Review of the evidence on the use of arbitration or consensus within breast screening: A systematic scoping review

Hackney, L, Szczepura, A, Moody, L & Whiteman, B

Original citation & hyperlink:
Hackney, L, Szczepura, A, Moody, L & Whiteman, B 2017, 'Review of the evidence on the use of arbitration or consensus within breast screening: A systematic scoping review' *Radiography*, vol 23, no. 2, pp. 171-176
https://dx.doi.org/10.1016/j.radi.2017.01.002

DOI 10.1016/j.radi.2017.01.002
ISSN 1078-8174
ESSN 1532-2831

Publisher: Elsevier

Review of the Evidence on the Use of Arbitration or Consensus within Breast Screening; A Systematic Scoping Review.

Lisa Hackney, Professor Ala Szczepura, Louise Moody, Becky Whiteman

Abstract

Objectives: A systematic scoping review was undertaken to establish the evidence base on arbitration and consensus in mammography reporting.

Database searches were supplemented with hand searching of peer –reviewed journals, citation tracking, key author searching, grey literature and personal contact with experts. A 3-stage process was utilised to screen a large volume of literature (601) against the inclusion and exclusion criteria. 26 papers were retained.

Key findings: A lack of guidance and underpinning evidence to inform how best to use arbitration or consensus to resolve discordant reads. In particular, a lack of prospective studies to determine effectiveness in real-life clinical settings.

Conclusion: The insufficiency of follow-up or reporting of true interval cancers compromised the ability to conclude the effectiveness of the processes.

Introduction

An estimated 1.6 million women were diagnosed with breast cancer worldwide in 2012, representing the most common cancer in developed and developing countries.[2] Cancer Registration statistics (2013)[3] confirm that 43.5% of UK female breast cancer cases are

diagnosed in the 50-59-age range and 34.3% in the 60-69-age range, with a 6% increase in incidence rates in UK females between 2002-2004 and 2011-2013. The combination of breast cancer prevalence and demographic trends contributed to the founding of the UK National Health Service (NHS) Breast Screening Programme (NHSBSP) in 1988 to facilitate early detection and reduce mortality rates. Although the incidence of breast cancer has continued to rise in the UK over the last decade the mortality rates have fallen.[3]

In order to increase cancer detection rates different reporting strategies are utilised in various regions of the world. In the United States, single radiologist reporting or single radiologist reporting with Computer Aided Detection (CAD) are commonly employed.[4] Double reporting by Radiologists specialised in breast screening is the European standard.[5] Unique to the UK is double reporting undertaken by trained mammographer's (Allied Health Professionals). This was validated in 2012 following an extensive NHSBSP research project (Non-Discordant Radiographer Only Reporting - NDROR).[6] The principal complexity for reporters is balancing the trade-off relationship of attaining a high sensitivity whilst minimising false positives[7], which impact adversely on patient wellbeing[8] and represent cost implications in time and resources.

Double reporting inherently results in discordant cases, which require resolution. The most common decision methods utilised are arbitration by a third independent reader or some form of consensus review. For the purpose of this review arbitration and consensus definitions are those detailed in Table 1. Complex pathways also exist where both consensus and arbitration are undertaken in the decision-making process.

Table 1 Definitions Used for Arbitration and Consensus

Process	**Definition**
Arbitration	solitary 3rd reader who made the final decision
Consensus	Group decision making process. Group members discuss and agree to support a decision even if not the "preferent" of each individual

Until recently, NHSBSP guidance stipulated that the independent third reader or lead of the consensus review must be a medical practitioner. Concerns about the future availability of specialist radiologists have been highlighted in a recent Royal College of Radiologists publication.[9] This predicts the retirement of 21% of breast radiologists in the next five years, together with a potential 2.2 million increase in women eligible for screening if the current age extension programme is implemented (based on current population figures). The NHSBSP arbitration guidance[1] was necessary as it was recognised that, to maintain the current quality standards and avoid delays in patient management, the extension of arbitration duties to non-medics had to be considered.

Whilst there was national momentum for delegation of arbitration to radiographers, there seemed to be little consolidated evidence available on the effectiveness of arbitration versus consensus and whether one strategy produces improved performance in a breast-screening unit. No systematic reviews in this area had been undertaken.

Review Aims

The primary aim was to establish what evidence there is to support different models of arbitration or consensus review in breast screening and evaluate the evidence to support the effectiveness of the different models. Specifically, effectiveness was defined in terms of recall rates, cancer detection rate, Positive Predictive Value (PPV) and programme sensitivity/specificity. The review did not aim to address cost-effectiveness.

Method

Literature searches of PubMed, Medline, CINAHL, EMBASE, Scopus, Web of Science and the Cochrane Library were supplemented by a broad Google scholar web search. Hand searching of key peer-reviewed breast and radiology journals, a manual search of reference lists and key author searching was undertaken. Grey literature was sourced by hand searching of conference proceedings and doctoral theses. Personal contact with experts internationally was also undertaken in locating relevant literature.

Table 2 lists the search terms and variations used in the database searches. Concepts of interest[10,11] were cross-referenced by searching Cochrane reviews for validation.

Table 2 Search Terms and Variations Used

Exploded terms	Alternative keywords
Breast neoplasm	breast adj3 (neoplasm* OR carcinoma* OR tumour* OR tumor* OR cancer*.
Mass screening	breast adj3 (scan* OR screen* OR radiograph* OR imaging OR visualise OR visualize OR exam* OR test* OR mammogra* OR routine* OR check* OR diagnos* OR detect*)
Mammography	mammogra* adj3 (scan* OR screen* OR visualise OR visualize OR exam* OR test* OR breast*)
Early detection of cancer	
National Health Service Breast Screening Program	OR "NHSBSP" or "UK breast screen* program*" "NHS breast screen* program*"
Negotiating	arbitration* OR discordan* OR discrepan* OR disparity* OR negotiat* OR disagree* OR conflict* OR differen* OR inconsisten* AND variation* OR consensus* OR uncertain*
Decision making	"decision mak* OR shared decision making" OR "medical decision making" OR "choice behaviour" OR "problem solving" OR "clinical decision analysis" OR "critical think*" OR "decision aids" OR "Task performance and analysis"
Interpersonal communication	

Inclusion/exclusion criteria

Inclusion and exclusion criteria detailed in Table 3 related to the intervention and population characteristics but there was no limitation on study design.

Table 3 Inclusion and exclusion criteria

Inclusion criteria
1. Provides an English abstract or summary (to assess content) or the title explicitly demonstrates relevance 2. Specifically mentions breast reporting arbitration, 3rd reader or consensus processes OR 3. Discusses reporting strategies – i.e. single reading, double reading, blinded or non-blinded reading. OR 4. Reports strategies for management of discrepant cases – i.e. higher reader recall, arbitrate all recalls, arbitrate discordant cases only. OR 5. Reports the grade of personnel undertaking the arbitration/consensus/3rd read task i.e. radiologist, radiographer, clinician, surgeon OR 6. Specifically, in relation to arbitration, 3rd reader or consensus mentions any attributes required by the personnel undertaking the task. In particular: *Volumes of films read per annum,* *Number of years' experience of the reporter,* *Attendance at MDT's,* *Decision making skills,* *Audit and reflective practice*

Exclusion criteria
1) Non English-language paper 2) Arbitration, consensus or 3rd reader 'mentioned in passing' but not a significant focus of the article.

Studies published from 1st January 2008 were considered for inclusion in this review, as it would give a 2-year lead in period from when relevant NHSBSP guidance was last revised (2010/2011). Initial searches retrieved small numbers of articles. Therefore, for subsequent searches either the start year was extended to 2005, or no date restriction was applied to ascertain if a seminal piece of work was produced earlier.

Two reviewers independently undertook a three-stage process for filtering the literature retrieved.[12,13] Reviewer one was a Masters in Research student and consultant radiographer, reviewer two is a Professor of Health Technology Assessment. First stage selection was based on an analysis of the titles and/or abstracts or summaries. In the second screening stage, abstracts were screened for all retained literature, against the agreed inclusion and exclusion criteria. Any disagreement was resolved after retrieval and review of the full text (five articles identified and arbitrated).

In stage three, the full text of all potentially eligible peer-reviewed papers /grey literature items were examined. A third reviewer (clinical research fellow) resolved any disagreements over the eligibility of a particular study (no articles identified). Articles that met the inclusion criteria were documented in a customised data extraction form (S1). Data extracted included:

- Article descriptors: author; year of publication; country where study performed;
- Study context (screening versus diagnostic);
- Sample size;
- Data analysis/metrics;
- Reporting strategy (double reading; blinded or non-blinded reading);
- Use of a test set versus prospective series of patient selection;
- Strategy utilised for discordant results;
- Readers (professions, number acting as arbitrator, years of experience, and specific training in mammogram reading);
- Strengths and weaknesses of the study (to include selection/measurement bias).

The data extraction form enabled raw data from multiple disparate studies to be amalgamated and compared, aiding in pattern recognition and providing a '*rapid and succinct summary of the literature for review*'.[10]

Quality assessment for methodological rigour was undertaken using criteria derived from the standardised Critical Appraisal Skills Programme (CASP)[14] questions where appropriate. Quality appraisal was undertaken independently by two reviewers, and in cases of disagreement, a third reviewer was consulted with the aim of reaching consensus through discussion. No weighting or ranking of the papers finally included was undertaken. The findings were summarised in a thematic narrative synthesis.

