Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate

Paul Taylor and Henry WW Potts

Centre for Health Informatics and Multiprofessional Education, University College London, United Kingdom

Corresponding author:

Paul Taylor

Centre for Health Informatics and Multiprofessional Education

University College London

Archway Campus, Highgate Hill

London N19 5LW

tel: +44 20 7288 3548

email: p.taylor@chime.ucl.ac.uk

Abstract

Background

There are two competing methods for improving the accuracy of a radiologist interpreting screening mammograms: computer aids (CAD) or independent second reading.

Methods

Bibliographic databases were searched for clinical trials. Meta-analyses estimated impacts of CAD and double reading on odds ratios for cancer detection and recall rates. Sub-group analyses considered double reading with arbitration.

Results

Ten studies compared single reading with CAD to single reading. Seventeen compared double to single reading. Double reading increases cancer detection and recall rates. Double reading with arbitration increases detection rate (CI: 1.02-1.15) and decreases recall rate (CI: 0.92-0.96). 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). However, there is considerable heterogeneity in the impact on recall rate in both sets of studies.

Conclusion

The evidence that double reading with arbitration enhances screening is stronger than that for single reading with CAD.

Keywords: Mammography; Diagnosis, Computer-Assisted; Image Interpretation, Computer-Assisted; Double Reading

Introduction

In many countries, including the UK, it is standard practice for each screening mammogram to be viewed independently by two readers who either confer on discordant cases or refer them for arbitration. It is sometimes argued that this 'double reading' is too expensive or too demanding of radiologists' time.(1) An alternative is to use computer programs that process digitised mammograms and alert readers to possible abnormalities. A systematic review identified six studies comparing CAD to double reading but concluded they were methodologically flawed and the evidence was limited.(2) This paper takes a different approach: two sets of studies are reviewed:

- studies comparing single reading with CAD to single reading without CAD;
- studies comparing double reading to single reading.

We assess the impact of both interventions on cancer detection and recall rate since an improvement in cancer detection rate at the cost of an increased recall rate may not present an enhancement of the screening test.

Methods

Criteria for considering studies for this review

Types of studies

Prospective and retrospective studies where the intervention was incorporated into routine screening work and all cases selected only on the basis of the usual screening criteria were included.

Types of participants

All studies of women in a screening age range (aged 40 and above) were considered.

Types of interventions

Only studies using commercially available CAD systems were included. Studies of double reading in which the second reader was a trained film reader but not a radiologist were included.

Types of outcome measures

Only studies reporting the impact of the interventions on cancer detection rate and recall rate, or for which these could be calculated or otherwise obtained, were included.

Strategy for identification of studies

The NLH PubMed database was searched using MeSH terms “Mammography” and either "Diagnosis, Computer-Assisted", “Image Processing, Computer-Assisted” or “Image Interpretation, Computer-Assisted”, or the text string “CAD”; and using MeSH term “Mammography” and text strings “double reading”, “second reading” or “second reader”. Google Scholar, Biotech, CINAHL, Embase, HMIC, Pyschinfo, Web of Science and Science Direct were searched using the strings “mammography” and “computer” or “CAD”, and “mammography” and “double reading”. The online catalogue of the British Library and recent proceedings of relevant conferences were searched. A previous systematic review of double reading was identified and its references checked,(3) as were references in retrieved papers. Immediately prior to publication (14th Feb 2008) we repeated the Medline searches with the same search terms, checking for articles added to Medline in the last six months and also hand-searched for current and future publications in the journals which had published the articles identified in the initial search.

Methods of the review

Retrieved articles were assessed against the pre-defined criteria. Full copies of papers potentially meeting the inclusion criteria were obtained. Each author separately extracted data and differences were reconciled.

Statistical analysis

Four meta-analyses were performed for the impact of CAD and double reading on cancer detection and recall.

Two designs are used in studies of CAD. In some, the radiologist's assessment before viewing the computer prompts is compared with their final assessment having seen the prompts. The assessment, before and after using CAD, is on the same mammogram, so we term these 'matched' studies. Other studies compare the performance of mammography facilities before and after the introduction of CAD. Different mammograms are interpreted in the two conditions. These studies are 'unmatched'. The meta-analyses should take into account this design difference and combine both types of study. We use Becker-Balagtas marginal estimated odds ratios (4). This method treats matched data as if it was unmatched, but with a correction to the estimated variance of the log odds. However, with large sample sizes (as here) the correction is trivial and results are presented as if unmatched. Key results were repeated using risk differences, with no correction for matching. Meta-analyses were performed using the "metan" command in Stata 8.2. (5) We fitted fixed effects models (using the Mantel–Haenszel method), but used random effects models (DerSimonian & Laird method) when heterogeneity was high.

