacing the second reader with AI will impact on the process of arbitration, and close examination of these effects will be critical to understanding the overall impact of AI in this role on women's outcomes from breast cancer screening.

ACKNOWLEDGMENT

Dr Taylor-Phillips is supported by an NIHR Career Development Fellowship (CDF-2016-09-018). The opinions are those of the authors and not the NIHR, the NHS or the Department of Health and Social Care.

REFERENCES

1. Fitzmaurice C, Allen C, Barber RM, Barregard L, Bhutta ZA, Brenner H, et al. Global, regional, and National cancer incidence, mortality, years of life lost, years lived with disability, and Disability-Adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. *JAMA Oncol* 2017; **3**: 524–48. doi: https://doi.org/10.1001/jamaoncol.2016.5688
2. Nelson HD, Cantor A, Humphrey L, et al. Preventive services Task force evidence syntheses, formerly systematic evidence reviews, screening for breast cancer: a systematic review to update the 2009 U. *S. Preventive Services Task Force Recommendation. Rockville (MD), Agency for Healthcare Research and Quality* 2016;.
3. Myers ER, Moorman P, Gierisch JM, et al. Benefits and harms of breast cancer screening: a systematic review. *Jama* 2015; **314**: 1615–34.
4. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M, et al. The benefits and harms of breast cancer screening: an independent review. *Br J Cancer* 2013; **108**: 2205–40. doi: https://doi.org/10.1038/bjc.2013.177
5. Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015; **175**: 1828–37. doi: https://doi.org/10.1001/jamainternmed.2015.5231
6. Wilson R. *Liston J: Quality Assurance Guidelines for Breast Cancer Screening Radiology NHS Breast Screening Programme Publication Number 59*. England: Sheffield; 2011.
7. Perry N, Broeders M, de Wolf C.et al. *European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition*. Office for Official Publications of the European Union: Luxembourg, European Commission; 2013.
8. Taylor P, Potts HWW. Computer AIDS and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. *Eur J Cancer* 2008; **44**: 798–807. doi: https://doi.org/10.1016/j.ejca.2008.02.016
9. Harvey SC, Geller B, Oppenheimer RG, Pinet M, Riddell L, Garra B, et al. Increase in cancer detection and recall rates with independent double interpretation of screening mammography. *American Journal of Roentgenology* 2003; **180**: 1461–7. doi: https://doi.org/10.2214/ajr.180.5.1801461
10. Ciatto S, Ambrogetti D, Bonardi R, Catarzi S, Risso G, Rosselli Del Turco M, et al. Second reading of screening mammograms increases cancer detection and recall rates. results in the Florence screening programme. *J Med Screen* 2005; **12**: 103–6. doi: https://doi.org/10.1258/0969141053908285
11. Deans HE, Everington D, Cordiner C, Kirkpatrick AE, Lindsay E, et al. Scottish experience of double reading in the National breast screening programme. *The Breast* 1998; **7**: 75–9. doi: https://doi.org/10.1016/S0960-9776(98)90060-1
12. Anttinen I, Pamilo M, Soiva M, Roiha M. Double reading of mammography screening films--one radiologist or two? *Clin Radiol* 1993; **48**: 414–21. doi: https://doi.org/10.1016/S0009-9260(05)81111-0
13. Duijm LEM, Groenewoud JH, Hendriks JHCL, de Koning HJ. Independent double reading of screening mammograms in the Netherlands: effect of arbitration following reader disagreements. *Radiology* 2004; **231**: 564–70. doi: https://doi.org/10.1148/radiol.2312030665
14. Brewer NT, Salz T, Lillie SE. Systematic review: the long-term effects of false-positive mammograms. *Ann Intern Med* 2007; **146**: 502–10. doi: https://doi.org/10.7326/0003-4819-146-7-200704030-00006
15. Anderson ED, Muir BB, Walsh JS, Kirkpatrick AE. The efficacy of double reading mammograms in breast screening. *Clin Radiol* 1994; **49**: 248–51. doi: https://doi.org/10.1016/S0009-9260(05)81850-1
16. 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; **5**: 195–201. doi: https://doi.org/10.1136/jms.5.4.195
17. Leivo T, Salminen T, Sintonen H, Tuominen R, Auerma K, Partanen K, et al. Incremental cost-effectiveness of double-reading mammograms. *Breast Cancer Res Treat* 1999; **54**: 261–7. doi: https://doi.org/10.1023/A:1006136107092
18. Thurfjell E. Mammography screening methods and diagnostic results. *Acta Radiol Suppl* 1995; **395**: 1–22.
19. Brown J, Bryan S, Warren R. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. *BMJ* 1996; **312**: 809–12. doi: https://doi.org/10.1136/bmj.312.7034.809
20. Séradour B, Wait S, Jacquemier J, Dubuc M, Piana L. Modalities of reading of detection mammographies of the programme in the Bouches-du-Rhône. results and costs 1990-1995. *J Radiol* 1997; **78**: 49–54.
21. Screening and Immunisations Team,. *Centre HaSCI: Breast Screening Programme,*