Results

The PRISMA flow chart in (Fig 1) details the review process. Details of the included studies, together with extracted data and quality assessment are summarised in S1.

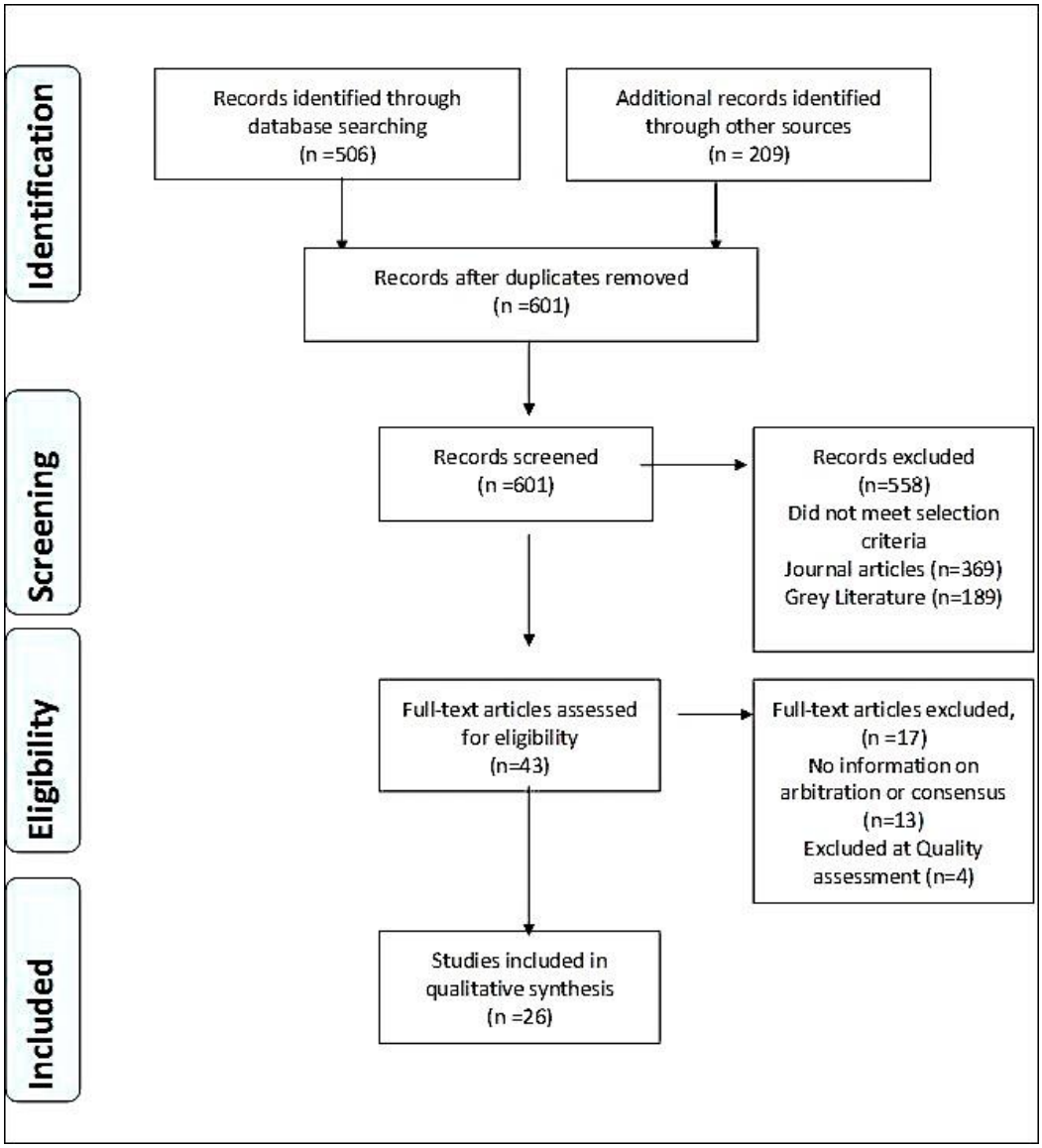

Figure 1. Flow Diagram of included articles.

The retained twenty-six studies consisted of a mixture of designs, but all were quantitative in nature. There were eight retrospective studies and twelve prospective studies with one[15] a mixed design of retrospective and prospective cases. The remaining study characteristics comprised of two audits[16,17], two systematic reviews[18,19] and one observational epidemiological study.[20] There were, only five prospective studies[21-25] predominantly looking at the effect of arbitration or consensus, the remainder focused on the transition

from screen film mammography to digital mammography[26], comparison of current reading protocols to CAD assisted reading[27], impact on the number of readers[28,7] and comparison of conventional Full Field Digital Mammography with tomosynthesis.[29-31] Two systematic reviews[18,19]were incorporated as arbitration or consensus was integrated within the reporting process although their primary remit was comparison of reading strategies i.e. double reading with single reading, and single reading with, and without CAD.

Publication dates ranged from 1998 to 2016 and were predominantly from the UK (n=11) with the last publication being a 2014 audit; prior UK studies relate to 2009 or earlier. The remaining publications were from the Netherlands (n=4), Norway (n=4) and Italy (n=2) with one publication from each of Australia, Finland, Sweden, Spain and Germany. It is notable that a number of studies have been undertaken prior to the start of the UK digital transition in 2006.

There was variability in both the experience and cohort of professionals' undertaking the reporting process; radiologists, radiographers, research fellows, senior radiology trainee, general radiologists and residents (equivalent to a UK House Officer). Internationally, specialist and general radiologists are representative of the workforce reporting screening mammography, which is disparate to current UK practice. Experience of the reporters ranged from 6 months to more than 30 years.

All studies were within population-based national or regional screening programmes, with sample sizes ranging from 182 test set cases to a retrospective review of 1,033,870 prevalent and incident screens. Study duration varied greatly dependent on study design, ranging from a 4 month prospective study[22] to a 9-year retrospective study.[32]

Double Reporting - Blinded and Non-blinded

From the data available, the percentage of cases in which double reading produced discordant results varied greatly, ranging from 0.5% [25] to 57.2 % [33] of cases. Klompenhouwer et al.[33] demonstrated that there was a significant difference in the number of discordant reads dependent on whether the double reading was performed blinded (57.2%) vs. non-blinded (29.1%).

Arbitration Studies

The final decision of the arbitrator resulted in a wide variation (27%[17,22] - 50%[34]) of cases that were subsequently recalled to assessment. Ciatto et al.[23] investigated the effectiveness of arbitration, but this was on non-consecutive cases limited to periods when a radiologist was available to undertake the third read. Follow-up data for 58% of the cases in which arbitration concluded a negative outcome were not available. Therefore, the effect of the arbitration process and subsequently cancer detection rates could only be estimated.

Overall, studies reported that compared to highest reader recall (non-arbitration), arbitration resulted in significant reductions in recall rates, with relative decreases in the range of 17.8%[35] to 40.9%[22]. However, the results of Caumo et al.[22] must be interpreted with caution as this study was conducted over a short (4-month) period, with a single experienced (>30yrs) individual arbiter. All cases were recalled to assessment irrespective of the arbitrator's decision, and therefore there was no direct impact on clinical care rendering the process futile. Variability in reducing recalls is also confirmed by Liston and Dall[16] reporting findings from a seven-year audit. With such variation in recall rates the PPV of assessment cases following arbitration is also unpredictable with low PPV's of 8.3%[17] to

31.2 %[33] reported.

There is disparity between the studies regarding the effect of arbitration on cancer detection rates. Klompenhouwer et al.[35] declared an overall decrease, albeit it (0.1-0.2%) not statistically significant, whilst the systematic review by Taylor and Potts[18] stated increased cancer detection rates. Dinnes et al.[19] systematic review affirmed there was '*insufficient evidence to detect any pattern in cancer detection based on recall policy*'.

Consensus Studies

Five papers [15,24,26,27,30] mentioned consensus as the method of resolving discordant cases but only two of the studies[15,24] were specifically looking at the effectiveness of the process. The three remaining studies were evaluating CAD and tomosynthesis. Therefore, limited data was available on recall rates to assessment following consensus with a range 31.1%[26] to 65.6%[15] reported. The high number of cases returned to routine recall in the Norwegian study[26] relates to the cumulative scoring system utilised where a score of 2 (defined as probably benign) or greater is referred for consensus discussion.

There was a supposition from some of the literature that fewer cancers will be missed by panel consensus compared to single reader arbitration. However, no evidence was found to support this. UK studies[15/17/24]have elucidated that centres may favour group consensus as it reflects the change in professional skill mix within the UK breast reporting system, offers an opportunity for educational learning from cases, or the perception that groups will miss fewer cancers. No evidence was found to support this.