Subgroup analyses

Matched and unmatched studies of CAD were analysed separately and together. Most UK centres do double reading with consensus or arbitration on discordant cases. However in some studies all discordant cases are retrieved, in others a mixed strategy or a mix of strategies are used. Results for these three subgroups (consensus/arbitration, unilateral and mixed) were analysed separately and together.

Results

Description of studies

Ten prospective studies comparing single reading with CAD to single reading without CAD were identified,(6-15) and 17 comparing single reading with double reading. (14-31)

Studies comparing single reading to single reading with CAD

The initial bibliographic searches for studies of CAD identified 2012 citations, from which 210 abstracts were reviewed and 19 papers retrieved. Of the retrieved papers not included, four were excluded since the results they reported were contained in other papers that were included,(32-35) three described studies comparing CAD to double reading rather than single reading,(36-38) and four were on selected cases not an unselected sequence

of screening cases.(39-42) Two of the included papers were published after the initial search and identified when the searches were repeated immediately prior to publication.(14;15)

Table 1 summarises the ten included studies: six matched and four unmatched. One unmatched study also includes comparative data on facilities that never adopted CAD.(11) We exclude this data but show the results of including it, and, since this study generated some criticism, (42) also show the results of excluding the study completely. In another paper, cancer detection was assessed using a matched design and recall rate using an unmatched design. (8) This paper noted that recall rate fluctuated over the study period: we used the figure for the period over which the cancer detection rate was measured.

Age of the screening population is given as a mean or median. Radiologists' experience is given as a mean or a range. Study duration is in months. There is one multi-centre study, for this the range of durations at each site is given.(11) All studies were conducted in the USA.

Studies comparing single reading to double reading

Initial bibliographic searches for studies of double reading found 335 citations, from which 72 abstracts were reviewed and 28 papers retrieved. Thirteen papers were excluded: four based on data reported in papers already included(43-46), one on selected cases and not an unselected sample of

screening cases (47) and six that did not report the recall rate under single reading. (47-53) A study using pre-screeners was excluded, since the intention was not to have all films double-read. (54) One study compared programmes using double and single reading using standardised detection rates.(55) This was excluded as the data are adjusted for prevalence and could not be compared with the cancer detection rates used elsewhere. Two of the included papers were published after the initial search and identified when the searches were repeated immediately prior to publication.(14;15) Three further studies identified in the updated search were excluded. One compared two approaches to double reading, one compared double reading with analogue vs double reading with digital and one reported the features of cancers detected by the second reader. (56-58)

Table 2 summaries the 17 included studies. All use a matched design: recall and cancer detection rate are measured under double reading and the performance of the first reader used as a proxy measure of single reading. In four studies, data on the performance of the first reader is not presented but recall and cancer detection rates for single reading can be calculated on the assumption that half the discordant cases can be attributed to the first reader. (17;19;22;29)

One paper provides recall rates under single reading for five readers who read 80% of the cases.(23) The mean of these values is used as the recall rate for single reading. The recall rate under double reading was

supplied on request by the study author. In another study, the recall rate for single reading was obtained from a subsequent review article.(24) Deans and colleagues report a recall rate for single reading for only a subset of data: only this data is included.(28) Two papers published by Ciatto and colleagues overlap, data from the first are included in the meta-analyses. (25;43)

Only two studies recorded the mean age of participating women (15;19); otherwise the age range is given. Only a few studies specified the years since qualification of participating readers, although others gave details of qualifications, special training or volume of films read. In two studies, radiographers were used as additional readers. In both, films could be third read by additional radiologists.(20;21) In two studies, only one reader was an experienced mammographer, the other a general radiologist. (15;30)

Data synthesis

Figure 1 shows forest plots summarising the four meta-analyses.