England—2012–13. Leeds, UK: The Health and Social Care Information Centre; 2014.
22. U.S. food and drug administration: mammography quality Standards act and program. 2017;.
23. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med Overseas Ed* 2005; **353**: 1773–83. doi: https://doi.org/10.1056/NEJMoa052911
24. Skaane P, Hofvind S, Skjennald A. Randomized trial of screen-film versus full-field digital mammography with soft-copy reading in population-based screening program: follow-up and final results of Oslo II study. *Radiology* 2007; **244**: 708–17. doi: https://doi.org/10.1148/radiol.2443061478
25. Pisano ED, Hendrick RE, Yaffe MJ, Baum JK, Acharyya S, Cormack JB, et al. Diagnostic accuracy of digital versus film mammography: exploratory analysis of selected population subgroups in DMIST. *Radiology* 2008; **246**: 376–83. doi: https://doi.org/10.1148/radiol.2461070200
26. Henderson LM, Miglioretti DL, Kerlikowske K, Wernli KJ, Sprague BL, Lehman CD, et al. Breast cancer characteristics associated with digital versus Film-Screen mammography for screen-detected and interval cancers. *American Journal of Roentgenology* 2015; **205**: 676–84. doi: https://doi.org/10.2214/AJR.14.13904
27. Taylor-Phillips S, Jenkinson D, Stinton C, Wallis MG, Dunn J, Clarke A, et al. Double reading in breast cancer screening: cohort evaluation in the CO-OPS trial. *Radiology* 2018; **287**: 749–57. doi: https://doi.org/10.1148/radiol.2018171010
28. 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; **11**: e0159806. doi: https://doi.org/10.1371/journal.pone.0159806
29. Euler-Chelpin Mvon, Lillholm M, Napolitano G, Vejborg I, Nielsen M, Lynge E, et al. Screening mammography: benefit of double reading by breast density. *Breast Cancer Res Treat* 2018; **171**: 767–76. doi: https://doi.org/10.1007/s10549-018-4864-1
30. Coolen AMP, Voogd AC, Strobbe LJ, Louwman MWJ, Tjan-Heijnen VCG, Duijm LEM, et al. Impact of the second reader on screening outcome at blinded double reading of digital screening mammograms. *Br J Cancer* 2018; **119**: 503–7. doi: https://doi.org/10.1038/s41416-018-0195-6
31. Weigel S, Heindel W, Heidrich J, Hense H-W, Heidinger O. Digital mammography screening: sensitivity of the programme dependent on breast density. *Eur Radiol* 2017; **27**: 2744–51. doi: https://doi.org/10.1007/s00330-016-4636-4
32. Jørgensen KJ, Gøtzsche PC, Kalager M, Zahl P-H. Breast cancer screening in Denmark: a cohort study of tumor size and overdiagnosis. *Ann Intern Med* 2017; **166**: 313–23. doi: https://doi.org/10.7326/M16-0270
33. Welch HG, Prorok PC, O'Malley AJ, et al. Breast-Cancer tumor size, overdiagnosis. *and Mammography Screening Effectiveness* 2016; **375**: 1438–47.
34. Yen M-F, Tabár L, Vitak B, Smith RA, Chen H-H, Duffy SW, et al. Quantifying the potential problem of overdiagnosis of ductal carcinoma in situ in breast cancer screening. *Eur J Cancer* 2003; **39**: 1746–54. doi: https://doi.org/10.1016/S0959-8049(03)00260-0
35. Holland K, van Gils CH, Mann RM, Karssemeijer N. Quantification of masking risk in screening mammography with volumetric breast density maps. *Breast Cancer Res Treat* 2017; **162**: 541–8. doi: https://doi.org/10.1007/s10549-017-4137-4