As with all group meetings, the dynamics within the consensus team can be a significant

factor affecting the final decision. Hukkinen et al.[28] although describing independent reading and conference consensus (the majority considered decisive) stated that they avoided readers discussing discordant cases to prevent the situation of one reader being overruled by another. Kerr and Tindale[36] and Bankier et al.[37] describe the complexities that exist within consensus discussions where one reader is the dominant and opinions are not equally weighted. The performance-reducing effects of '*group think*'[37] are also an important consideration in consensus where it is evidenced that individuals may change their judgment to what they '*believe others want to hear*'.[37]

Hukkinen et al.[28] calculated consensus by averaging sensitivities and this achieved maximum results when combining the readings of the four best performers. This is similar in principle to the unique Collective Intelligence (CI) study[7] which provided an interesting perspective as it removed the hierarchy and difficulties associated with group decision-making. A majority, quorum and weighted quorum rule was tested against an individual radiologists performance. In accordance with Hukkinen et al.[28] as group size increased all three CI rules achieved increases in true positives and decreases in false positives. Larger groups were declared to make more accurate decisions (concurrent improvements in true and false positives), but relatively small group sizes demonstrated improvements in achieving more true positives, fewer false positives and therefore greater overall accuracy. However, this was a test set scenario with no influence on real-life cases. As this model requires multiple reads to evaluate a mammogram, this may be problematic if units are struggling to achieve screen to results within a two-week period, as required in the UK.

A significant message from Jenkins et al.[17] and Hofvind et al.[32] was that the interval cancer rate was substantially higher in cases that had undergone arbitration or consensus relative

to the rate among concordant negative screenings. Jenkins et al.[17] report that 19.4% of interval cancers categorised as uncertain and suspicious were not initially called by any reader compared to 36.1% that had been recalled by at least one film reader ($p < 0.001$). This raises the question of whether arbitration or consensus could be refined to aid earlier detection in such cases.

Mixed Studies/Reviews

Within a number of studies[18-20,25,,32,38,39] it is not possible to differentiate the effect of arbitration versus consensus as the processes are either integrated in the discussion, or both are undertaken within the decision making strategy i.e. mutual consensus between the two readers with persistent discordant case being reviewed by an arbitration panel. The Duijm et al.[25] study reports that this strategy resulted in 45% of cases being resolved by mutual discussion and 55% still requiring arbitration by a panel. The panel recalled if at least one arbitration member considered it necessary, which may have resulted in higher recall rates comparative to a majority decision and the subsequent effect on PPV remains unknown.

Groenewoud et al.,[38] although a paper primarily concerned with cost effectiveness of different reporting strategies, stated that referral rates were highest with decision-making by consensus (73.8%) compared to arbitration (52.7%). However, this was an experimental study with test cases and therefore again not reflective of clinical practice. Conversely, Blanks et al.[20] studied cancer detection rates for a variety of reading strategies and concluded that although consensus had a lower recall rate, the Standardised Detection Ratio (SDR) was higher for double reading with arbitration compared to double reading and

consensus for both prevalent and incident screens. Also, for incident screens the SDR for small (<15mm) invasive cancers was also higher (Double consensus =1.00 vs. Double arbitration =1.18). It is noteworthy that this study is 18 years old and with improved technologies the SDR is now much higher, and there may be value in repeating this study to ascertain the impact on current practices.

A further variation in recall policy was discussed by Hofvind et al.[32] and Matcham et al.[15] who performed consensus on all recalls (concordant and discordant) resulting in 17.9% and 10.7% of the concordant readings to recall being over-ridden at consensus.

Follow-Up/False Negative Cases

Regardless of the strategy used, cancer cases were incorrectly dismissed to routine recall by both processes. Only twelve studies provided information regarding interval cancers. The length of follow-up was variable ranging from four months to seven years, and as a full screening interval (2 or 3years dependent upon country) was not complete prior to the reporting of some studies, the true effect of cases returned to routine screening is unknown. Shaw et al.[24] and Duijm et al.[25] report fairly low rates of cancer cases dismissed at consensus 1.1% and 3% respectively. More significantly, Jenkins et al.[17] showed 4.1% of false negative interval cancers were double reported as normal, which was significantly lower than cases where at least one reader had indicated recall (10.9%; $p< 0.005$).

Tumour/Mammographic Characteristics of Discordant Cases.

Three studies investigated the mammographic features of tumour's detected at discordant reading. Klompenhouwer et al.[33] described no difference in the proportion of DCIS, smaller tumours, lymph node involvement or advanced tumours between screen-detected cancers

and those missed at arbitration. Conversely Cornford et al.[34] indicate arbitration cancers more frequently presented as parenchymal distortions and were smaller in size ($p<0.045$), a finding also supported by Cawson et al.[39] Lobular cancers which are often mammographically difficult to detect were reported to be *more common in the arbitration group, albeit of borderline significance*.[34]

Emerging Technologies

Several studies included an assessment of CAD or tomosynthesis. Although an evaluation of these was beyond the scope of the review it was notable that both technologies impacted on the number of arbitration cases and subsequent recalls.

The Skaane et al.[31] study demonstrated that although 62% of radiologists referred fewer patients for arbitration with the use of FFDM and tomosynthesis the overall number of women recalled after arbitration was larger for this cohort (351 versus 265), which was also supported by Lang et al.[29] and Skaane et al.[30]

An important factor related to the use of new technologies is that they may improve the cancer detection rates and hence produce more recalls. Therefore, the role of arbitration and consensus will be paramount in reducing false positives, as resources within assessment clinics are already limited in some services. The CAD studies identified were primarily concerned with aiding detection of lesions rather than assisting the decision making process. The James and Cornford[40] study was unique in investigating the potential of CAD as an arbitrator, but this study as with others indicated that CAD produced too many false prompts. However, these studies were undertaken in 2009 or earlier, and CAD systems are evolving with the next generation of CADx a possibility for aiding diagnosis.

Conclusions and future work

This review has revealed a dearth of literature relating to either strategy. No research was identified comparing the accuracy of an independent 3rd reader (arbitrator) versus consensus (group/panel review) of discordant cases. There is a lack of guidance and underpinning evidence to inform how best to use arbitration or consensus, but no current system recalls all discordant cancer cases.

Definitions of consensus and arbitration are not clear-cut. The two terms are used interchangeably and often confusing with some studies reporting 'arbitration by an individual', others 'arbitration by a panel', and 'consensus based arbitration'. The lack of clear definitions makes it not only difficult to review the literature and synthesise the findings, but it also adds to confusion in a clinical setting when discussing processes with no clear delineations. Internationally there is disparity in the scoring systems used to grade the mammographic images and the guidelines for recall rates. Overall, screening outcome is influenced by many interrelated factors and the disparities in screening interval; classifications, reading strategies and reader performance make international comparisons problematic. Breast screening units have implemented variances in practice when deciding which cases are sent for review, strategies to resolve discordant cases and structure and scheduling of the processes.

Overall, either the short follow-up period, lack of complete data, absence of reporting of true interval cancers versus false negatives and the retrospective nature of many studies means there is insufficient evidence to assess the effectiveness of one strategy versus the other. Given the current workforce shortages in the UK, the use of a 3rd reader arbiter

versus a consensus meeting involving a group of individuals is an important consideration in terms of available skills as well as costs.

The primary aim of future research would be to establish current practice and to develop clear precise definitions and guidance on the processes. Further research would be required to:

1. Explore the clinical implications (time/resources/benefits) of a consensus panel reviewing all recall cases (concordant and discordant).
2. Explore the dynamics of the professionals that constitute consensus meetings and determine how the final decision is made.
3. Ascertain why some sites will be early adopters and some sites possibly non-adopters of the NHSBSP guidance[1] and determine the consequences of disparate practice not just for professionals, but service users? In particular, the impact on outcomes of performance measures (recall rates, PPV, screen to routine recall and screen to assessment).

References

1. NHS Breast Screening Programme. *Guidance on who can undertake arbitration.* London: Public Health England; 1st edition August 2016
2. GLOBOCAN 2012 v1.0. *Cancer incidence and mortality worldwide*: *IARC Cancer-Base No. 11 [Internet].* International Agency for Research on Cancer; 2013; 2012. Accessed May 2016, at, http://globocan.iarc.fr.
3. Office for National Statistics. *Breast cancer: incidence, mortality and survival;* 2013.
4. National Academies of Sciences, Engineering, and Medicine. Assessing and improving the interpretation of breast images: Workshop Summary. 2015 Washington, DC: The National Academies Press.
5. Perry N, Broeders M, de Wolf C, Tornberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition–summary document. *Ann Oncol* 2008; **19**(4): 614–22.

6. Bennett RL, Sellars SJ, Blanks RG, Moss SM. An observational study to evaluate the performance of units using two radiographers to read screening mammograms. *Clinical Radiology*. 2012;**67**:114-121
7. Wolf M, Krause J, Carney PA, Bogart A, Kurvers RHJM. Collective intelligence meets medical decision-making: The collective outperforms the best radiologist. *PLoS ONE* 2015;**10**.
8. Bond, M, Garside, R, Hyde, C. Improving screening recall services for women with false-positive mammograms: A comparison of qualitative evidence with UK guidelines. *BMJ Open* 2015; **5**: e005855
9. The Royal College of Radiologists, The breast imaging and diagnostic workforce in the United Kingdom. *Results of a survey of NHS Breast Screening Programme units and radiology departments.* London: The Royal College of Radiologists; 2016
10. Kable A, Pich J, Maslin-Prothero S. A structured approach to documenting a search strategy for publication: A 12 step guideline for authors'. *Nurse Education Today* 2012;32:878-886
11. Lloyd-Jones N, Masterson A. 'Writing skills and developing an argument' in Bailliere's study skills for nurses and midwives. Edinburgh: Bailliere Tindall; 2010.
12. Bettany-Saltikov J. Learning how to undertake a systematic review: part 1. *Nursing standard* 2010;**24**(50):47–55
13. Bettany-Saltikov J. Learning how to undertake a systematic review: part 2. *Nursing Standard* 2010;**24**(51):47–56
14. CASP (2014) *Critical appraisal skills programme* [online] Accessed October 2015, at, http://www.casp-uk.net/
15. Matcham NJ, Ridley NT, Taylor SJ, Cook JL, Scolding J. Breast screening: the use of consensus opinion for all recalls' *Breast* 2004;**13**(3):184-7
16. Liston JC, Dall BJG. Can the NHS Breast Screening Programme afford not to double read screening mammograms? *Clin.Radiol*. 2003;**58**:474-7.
17. Jenkins J, Murphy AE, Edmondson-Jones M, Sibbering DM, Turnbull AE. Film reading in the East Midlands Breast Screening Programme - Are we missing opportunities for earlier diagnosis? *Clin.Radiol*. 2014;**69**:385-90.