Impact of CAD on cancer detection rate

Studies of the impact of CAD on cancer detection rate are shown in Figure 1a. There is no evidence of heterogeneity between or within the matched and unmatched studies: overall test, $\chi^2(9) = 1.07$, $p = 1.00$, $I^2 < 0.1\%$; testing between the two sub-groups, $\chi^2(1) = 0.40$, $p = 0.53$. None of the studies shows a statistically significant increase in cancer detection rate and neither group shows a pooled effect. The overall estimate of the effect is an

odds ratio of 1.04 (95% confidence interval: 0.96, 1.13) that is not significant ($\chi^2(1) = 0.86$, $p = 0.35$). A similar result is found using a risk difference metric: the overall pooled estimate is 0.16 *per* 1000 (95% confidence interval: -0.17, 0.48; $z = 0.93$, $p = 0.35$). Using figures for all clinics in Fenton and colleagues produced a similar result, as did omitting the study entirely. (72)

Impact of double reading on cancer detection rate

Figure 1b shows the impact of double reading on cancer detection rate. There is no evidence of heterogeneity: overall test, $\chi^2(16) = 5.1$, p = 1.0, $I^2 <$ 0.1%; testing between the three sub-groups, $\chi^2(2) = 1.4$, $p = 0.50$. Although individually the reported effects are mostly not significant, the pooled estimate is significant (95% confidence interval: 1.06-1.14; $\chi^2(1) = 23.5$, $p < 0.001$). A similar result is found using a risk difference metric: overall pooled estimate is 0.44 per 1000 (95% confidence interval: 0.26, 0.62; $z = 4.84$, $p < 0.001$). For arbitration/consensus studies, the overall pooled estimate for the odds ratio is 1.08 (95% confidence interval: 1.02, 1.15; $\chi^2(1) = 6.2$, $p = 0.012$) and the risk difference is 0.44 per 1000 (95% confidence interval: 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.

Impact of CAD on recall rate

The evidence on the impact of CAD on recall rate (Figure 1c) is less clear. All the studies showed increased recall rates, but there is strong evidence of heterogeneity: overall test, $\chi^2(9) = 148.1$, $p < 0.001$, $I^2 = 94\%$. The matched studies do not show heterogeneity: $\chi^2(4) = 3.6$, $p = 0.47$, $I^2 < 0.1\%$. However, the unmatched studies do: $\chi^2(4) = 143.6$, $p < 0.001$; $I^2 =$ 97%. A

significant result is found on the test for heterogeneity if either the studies of Fenton and colleagues (11) or Gur and colleagues (12) are included, but the other papers are mutually consistent.

The overall pooled estimate for the odds ratio is 1.10 (95% confidence interval: 1.09, 1.12), which is significant ($\chi^2(1) = 130.3$, $p < 0.001$), as are the estimates for the matched and unmatched studies separately. The marked difference between Fenton and colleagues, Gur and colleagues and the other large studies, is unexplained. However, the matched studies clearly show an increased recall rate and the sub-total pooled result is our best estimate of that effect (OR = 1.13; 95% CI: 1.08, 1.17), or expressed as a risk difference, 10.08 per 1000 (95% confidence interval: 6.59, 13.56).

Using the data for all clinics in Fenton and colleagues, pooled estimates and heterogeneity are reduced, but remain significant. The odds ratio for the unmatched studies would be 1.04 (95% CI: 1.01, 1.06), the overall pooled estimate 1.05 (95% CI: 1.03, 1.06). A similar result occurs if we omit this study entirely.

Given the remaining unexplained heterogeneity, a random effects model was also fitted. All the pooled estimates (matched, unmatched and overall) remain significant, the overall pooled estimate of the odds ratio being 1.13 (95% CI: 1.05, 1.23). A similar result is found if Fenton and colleagues is omitted.

Impact of double reading on recall rate

Studies of the impact of double reading on recall rate are summarised in Figure 1d. There is clear evidence of heterogeneity: overall test, $\chi^2(16) = 925.7$, $p < 0.001$, $I^2 = 98\%$. There is heterogeneity between the three groups ($\chi^2(2) = 513.5$, $p < 0.001$) and within each of the groups (for arbitration/consensus studies, $\chi^2(7) = 306.5$, $p < 0.001$, $I^2 = 98\%$; for mixed studies, $\chi^2(2) = 8.6$, $p = 0.014$, $I^2 = 77\%$; for unilateral studies, $\chi^2(5) = 97.2$, $p < 0.001$, $I^2 = 95\%$). It appears that different centres have different attitudes towards recall rate.