36. Wanders JOP, Holland K, Veldhuis WB, Mann RM, Pijnappel RM, Peeters PHM, et al. Volumetric breast density affects performance of digital screening mammography. *Breast Cancer Res Treat* 2017; **162**: 95–103. doi: https://doi.org/10.1007/s10549-016-4090-7
37. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, et al. Mammographic density and the risk and detection of breast cancer. *N Engl J Med* 2007; **356**: 227–36. doi: https://doi.org/10.1056/NEJMoa062790
38. Brennan PC, Ganesan A, Eckstein MP, Ekpo EU, Tapia K, Mello-Thoms C, et al. Benefits of independent double reading in digital mammography: a theoretical evaluation of all possible pairing methodologies. *Acad Radiol* 2019; **26**: 717–23. doi: https://doi.org/10.1016/j.acra.2018.06.017
39. Rutter CM, Taplin S. Assessing mammographers' accuracy. A comparison of clinical and test performance. *J Clin Epidemiol* 2000; **53**: 443–50. doi: https://doi.org/10.1016/s0895-4356(99)00218-8
40. Gur D, Bandos AI, Cohen CS, et al. The "Laboratory" Effect: Comparing Radiologists' Performance and Variability during Prospective. *Clinical and Laboratory Mammography Interpretations* 2008; **249**: 47–53.
41. Klompenhouwer EG, Voogd AC, den Heeten GJ, Strobbe LJA, de Haan AFJ, Wauters CA, et al. Blinded double reading yields a higher programme sensitivity than non-blinded double reading at digital screening mammography: a prospected population based study in the South of the Netherlands. *Eur J Cancer* 2015; **51**: 391–9. doi: https://doi.org/10.1016/j.ejca.2014.12.008
42. Weber RJP, Klompenhouwer EG, Voogd AC, Strobbe LJA, Broeders MJM, Duijm LEM, et al. Comparison of the diagnostic workup of women referred at non-blinded or blinded double reading in a population-based screening mammography programme in the South of the Netherlands. *Br J Cancer* 2015; **113**: 1094–8. doi: https://doi.org/10.1038/bjc.2015.295
43. Smith MJ. *Error and variation in diagnostic radiology*: CC Thomas; 1967.
44. Cialdini RB, Goldstein NJ. Social influence: compliance and conformity. *Annu Rev Psychol* 2004; **55**: 591–621. doi: https://doi.org/10.1146/annurev.psych.55.090902.142015
45. Kaba A, Beran TN. Impact of peer pressure on accuracy of reporting vital signs: an interprofessional comparison between nursing and medical students. *J Interprof Care* 2016; **30**: 116–22. doi: https://doi.org/10.3109/13561820.2015.1075967
46. Kaba A, Beran TN, White D. Accuracy of interpreting vital signs in simulation: an empirical study of conformity between medical and nursing students. *Journal of Interprofessional Education & Practice* 2016; **3**: 9–18. doi: https://doi.org/10.1016/j.xjep.2016.03.002
47. Beran TN, McLaughlin K, Al Ansari A, Kassam A, et al. Conformity of behaviors among medical students: impact on performance of knee Arthrocentesis in simulation. *Adv in Health Sci Educ* 2013; **18**: 589–96. doi: https://doi.org/10.1007/s10459-012-9397-5
48. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PMM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ* 2012; **344**(feb21 1): e686 doi: https://doi.org/10.1136/bmj.e686
49. Tagliafico AS, Calabrese M, Bignotti B, Signori A, Fisci E, Rossi F, et al. Accuracy and reading time for six strategies using digital breast tomosynthesis in women with mammographically negative dense breasts. *Eur Radiol* 2017; **27**: 5179–84. doi: https://doi.org/10.1007/s00330-017-4918-5