18. 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.
19. Dinnes J, Moss S, Melia J, Blanks R, Song F, Kleijnen J. Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: Findings of a systematic review. *Breast* 2001;**10**:455-63.
20. 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.
21. Mucci B, Athey G, Scarisbrick G. Double read screening mammograms: The use of a third reader to arbitrate on disagreements. *Breast* 1999;**8**:37-9.
22. Caumo F, Brunelli S, Tosi E, Teggi S, Bovo C, Bonavina G, et al. On the role of arbitration of discordant double readings of screening mammography: experience from two Italian programmes. *Radiol.Med*. 2011;**116**:84-91.

23. Ciatto S, Ambrogetti D, Risso G, Catarzi S, Morrone D, Mantellini P, et al. The role of arbitration of discordant reports at double reading of screening mammograms. *J.Med.Screen.* 2005;**12**:125-7.
24. Shaw CM, Flanagan FL, Fenlon HM, McNicholas MM. Consensus Review of Discordant Findings Maximizes Cancer Detection Rate in Double-Reader Screening Mammography: Irish National Breast Screening Program Experience. *Radiology* 2009;**250**:354-62.
25. 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.
26. 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.
27. Khoo LAL, Taylor P, Given-Wilson R. Computer-aided detection in the United Kingdom National Breast Screening Programme: Prospective study. *Radiology* 2005;**237**:444-9.
28. Hukkinen K, Kivisaari L, Vehmas T. Impact of the number of readers on mammography interpretation. *Acta Radiol*. 2006;**47**:655-9.
29. Lang K, Andersson I, Rosso A, Tingberg A, Timberg P, Zackrisson S. Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the Malmo Breast Tomosynthesis Screening Trial, a population-based study. *Eur.Radiol*. 2016;**26**:184-90.
30. Skaane P, Bandos AI, Gullien R, Eben EB, Ekseth U, Haakenaasen U, et al. Prospective trial comparing full-field digital mammography (FFDM) versus combined FFDM and tomosynthesis in a population-based screening programme using independent double reading with arbitration. *Eur.Radiol.* 2013;**23**:2061-71.
31. Skaane P, Bandos AI, Gullien R, Eben EB, Ekseth U, Haakenaasen U, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a populationbased screening program. *Radiology* 2013;**267**:47-56.
32. Hofvind S, Geller BM, Rosenberg RD, Skaane P. Screening-detected Breast Cancers: Discordant Independent Double Reading in a Population-based Screening Program. *Radiology* 2009;**253**:652-60.
33. Klompenhouwer EG, Voogd AC, den Heeten GJ, Strobbe LJA, Tjan-Heijnen V, Broeders MJM, et al. Discrepant screening mammography assessments at blinded and non-blinded double reading: impact of arbitration by a third reader on screening outcome. *Eur.Radiol*. 2015;**25**:2821-9.
34. Cornford EJ, Evans AJ, James JJ, Burrell HC, Pinder SE, Wilson ARM. The pathological and radiological features of screen-detected breast cancers diagnosed following arbitration of discordant double reading opinions. *Clin.Radiol*. 2005;**60**:1182-7.
35. Klompenhouwer EG, Weber RJP, Voogd AC, den Heeten GJ, Strobbe LJA, Broeders MJM, et al. Arbitration of discrepant BI-RADS 0 recalls by a third reader at screening mammography lowers recall rate but not the cancer detection rate and sensitivity at blinded and non-blinded double reading. *Breast* 2015;**24**:601-7.
36. Kerr NL, Tindale RS. Group performance and decision making. *Annual Review of Psychology* 2004;**55**:623–655
37. Bankier AA, Levine D, Halpern EF, Kressel HY. Consensus interpretation in imaging research: is there a better way? *Radiology* 2010;**257**:14-7.

38. Groenewoud JH, Otten JDM, Fracheboud J, Draisma G, Van Ineveld BM, Holland R, et al. Cost-effectiveness of different reading and referral strategies in mammography screening in the Netherlands. *Breast Cancer Res.Treat*. 2007;**102**:211-8.
39. Cawson JN, Nickson C, Amos A, Hill G, Whan AB, Kavanagh AM. Invasive breast cancers detected by screening mammography: A detailed comparison of computer-aided detection-assisted single reading and double reading. *Journal of Medical Imaging and Radiation Oncology* 2009;**53**:442-9.
40. James JJ, Cornford EJ. Does computer-aided detection have a role in the arbitration of discordant double-reading opinions in a breast-screening programme? *Clin.Radiol.* 2009;**64**:46-51.

S1. Articles included in the review

1. Klompenhouwer et al (a) (2015) *Netherlands – Quality CASP criteria met*						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Effect of arbitration by a 3rd reader discrepant reading for blinded and non-blinded double read- screening mammography Mammographic abnormalities and tumour characteristics of cancers missed after arbitration.	Retrospective review Quantitative design	Double reading Blinded and non-blinded alternated on a monthly basis Discrepant readings were always recalled Retrospectively reviewed by a 3rd radiologist – blinded to outcome. Used BI-RADS classification	Consecutive series of 84,927 mammograms 1st July 2009 -1st July 2011. 3 units – 12 radiologists, 1-15 years of screening mammography experience. FFDM Discrepant cases randomly assigned	Recall rate, cancer detection rate, proportion of BI-RADS 0 among all recalls, PPV, programme sensitivity. Cancers not recalled after arbitration by a third reader calculated as interval cancers. Independent-sample t- test. (95 % CI). Chi square and Fisher's exact tests - differences in tumour and mammographic characteristics of the reading strategies, differences in surgical treatment. P-value < 0.05	Discrepant readings =57.2 % blinded vs. 29.1% non-blinded, (p< 0.001), Blinded double reading, arbitration= 1. Decreased recall rate (3.4 to 2.2 %, p< 0.001) 2.decreased sensitivity (83.2 to 76.0 %, p = 0.013) 3.No influence on cancer detection rate (CDR; 7.5 to 6.8 per 1,000 screens, p = 0.258) 4. Increased the PPV; 22.3 to 31.2 %, p <0.001). Non-blinded double reading, arbitration = 1. Decreased recall rate (2.8 to 2.3 %, p < 0.001) 2.increased PPV (23.2 to 27.5 %, p=0.021) 3.no affect on affected CDR (6.6 to 6.3 per 1,000 screens, p=0.604) 4.no affect on sensitivity (76.0 to 72.7 %, p=0.308). No differences in the proportion of DCIS, smaller tumours, lymph node Involvement or advanced tumours among SDCs and cancers missed at arbitration. Invasive cancers with axillary lymph node metastasis were less often seen among cancers Missed at arbitration (20.3 % vs. 11.1 %,	Weakness – Acknowledged by the author arbitration outcome did not affect "real-life". Discrepant cases were recalled regardless. Therefore, the arbitrator's role did not have clinical implications for the screening. Strengths - Waited 2 yr. screening period to capture "interval cancers". True sensitivity calculated. Prior films available Number of radiologists with variable experience reflects clinical practice Large case series

					p<0.001	

2. Klompenhouwer et al (b) (2015) *Netherlands* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Evaluate PPV, discrepant rate, and characteristics of BI-RADS 0 recalls screening program. Determine the effect of arbitration by a 3rd reader of discrepant BI-RADS 0 readings at both reading strategies.	Retrospective review Quantitative design	Double reading Blinded and non-blinded -alternated on a monthly basis. Discrepant readings were always recalled Retrospectively reviewed by a 3rd radiologist – blinded to outcome. Used BI-RADS classification	Consecutive series of 84,927 1st July 2009 – 1st July 1 2011. 3 units – 12 radiologists 1-15 years of screening mammography experience. FFDM Discrepant cases randomly assigned	Chi square or Fisher exact test - differences in categorical variables PPV of recall of BI-RADS categories. Cancers not recalled after arbitration by a third reader were calculated as interval cancers. Continuous variables - double sided t-test for independent samples P-value < 0.05	Arbitration of discrepant BI-RADS 0 recalls = lowered recall rate (from 3.4% to 2.8% at blinded double reading, p < 0.001, and from 2.8% to 2.5% at non-blinded double reading, p 1⁄4 0.008), without a decrease in cancer detection rate (from 7.5‰ to 7.3‰, p 1⁄4 0.751, and from 6.6‰ to 6.5‰, p 1⁄4 0.832, respectively) and program sensitivity (from 83.2% to 81.2%, p 1⁄4 0.453, and from 76.0% to 74.6%, p 1⁄4 0.667, respectively). Arbitration would have significantly increased the PPV at blinded double reading (from 22.3% to 26.3%, p 1⁄4 0.015). 13 cancers missed by arbitration - overall decrease in cancer detection rate is very small, 0.1-0.2% at both reading strategies No differences in mammographic and tumour characteristics of BI-RADS 0 Recall at blinded and non-blinded reading	Weakness – Acknowledged by the author arbitration outcome did not affect "real-life". Discrepant cases were recalled regardless. Therefore, the arbitrator's role did not have clinical implications for the screening. No cost-effectiveness Strengths– waited 2 yr. screening interval to capture "interval cancers". Large case series Number of radiologists with variable experience reflects clinical practice