All the mixed and unilateral studies show increases in recall rate. Overall, arbitration studies show a decrease, but two, including one of the largest studies, (23) show a significant increase. For this study, recall rates under single reading were based on the five readers who read 80% of the cases. These, presumably more experienced, readers may have had a lower recall rate, biasing the comparison in favour of single reading which would explain why the result stands out. However if this study is omitted, the arbitration studies still show heterogeneity ($I^2 = 97\%$).

Just considering arbitration studies, the overall pooled estimate for the odds ratio is 0.94 (95% confidence interval: 0.92, 0.96; $\chi^2(1) = 30.1$, $p < 0.001$). As a risk difference, this is a reduction of 2.67 per 1000 (95% confidence interval: -1.72, -3.62; $z = 5.49$, $p < 0.001$). These analyses were repeated using random effects models. The pooled estimate for arbitration/consensus studies is lower, but a larger confidence interval means

the result is marginally not significant (OR = 0.87; 95% CI: 0.75, 1.02; z = 1.67, p = 0.095).

Discussion

Impact of CAD and double reading on cancer detection and recall rate

Matched CAD studies measure its impact more directly, comparing assessments on individual images before and after looking at prompts. Although all of these studies show an improvement in cancer detection rate, none shows a statistically significant improvement and their combined effect is not statistically significant. Since it is impossible to detect fewer cancers after taking a second look than were detected initially, these studies are biased in favour of CAD. They will not detect if unprompted cancers that might otherwise be detected are missed. Improvements which fail to achieve significance are therefore not necessarily promising.

The unmatched studies seem a more rigorous test. These studies however are susceptible to criticism. If the impact of CAD is assessed too soon after its introduction, results may be affected by a temporary drop in specificity as readers adjust to working with the prompts. Studies with longer assessment periods might also fail to detect a benefit since the extra cancers detected when CAD was introduced are not available to be detected later and the earlier detection is not revealed by a comparison of detection rates.

Particular criticism has been levelled at Fenton and colleagues (42) We found that the cancer detection rates in this study are consistent with others and its omission does not change our conclusion. Fenton and colleagues do find an unusually high recall rate, but omission of the study still produces a significant increase in recall.

The design of the double reading studies is similar to that of the matched studies of CAD: an audit in centres using double reading identifies cancers only detected by the second reader. Two studies record cancers which would have been recalled under single reading that were missed under double reading with arbitration. (16;19) Only two studies show a statistically significant improvement in cancer detection rate due to double reading, (28;30) however the meta-analysis shows a clear, statistically significant improvement. The pooled estimate suggests an extra 0.44 cancers detected *per* 1000 women screened.

Comparing the effects of CAD and double reading

Comparing pooled estimates of the effect sizes, the overall picture for double reading is that recall rate is increased, but it is lowered for double reading with consensus/arbitration. Figure 2 shows the confidence intervals for double reading with consensus/arbitration and for CAD. For cancer detection rate, the confidence interval for CAD mostly overlaps that for double reading with arbitration. However, there is a clear difference on recall rate,

which is significantly better for double reading with arbitration than for CAD. Even if CAD and double reading produce similar improvements in cancer detection rates, the reduced recall rate is a substantial advantage for double reading with arbitration.

The review also demonstrates the importance of arbitration/consensus in double reading. The introduction of an arbitration step allows readers to identify cases with minimal signs knowing that they will be reviewed and discussed by colleagues and only a proportion recalled. Eliciting extra assessments for difficult cases in this way allows a more efficient decision threshold to be maintained. A unit staffed by readers with different levels of experience should think carefully about which readers should work together and who should do the arbitration. Brown and colleagues carried out a cost-effectiveness analysis following their comparison of single reading and double reading with consensus. (18) They extended their analysis to include double reading without consensus, assuming that all women marked for recall by either reader would have been recalled. Consensus double reading was cheaper than single reading, saving £4853 per 10,000 women screened, whereas simple double reading cost £19529 more than single reading (costs based on 1994 prices).

There is unexplained heterogeneity in recall rate effects. Analysis is shown using fixed effects and a random effects model. A random effects approach yields enlarged confidence intervals. The increase in recall rate for

CAD is significant, but not the decrease for double reading with arbitration. However, the effect of double reading remains better than that for CAD.