3. Hofvind et al (2009) *Norway* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Analyse discordant and concordant screen detected breast cancers using independent double reading with consensus. Arbitration only if consensus not reached by initial reporters	Retrospective review Quantitative design	Double reading Blinded reading Score 1-5 1, normal; 2, probably benign; 3, indeterminate; 4, probably malignant; and 5, malignant. Initial score of 2 or higher by either reader = a consensus meeting Initial score of 3 or higher – can't be dismissed without agreement from initial reporter Arbitration only if consensus not reached by initial reporters	1 033 870 prevalent and incident screens 5611 screen detected cancers (DCIS + invasive) 1996–2005 Radiologists Average experience = 4.3 years (range, 1–11 years), average volume for the whole study period (9 yrs.) = 19, 745 screening mammograms range, 525–107 161. SFM= 97% FFDM = 3%	Differences in rates and proportions tested with a x2 test. All tests were two-sided. *P* values <0.05. Logistic regression to estimate the odds that a discordant cancer was associated with mammographic density. Odds ratios (ORs) with 95% CI - adjustment for age at screening and prevalent vs. incident screening K Statistics - for agreement between two readers. Unweighted K values for 2 x 2 table analyses (positive and negative scores) Quadratic weighting for five-point interpretation scale. Observer agreement, k values < 0.20 =poor agreement; 0.21– 0.40, fair agreement; 0.41– 0.60, moderate agreement; 0.61–0.80, good agreement; and more than 0.81, very good agreement SPSS	Discordant scores = 5.3% Concordant positive scores = 2.1% At consensus, 66.8% (36 380 of 54 447) of the discordant and 17.9% (3932 of 21 928) of the concordant screenings were dismissed. Recall rate = 3.5% 23.6% (1326 of 5611) of CA had discordant interpretation. Varied from 16.9% (148 of 874 cancers) to 28.6% (265 of 928 cancers) according to county 117 interval breast cancers were diagnosed among the 40 312 screenings that were dismissed at consensus = 6.5% of all interval cancers.	Weakness – Acknowledged by author - Don't know if score correlates with actual CA and if the 2 reporters recalled for the same abnormality as quadrant and lesion characteristics not specified at initial interpretation 2 radiologists read less than 500 screening mammograms during 1 year in study period. Against the exclusion criteria No cost effectiveness Strengths - Large case series Specialist and general radiologists – representative of a community setting, but no information provided on the amount of time non-specialists dedicate to breast
4. James and Cornford (2009) *UK*. - Quality CASP criteria met						

Research question/aim	Study design	Reporting/ Arbitration strategies	Method Data Collection Sample size	Data Analysis/ Metrics	Main findings/results	Strengths/ Weaknesses
Can computer-aided detection (CAD) act as an arbitrator of discordant double-reading opinions, replacing the need for an independent 3rd film reader.	Retrospective review Quantitative design	Double reading Not completely blind Original arbitration by independent 3rd reader – radiologist Arbitration Mammograms digitised and analysed by CAD system – compared to radiologist CAD algorithms set to operate at a detection sensitivity of 88% for masses and 95% for micro calcifications.	240 cases underwent arbitration from 16,629 cases July 2003-April 2004. 5 radiologists, 1 research fellow, 1 radiographic film reader Radiologists experience ranged - 5-18 yrs. radiographer - 5 years	Statistical significance - McNemar test to take into account the matched nature of the data.	Arbitration cases accounted for 22% (112/518) of total cases recalled for assessment. 47% cases recalled to assessment following the opinion of the arbitrator 21 cancers in arbitration set, 13 diagnosed at the time of the original screening mammogram, 8 diagnosed subsequently. 3 were not the arbitrated lesion, **5 were** – 2 of these were assessed and returned to RR. CAD correctly prompted in these 5 cases. 2 cancers recalled by arbitrator and not CAD Independent 3rd reader recalled 15/18 (83%) of the cancers that corresponded with the arbitrated lesion. CAD as the arbitrator would have recalled 16/18 (89%) of the cancers that corresponded to the arbitrated lesion. CAD= **significant increase in normal women being recalled to assessment in the arbitration group (P < 0.001).** Extra 50 recalls. Recall rate increase from 3.1 to 3.4%; increase of 10%. Overall –No. Of cancers detected were broadly similar with 1 additional cancer recalled by CAD	Strengths - Reporters included radiographer – represents current UK practice Weakness – acknowledged by author -Small number of cancers in the series (18) Retrospective - can only give an indication as to the potential effect of CAD acting as an Arbiter No cost effectiveness Not completely blinded reading - may influence the proportion of discordant cancers.

5. Mucci et al (1999) *UK* - Quality CASP criteria met	

Research question/aim	Study design	Reporting/ Arbitration strategies	Method Data Collection Sample size	Data Analysis/ Metrics	Main findings/results	Strengths/ Weaknesses
Experience of double reading – breast screening 3rd person arbitrator	Prospective study Quantitative Design	Double reading Non-blinded 3rd reader decision= final decision. Non-blinded.	398 arbitration cases 1992-1994 3 radiologists	% Calculated for recall rates	398 arbitration cases - final reader recalled 196 (49%) and returned 202 (51%) to routine recall – 1 true interval CA subsequently Of 196 assessed - 4 malignant. Estimated cost saving by arbitration £20,000– 202 women returned to normal screening Assessment episode is £101, 3rd read=£1 (1999 figures) 3rd reader =reduction in no. Of recalls and no reduction in cancer detection.	Weakness – acknowledged by author -non-blinded 2nd reader knew the opinion of the first and was influenced. Therefore, underestimate the benefits of double reading to cancer detection. Strength - 3rd reader was aware of the opinion of the first two; simply asked to arbitrate on the action to be taken on an identified lesion – real clinical practice
6. Liston and Dall. (2003) *UK* - N/A for CASP -audit						
Method for assessing performance of new readers Arbitration	7yr Audit	Double read Non blinded Independent review by 3rd reader. Majority opinion is acted upon.	1/4/95 - 31/3/02 5 radiologists Varying experience	% Calculated for Cancers incorrectly returned to RR by 1st and 2nd reader Total no. Of cancers detected through double reading	The % of cancers detected with double reading + 3rd reader arbitration varied each year -3.6 and 11.4% Overall 87 (8.1%) of the 1072 cancers were detected following 3rd reader arbitration.	Strength - Robust audit

7. Cornford et al (2005) *UK* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Compare the mammographic background pattern, mammographic and pathological features of screen-detected cancers diagnosed following arbitration of discordant double reading opinions with cancers diagnosed following concordant double reading.	Retrospective review Quantitative design	Double reading Not entirely blinded 3rd reader arbitrator – had final decision. Independent decision – but not blinded to initial reports	April 2002 - December 2003 32,613 screened 431 arbitration cases 5 radiologists, 1 research fellow 1 radiographic film reader. Radiologists' experience ranged from 5–18 yrs. Film reader =5 yrs. experience.	Chi-square and Fisher's exact tests. Comparison of normally distributed, continuous variables, such as patient age, was analysed with unpaired t-test with Stat- View	287 malignancies. 38 (14%) had undergone arbitration and 249 (86%) had concordant double reading. 50% of arbitrated cases were recalled for assessment -38 malignant [PPV=18%]. Arbitration cases accounted for 20% of the total recalls. Arbitration group – 1st reader did not recall 27 malignancies; 2nd reader did not recall 11 malignancies. Arbitration group =27 invasive cancers and 11 DCIS. Concordant group = 196 invasive cancers and 47 DCIS. = No significant difference between 2 groups. No significant difference in proportion detected through a first or subsequent screen in the two groups ($p<0.7$). Cancers detected following arbitration were more likely to manifest as parenchymal distortions $p<0.001$ and less likely to manifest as spiculate	Weakness - 2nd reader not entirely blinded – may affect cancer detection rates, but does reflect normal clinical practice. Only 2/5 radiologists as arbitrators Only 1-year f/u – too short to assess all interval cancers Strength - Arbitrator not blinded –reflects normal clinical practice Reader workforce representative of UK practice, radiographer included. All with substantial experience.

					masses p<0.014). Less likely to be detected in fatty breasts p<0.01). Were smaller (p<0.045). Lobular cancers were commoner in the arbitration group, although this was of borderline significance, p<0.057 Estimated -11% more cancers are detected as a result of double reading with arbitration compared with single reading alone, after taking into consideration second reader bias.	

8. Caumo et al (2011) *Italy* - Quality CASP criteria not met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Role of arbitration by 3rd reader of discordant double readings to reduce recall rates	Prospective Quantitative Design	Double reading 3rd reader – **only 1 person used** Assessment performed irrespective of arbitration results	15/9/09 - 15/1/10 298 arbitrated cases **Only 1 radiologist as the arbitrator** >30 years' experience FFDM	Observed differences were checked by the chi-square ($\chi 2$) test, p *value <0.05.*	Recalls rate at double reading =6.8%. 230 (43.5%) were concordant + 298 (56.5%) were discordant. After arbitration classified – 216 (72.4%) negative + 82 (27.6%) positive 43 (18.6%) cancers were in concordant group 6 (2%) discordant recalls 5 were recalled 1 CA would have not been recalled Arbitration = reduced 216 assessment procedures (2.8% absolute, 40.9% relative reduction of recall rate) missed 1 CA (0.13‰ absolute, 2.0% relative reduction of cancer detection rate). Arbitration had a sensitivity of 83.3% Arbitration cost calculated as adding 3rd reader = 0.25 euros Assessment cost = 67.4–110.4 euros per Discordant readings, often resolved by additional views or ultrasound = lower cost to concordant recalls, more likely to require a biopsy. Based on above - Arbitration cost = 74 euros, 216 spared assessment =14,558.4–23,346 euros. Bias adjusted for by doubling the cost per mammography reading to 0.50 euros and by reducing the cost per assessment procedure to 50 euros. Arbitration = saved cost of 10,651 euros.	Weakness - Only 4-month period in study Only used 1 radiologist as the 3rd reader who had extensive experience >30yrs –not representative of the majority All cases were assessed and therefore the arbitrator's role did not have clinical implications for the screening. Comment Author acknowledged, "some imprecision of cost estimates might have occurred". 1st reading-cost estimates calculated from an excellence centre – does not reflect the average National scenario.

9. Ciatto et al (2005) *Italy* - Quality CASP criteria not met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Effectiveness of arbitration of discordant double readings in mammography screening	Prospective study Quantitative design	Double reading Does not state if blinded Arbitration – 3rd reader	2000–4, 1217 cases 9 radiologist readers 7 radiologist arbitrators Experience - mammograms (at least 10,000 mammograms read and at least three years of screening experience).	% Of sensitivity/ NPV /recall rates	1217 discordant double readings 476 cases (39.2%) arbitrated to assessment, detecting 30 cancers (6.3%). Of 741 negative arbitrations (60.8%), 311 F/U thus far = 2 cancers (0.64%) occurred in the site previously suspected at one of the two independent readings. Assumed Arbitration sensitivity = 86.3% NPV 99.3%. Arbitration reduced the overall referral rates from 3.82% to 2.59% (relative decrease 32.1%). false-negative arbitration, cancers detected per 1000 women screened would decrease from 4.58 to 4.50 (relative decrease 1.7%). 2005 standards: cost per arbitration = 4 euros, assessment 147 euros. For every 1 cancer missed due to arbitration - 151 recalls and 21,248 euros would have been saved, whereas the saved cost per screened woman due to arbitration was 1.72 euros.	Weakness - Only followed up 42% so far so estimated cancer detection rate. Rates transposed to full population screening to give the sensitivity/NPV recall etc. NOT continuous cases -limited to periods when radiologists were available to perform a 3rd third read Strengths - Acknowledged by author - cost analysis cannot be generalized to any other setting, as costs may vary substantially from one country to another and possibly among different centres.

10. Cawson et al (2009) *Australia* - Quality CASP criteria met	

Research question/aim	Study design	Reporting/ Arbitration strategies	Method Data Collection Sample size	Data Analysis/ Metrics	Main findings/results	Strengths/ Weaknesses
Compare double reading and arbitration (BP) for discordance, with CAD Invasive CA only	Retrospective cases. Quantitative design	1. Single read 2. CAD-assisted single Reading 3. Double reading - blinded	January 1998 to December 2001 Total 1569 cases 157 randomly selected double-read Invasive cancers were mixed 1:9 with normal cancers. 2 Radiologists Reader A - (>5000 cases/year) 7 years screening experience Reader B - senior radiology Trainee - 6 months training 3rd reader (10 years' experience Reading >5000 cases/year) Verified whether lesions recalled by the readers corresponded to cancers.	95% CI Comparison of sensitivities of 2 reading methods - Stata 'prtest' T-tests - to compare mammographic diameters. ROC curves plot sensitivity against specificity	The CAD system was highly Sensitive (93%, 95% CI 87.8–96.5), detecting many cancers overlooked by the readers, but the readers rejected most TP prompts CAD prompts are numerous and mostly FP. BP sensitivity = 90.4% CAD+RA sensitivity =86.6% (P = 0.12) CAD+RB 94.3% (P = 0.14). CAD-RB specificity was less than BP (P = 0.01). After CAD, reader's sensitivity increased 1.9% and specificity dropped 0.2% and 0.8%. **Arbitration decreased specificity 4.7%.** ROC analysis = BP accuracy better than CAD+RA, borderline significance (P = 0.07), but not CAD-RB. Cancers recalled after arbitration (P = 0.01) and CAD-R (P = 0.10) was smaller. No difference in cancer size or sensitivity between reading methods was found with increasing breast	Weakness - Prior mammograms were not available – may affect a reader's decision to recall Relatively high ratio of cancers to normal cases in the test set Readers had no prior inexperience with CAD Don't know what level of sensitivity the CAD system was set to. Only 2 readers utilised. Trainee as 1 of readers although sensitivity higher than experienced radiologist Strengths - Excluded cancer cases that were previously detected by the readers to eliminate bias due to recollection. Waited 2 yr. screening interval to capture "interval cancers".

					density. CAD-R and BP sensitivity and cancer detection size were not significantly different.	

11. Taylor and Potts (2008) *UK*.- Quality CASP criteria met						
Research question/aim	Study design	Reporting/ Arbitration strategies	Method Data Collection Sample size	Data Analysis/ Metrics	Main findings/results	Strengths/ Weaknesses
Compare single reading with CAD to single reading without CAD Compare double reading to single reading Arbitration and consensus.	Systematic review	1.Single reading 2. Double reading 3. Consensus 4. Arbitration studies	1991-2008 27 studies in total	Meta-analysis using the 'metan' command in Stata 8.2. Becker–Balagtas marginal estimated odds ratios Fixed effects models (using the Mantel–Haenszel method), random effects models (DerSimonian and Laird method) when heterogeneity as high.	Heterogeneity within each of the groups for recall rates. Arbitration/consensus studies, $p < 0.001$ Overall, arbitration studies show a **decrease in recall rates,** but two, including one of the largest studies, show a significant **increase.** Double reading – recall rates with arbitration - overall pooled estimate for the odds ratio is 0.94 (95% CI: 0.92, 0.96; **v**2 (1) = 30.1, $p < 0.001$). As a risk difference, this is a reduction of 2.67 per 1000 (95% CI: –1.72, –3.62; $z = 5.49$, $p < 0.001$). Random effects models - pooled estimate for arbitration/consensus studies is lower, but a larger confidence interval means that the result is marginally not significant (OR = 0.87; 95% CI: 0.75, 1.02; $z = 1.67$, $p = 0.095$). Double reading with arbitration **increased detection rate** (confidence interval (CI): 1.02, 1.15) and decreases recall rate (CI: 0.92,	Strengths - Met all the CASP criteria – transparent methodology

					0.96). Double read – cancer detection rates with arbitration/consensus – overall pooled estimate for the odds ratio is 1.08 (95% CI: 1.02, 1.15; V 2(1) = 6.2, p = 0.012) and the risk difference is 0.44 per 1000 (95% CI: 0.10, 0.79; z = 2.50, p = 0.012). For double reading with arbitration, the number needed to treat is 2222 women screened for each additional cancer detected. CAD does not have a significant effect on cancer detection rate (CI: 0.96, 1.13) and increases recall rate (95% CI: 1.09, 1.12). Evidence that double reading with arbitration enhances screening is stronger than that for single reading with CAD.	

12.Groenewoud et al - (2007) *Netherlands* - Quality CASP criteria not met						
Research question/aim	Study design	Reporting/ Arbitration strategies	Method Data Collection Sample size	Data Analysis/ Metrics	Main findings/results	Strengths/ Weaknesses
Compare reporting strategies – cost effectiveness 1.decision by one of the readers 2. Refer if both agree (consensus) 3.arbitration by a 3rd reader	Retrospective cases Quantitative design	Blinded reading 1.single reading; 2.double reading with referral if any Reader suggests 3. Double reading with referral only if both radiologists agreed	26 radiologists volunteered 10 read all films 18 read sub-sets Test set of 500 cases 250 controls 125 screen-detected Cancers 125 interval cancers	MIcrosimulation SCreening ANalysis (MISCAN) to estimate cost-effectiveness	Double reading with referral if any reader suggests resulted in a 1.03 times **higher sensitivity** (76.6%) and a 1.31 times **higher referral rate** (1.26%) than double reading with consensus. **Figured assumed – extrapolated** Assuming a relative increase of the detection rate by 2% and a relative increase of the referral rate by 30% double reading with referral if any reader suggests is comparably cost-effective to double reading with consensus (**e** 2,168 and **e** 2,207 per life-year gained, respectively). Control cases concordant =90.2% 89.4% both readers=normal case. 0.8% they both recommended referral. Cases concordant =75.2% 59.3% both readers=normal case 15.9% they both recommended referral. Of all readings by the 153 radiologist pairs, 17.7% were discrepant. Referral rates were highest with decision-making by consensus =73.8% decision by 1 reader = 57.4% arbitration = 52.7%	Weakness - Experimental setting not reflective of daily practice Used published regional Data to estimate the distribution of concordant and discrepant readings Assumed that each referral of a case would lead to the diagnosis of cancer

13. Lång et al (2016) *Sweden* - Quality CASP criteria met						
Research question/aim	Study design	Reporting/ Arbitration strategies	Method Data Collection Sample size	Data Analysis/ Metrics	Main findings/results	Strengths/ Weaknesses
Performance of one-view digital breast tomosynthesis (DBT) in breast cancer screening. Arbitration	Prospective one-arm single-institution study Quantitative design	Blinded reading Double reading and scoring Arbitration = **at least two readers** decided on recall irrespective of the score on the other modality Conventional 2 view DM 1 view (MLO) DBT	January 2010 to December 2012 Aim for 15,000 this study reports first half - 7500 cases 6 radiologists 5 = > 10 years' experience 1 reader =< 10 years' experience Mean 26 years, range 8 to 41 years) Individual training in interpretation of DBT images	McNemar's test for paired data of DBT and DM screens for differences in detection and recall rates with 95 % CIs. Differences in characteristics between cancers detected solely by DBT and all DM-detected cancers tested using chi-2 test and Fisher's Exact test, if the sample size was small. Analyses -Stata software (version 13). 80% power ROC analysis	Recall rate after arbitration was 3.8 % (3.3 to 4.2) for DBT and 2.6 % (2.3 to 3.0) for DM ($p<0.0001$). The PPV was 24 % for both DBT and DM.	Strength - Large prospective cohort Readers had DBT experience Weakness - Interim analysis - does not have 80% power at this stage

14.Duijm et al, (2004) *Netherlands* - Quality CASP criteria met						
Research question/aim	Study design	Reporting/ Arbitration strategies	Method Data Collection Sample size	Data Analysis/ Metrics	Main findings/results	Strengths/ Weaknesses
Determine the value of arbitration by a panel for discordant screen reads Arbitration and consensus	Prospective design Quantitative design	Blinded reading Double reading Mutual consensus between 2 readers. Persistent discordance went to arbitration panel = 3 Radiologists different to original reporters Referred to assessment if at least one arbitration member considered necessary. 3 panel radiologists aware of discordant reads but Blinded to results of the other arbitration panellists.	July 1, 1998, and January 1, 2001. 65,779 cases screened 332 discrepant cases 8 radiologists Experience in reading screening mammograms varied from 15 to 36 months (mean, 31 months).	% Or recall rates, cancer detection rates	Concordant referral = 498 (0.8%) of 65,779 screened Concordant normal = 64,949 (98.7%) women. Initial Discordant = 332 (0.5%) cases. After a mutual consultation, disagreement persisted 183 (0.3%) mammograms. Arbitration panel referred 89 of 183 cases. CA = 20 (22%) cases. 3 (3%) of the 94 not referred by the panel, breast cancer was detected at the site of previously discrepant mammographic findings seen at subsequent screening performed 2 years later. **Arbitration panel missed** If all 183 discrepant cases had been referred, the referral rate would have increased from 0.8% to 0.9% at subsequent (incident) screenings and from 1.5% to 1.7% at initial screenings. At subsequent screenings, the number of cancers detected per 1,000 women screened would have increased from 4.4 to 4.5.	Strength - 2yr. screening interval complete Able to assess no. Of interval cancers. Prior films available Blinding of arbitrator to other arbitrators

15. Khoo et al *(2005) UK* - Quality CASP criteria met	

Research question/aim	Study design	Reporting/ Arbitration strategies	Method Data Collection Sample size	Data Analysis/ Metrics	Main findings/results	Strengths/ Weaknesses
Recall and cancer detection rates with and without computer-aided detection (CAD) of discrepant cases-screening Consensus	Prospective design Quantitative design	Blinded reading Double reading - by at least 1 radiologist Each reader viewed current and available prior mammograms for each case – recorded an opinion CAD prompts for the current mammograms displayed - reader reassessed the prompted areas before recording a revised assessment Arbitration cases - discussed by an additional 2 consultant radiologists reviewed current/prior images, CAD prompts, and proforma	March 21, 2003, and January 9, 2004, 6111 case – images digitized 1639 cases arbitrated 12 readers – 7 radiologist + 5 radiographers 4 to 23 years' experience - Mean of 11 years	Relative sensitivity was calculated for each of three protocols (i.e., single reading, single reading with CAD, and double reading) Recall and cancer detection rates 95% CI Estimates for the time spent on arbitration per reader by monitoring time taken and number of cases arbitrated over a 3-week period	62 CA detected. CAD prompted 51 (84%) of 61 radiographically detected cancers. Of 12 cancers missed on single reading, 9 were correctly prompted; 7 prompts were overruled by the reader. Sensitivity Single reading was 90.2% Single reading with CAD was 91.5% Double reading without CAD was 98.4% 1639 cases arbitrated 39% recalled to assessment 61% - routine recall More women were allocated to arbitration when mammograms were read with CAD -13.8% to 10.5% non CAD More women were recalled for assessment in the CAD group -6.1% to 5% non-CAD Cancer detection rates = no difference	Strength - Prior mammograms available if possible Weakness - The sensitivity the CAD system was set to is not mentioned True false-negative rate – can't be calculated 3 years of follow-up needed. Unable to assess if any cancers were arbitrated to normal and have developed since

16. Posso et al (2016) Spain - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Costs and health-related outcomes of double versus single reading of digital mammograms in a breast cancer-screening programme. Arbitration and consensus	Retrospective cases	Blinded Double reading Discrepant reads first discussed by consensus persistent discrepant cases went for arbitration by 3rd third senior radiologist	June 2009 until May 2013, 57,157 cases 4 radiologists (2010 value for costings)	Student's t-test, Chi-square Test, and Fisher exact test. Statistical tests were two sided P values < 0.05 Analyses were performed using Microsoft Excel (2011) and IBM SPSS software version 21.0 (SPSS, 2013).	Discordance between radiologists in 4.5 %(**N**= 2,556) cases 98.1 % (**N**= 2,508) resolved by consensus and 1.9 % (**N** = 48) by arbitration **Estimate affect** Cost. Double reading without consensus and arbitration was 14 % (€ 36,341) more expensive than double reading with consensus and arbitration. Health-related outcomes. Double reading without consensus and arbitration had 1.5 % more false positive results than double reading with consensus and arbitration ($p < 0.001$). Both reading strategies had similar cancer detection rates ($p = 0.986$). Double reading with consensus and arbitration was 15%(Euro 334,341) more expensive than single reading with first reader only. False-positive results were more frequent at double reading with consensus and arbitration than at single reading with first reader only (4.5 % and	Weakness - No interval cancer rates -results are not conclusive Did not calculate the cost-effectiveness of reading strategies

					4.2 %, respectively; P <0.001). Single reading could reduce the frequency of false positive results without changing the cancer detection rate.	

17.Dinnes et al (2001) UK - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Compare double reading with single reading of mammograms for screening accuracy, patient outcomes and costs. Arbitration and consensus	Systematic review	Single reading and Double reading For double reading recall policies 1. Recall if 1 suggests 2. Arbitration 3.consensus 4. Mixed Mixture of blinded and non-blinded Double reading	April 1991 -July 1999 10 cohort studies met inclusion criteria Only 3 studies evaluated for sensitivity and specificity		Consensus or arbitration or a mix of the two, decreased recall rates (by between 61 and 269 per 10,000 women screened). Insufficient evidence was available to detect any pattern in cancer detection according to recall policy. Specificity increased with consensus or mixed recall. Unable to analyse cost effectiveness as significant variation between the organisation of services from different countries Unable to quantify a difference on cancer detection rates from the results.	Strength - Met CASP criteria

18. Skaane et al (2013) *Norway*- Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Assess cancer detection rates, false-positive rates before arbitration, PPV for women recalled after arbitration, and the type of cancers detected with use of FFDM alone and combined with tomosynthesis	Prospective trial Quantitative design Interim analysis – phase 1	Blinded Double reading Consensus based arbitration meeting. 1. Mammography alone, 2.mammography + CAD 3.mammography + tomosynthesis 4. Synthesized mammography + Tomosynthesis	November 22, 2010, to December 31, 2011. 12631 cases 8 radiologists with 2–31 yrs. of experience in screening Images scored 1-5 One score of 2 or greater in at least one arm were discussed at arbitration before a consensus-based decision was made. Consensus-based arbitration meetings = min 2 radiologists	Analyses were based on marginal log linear models for binary data, accounting for correlated interpretations and adjusting for reader-specific performance levels by using a two-sided significance level of .0294 Cancer detection rates, false positive rates before arbitration, and PPV for patients recalled after arbitration.	**False-positive rates before arbitration were 61.1 per 1000 examinations with mammography alone and 53.1 per 1000 examinations with mammography + tomosynthesis (15% decrease, adjusted for reader; P, .001).** 5 of 8 radiologists referred proportionally more patients for arbitration with use of mammography alone than with use of mammography + tomosynthesis. **Overall number of women recalled as a result of arbitration was larger for those initially assigned a positive score at mammography + tomosynthesis (351 vs. 265 women).** However, the concordant increase in the detection of 24 additional Cancers resulted in a similar PPV for the cases ultimately recalled after arbitration (29.1% mammo alone and 28.5% + tomo)	Weakness - Only limited data about interval cancers -cannot estimate conventional absolute sensitivity or specificity. Estimate relative performance levels Potential candidates were selected on the basis of whether technical staff members and imaging systems were available to perform the additional imaging examination

19. Wolf et al (2015) *Germany* -Quality CASP criteria met	

Research question/aim	Study design	Reporting/ Arbitration strategies	Method Data Collection Sample size	Data Analysis/ Metrics	Main findings/results	Strengths/ Weaknesses
Performance of 3 collective intelligence rules ("majority", "quorum", and "weighted quorum") applied to mammography screening	Prospective Quantitative design	Majority, quorum and weighted quorum against individual radiologist performance	182 test set cases Images from 2000-2003 from 6 centres 101 radiologists randomly grouped into sizes (range: 1 to 15)	Average true and false positive rate of the no. of radiologists determined by a training set to give the quorum threshold Weighted quorum	As group size increased, all three CI rules achieve increases in true positives and decreases in false positives. Larger groups made more accurate decisions Marginal affect when group size exceeds 9 relatively small group sizes achieved performance improvements Overall decision accuracy = Weighted quorum rule slightly outperforms the quorum rule and that the quorum rule outperforms the majority rule	Strength - Large number of radiology participants – representative of diverse experience Unique, transparent system of consensus without 'over-ruling' of a group face-to-face setting. Weakness – Test set, no influence on "real-life" cases.

20. Blanks et al (1998) U*K* - Quality CASP criteria not met (No for Q6 against cohort study)						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Cancer detection rates for different reading strategies. Breast screening Consensus and arbitration	Observational epidemiological study Quantitative	1. Single reading 2. Double reading (with recall if any reader suggests) 3.double reading (With recall if both readers agree, consensus) 4. Double reading (with arbitration by a third or more radiologists) 5. Double (complex)	1 April 1996 to 31 March 1997. 87 screening units	Cancer detection rate adjusting for confounding by age using Poisson regression 95% CI	Prevalent screen **Double (consensus) = 1.26 SDR** **referral rate = 6.8** **Double (arbitration) = 1.28 SDR** **Referral rate =7.3** Incident screen invasive cancer SDR - **Double (consensus) = 0.98 SDR** **Referral rate = 3.1** **Double (arbitration) = 1.10 SDR** **Referral rate =4.0** Incident screen invasive cancer SDR <15 mm Double (consensus) =1.00 Double (arbitration) =1.18	Strength - Multi –Centre study Weakness - 1yr study

21. Skaane et al (2013) *Norway* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Compare double readings for FFDM (2D) and tomosynthesis (3D) during mammographic screening.	Prospective study Quantitative design	5-point rating scale for probability of cancer: 1=normal or definitely benign; 2=probably benign; 3=Indeterminate 4=probably malignant 5=malignant. Scores of 2 or greater in at least one reading arm =discussed at arbitration, with at least two radiologists Consensus-based decision for all cases with a least one rating of 2 or 3. Cases with a score of 4 or 5 were recalled and **could not be dismissed at consensus.**	22/11/10 – 31/12/11 8 Radiologists - 2– 31 years of experience (average 16 years) in screening mammography	P<0.05 Type III test -in generalised linear mixed Model (proc glimmix, v. 9.23) Heterogeneity of performance - addressed using G-side random effects	74% of mammo only cases – returned to routine recall at consensus. 26% recalled. 75% of these negative at assessment 61% of mammo +tomo – returned to routine recall at consensus. 39% recalled. 74% of these negative at assessment Pre-arbitration false-positive scores were 10.3 % mammo only and 8.5 % for 2D+ 3D (P<0.001). Recall rates were 2.9 % (365/12,621) and 3.7 % (463/12,621), respectively (P=0.005). PPV Mammo only before arbitration= 6.5% after = 24.7 % 2D+ 3D before arbitration= 10% after = 25.5 %	Strength - Scores recorded directly into the NBCSP database -results locked at the end of each reading Weakness - Unable to assess outcome of cases dismissed at arbitration – 1 yr. study

22. Hukkinen et al (2006) *Finland* - Quality CASP criteria not met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Conference consensus (the Majority considered decisive) Or Independent reading of several radiologists (the positive opinion of at least a single reader considered Decisive).	Prospective Quantitative	Double reading Conference consensus = the majority opinion in the group	1997 – 2001 200 Test cases 4 radiologists 5 -18 yrs. screening experience 2 general radiologists, 2 residents, 6 months - 4yrs. of experience in Clinical mammography.	Sensitivity/ Specificity	The greatest sensitivity of 74.5% = readings of the four best-performing readers were combined. Sensitivity very variable Sensitivity maximal when any positive opinion within a pair or a group of readers is taken into consideration. Conference reading = improved specificity	Weakness - Small number – test cases High ratio 1:4 cancers to normal cases – not representative of normal practice Actual consensus where Readers discuss discordant findings did not happen in order to avoid a situation in which one reader is overruled by another. Worked out by calculating average sensitivities

23. Matcham et al (2004) *UK* - Quality CASP criteria met						
Research question/aim	Study design	Reporting/ Arbitration strategies	Method Data Collection Sample size	Data Analysis/ Metrics	Main findings/results	Strengths/ Weaknesses
Affect of consensus on all discordant **and** concordant recalls	Retrospective and Prospective Quantitative	Consensus for all cases even if both initial readers 'recalled'	April 1997 - March 2002. 2 years prior to the start of the consensus meeting, and the 3 completed years since. 3 radiologists – 3-12 yrs. Experience 1 film reader – 4yrs experience	PPV, cancer detection rates SDR	5% of screening cases discussed at consensus meeting (n=2637) 65.6% recalled after consensus 3 interval cancers subsequently diagnosed after RR outcome following consensus – 1 true and 2 minimal signs 97 (10.7%) of the women returned to routine screening had been marked for recall by both original film readers. Consensus of all cases - Reduction in recall rates Increase in Specificity	Strength - Sufficient follow-up period to assess interval cancers and true sensitivity

24. Jenkins et al (2014) *UK* - N/A for CASP Audit						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Assess differences in the film-reading histories of interval or screen detected cancers Arbitration	Audit – 3 year period	Double reading Not completely Blind reading Arbitration by 3rd reader – radiologist – not blinded has access to previous opinions	2004 -2007 4 programmes within the East Midlands Film readers – radiologists and radiographers Analogue films	Cancer detection rates, confidence intervals, and chi square Tests with Monte Carlo simulation.	Double reading= discordance in 13,279 cases (5%) underwent arbitration. 9726 (73%) were returned to routine rescreen, 3553 (27%) were recalled PPV for unanimous recall = 22.7% PPV for recall following arbitration = 8.3% 4.1% of interval cancers with no previous recall outcomes were false negatives, which was significantly lower compared to the groups where at least one reader had indicated recall (10.9%; p. 0.005). Cancers detected at the subsequent screen demonstrated no significant difference in prognosis dependent on previous film-reading history (P. 0.503).	Strengths - Robust method for identifying interval cancers.

25. Shaw et al (2009) *Ireland* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Consensus review of discordant Screening mammography	Prospective Quantitative	Double reading Blinded reading Consensus panel = Three to five consultant radiologists and usually included one or both of the original readers. Recall -if any member of the Consensus panel recommended after discussion.	2000-2005 5 radiologists 3–10 years of screening experience. Two consultants who had just completed fellowship training participated for 2 years of the study period.	Sensitivity/specificity Z test (95% CI P<0.05	Discordant cases = 1.04% After consensus, 45.39% recalled **11.7% of these were cancer** Highest reader recall = could potentially increase the cancer detection rate by 0.6 per 1000 women screened but would increase the recall rate by 12.69% and the number of False-positive findings by 15.37%. Conclusion: The consensus panel identified 71 (7.33%) of 968 cancers diagnosed. Consensus review substantially reduced the number of cases recalled and was associated with a low false-negative rate. **1.1% of known cancers missed by consensus review**	Weakness - 44 (6%) cases at consensus sent to RR with no follow-up. False-negative findings was predicted by multiplying the number of patients who did not return for a follow-up visit (n -44) by the percentage of false-negative findings in patients with follow-up screening data

26. Per Skaane et al (2007) *Norway* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Compare SFM and FFDM in a population-based screening program. Consensus	Prospective Quantitative	Double reading Blinded 5-point rating scale for probability of cancer: 1=normal or definitely benign; 2=probably benign; 3=indeterminate 4=probably malignant 5=malignant. Consensus meeting. Free to dismiss cases with scores no higher than 2 by one or both readers.	November 2000, and December 2001. Radiologists	Recall rate, cancer detection rate, PPV	74.1% of SFM case dismissed at consensus meeting; 68.9% of FFDM were dismissed at consensus 10.9 – 11.1% cancers missed by consensus 25-30% cancers only recalled by 1 reader	Weakness - 45-49 age group not complete follow-up? Accurate interval cancer rate reported