CAD might be preferable to double reading on cost grounds. However, even slight increases in recall rates weaken this argument if it rests on the value of time saved by not double reading. It takes approximately 20 seconds to read a mammogram, but one hour to deal with a woman recalled from screening. (36)

It is worth noting that all the included studies of CAD but only three of the studies of double reading were conducted in the United States. There are differences in how screening operates in different countries and these might affect the impact of interventions such as CAD or double reading. Smith-Bindeman and colleagues reviewed differences between the UK and US screening programmes (of the 17 studies of double reading, five were conducted in the UK).(59) They found that recall rates were twice as high in US but that the cancer detection rates in the two countries were similar. Women are screened more frequently in the US than the UK (between the ages of 50 and 60, a women being screened in the US will average 7 screening visits compared to 3 in the UK) and more small and in situ cancers are detected. In addition to practising double reading, the UK programme enforces strict quality assurance criteria (UK radiologists read 5 to 7 times as many films annually as their US counterparts) and is under less pressure from malpractice litigation. It is unclear how these differences might affect the impact of CAD or double reading.

Implications for future research

Researchers have argued for an RCT to determine whether single reading with CAD is equivalent to double reading. Such a trial would provide more direct evidence than our review. However, a trial is only justified if we are in a state of equipoise about the two approaches. This review suggests otherwise.

Cancer detection rate is correlated with recall rate. One estimate is that each 1% added to the recall rate leads to 0.22 extra detections per thousand. (60) Our pooled estimates are in line with this, suggesting that CAD may change the threshold for recall rather than improve the accuracy of screening. The unexplained heterogeneity observed in recall rates should also be investigated.

The limited impact of CAD is surprising. It prompts for cancers that radiologists miss, but the prompts do not always affect decision-making. It is often assumed their impact is diminished by the high number of false positive prompts and that CAD developers must improve specificity. That assumption should be addressed in future research.

Conclusion

There is evidence that double reading increases cancer detection rate and that double reading with arbitration does so while lowering recall rate. There is insufficient evidence to claim that CAD improves cancer detection rates, but it does increase recall rate. Comparing CAD and double reading with arbitration, there is no difference in cancer detection rate, but double reading with arbitration shows a significantly better recall rate. Therefore, the best current evidence shows grounds for preferring double reading to single reading with CAD.

Acknowledgements

Dr Given-Wilson provided advice. Dr Liston supplied additional data. The work was partly supported by the NHS Breast Screening Programme which had no input into the research or presentation of results.

Conflict of interest statement

The authors have no conflicts of interest.

References

(1) Denton ER, Field S. Just how valuable is double reporting in screening mammography? *Clin Radiol* 1997;**52**:466-8.

(2) Bennett RL, Blanks RG, Moss SM. Does the accuracy of single reading with CAD (computer-aided detection) compare with that of double reading?: A review of the literature. *Clin Radiol* 2006;**61**:1023-8.

(3) Dinnes J, Moss S, Melia J, Blanks R, and FS. Kleijnen. Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: findings of a systematic review. *Breast*. 2001;**10**:455-63.

(4) Becker MP, Balagtas CC. Marginal modeling of binary cross-over data. *Biometrics* 1993;**49**:997-1009.

(5) Bradburn MJ, Deeks JJ, Altman DG. metan - an alternative meta-analysis command. *STATA Technical Bulletin Reprints* 1998;**8**:86-100.

(6) Morton MJ, Whaley DH, Brandt KR, Amrami KK. Screening mammograms: interpretation with computer-aided detection--prospective evaluation. *Radiology* 2006;**239**:375-83.

(7) Freer TW, Ulissey MJ. Screening Mammography with Computer-aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center *Radiology* 2001;**220**:781-6

(8) Dean JC, Ilvento CC. Improved cancer detection using computer-aided detection with diagnostic and screening mammography: prospective study of 104 cancers. *Am J Roentgenol* 2006;**187**:20-8.

(9) Ko JM, Nicholas MJ, Mendel JB, Slanetz PJ. Prospective assessment of computer-aided detection in interpretation of screening mammography. *Am J Roentgenol* 2006;**187**:1483-91.

(10) Birdwell RL, Bandodkar P, Ikeda DM. Computer-aided detection with screening mammography in a university hospital setting. *Radiology* 2005;**236**:451-7.

(11) Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med* 2007 ;**356**:1399-409.

(12) Gur D, Sumkin JH, Rockette HE, et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. *J Natl Cancer Inst* 2004 ;**96**:185-90.

(13) Cupples TE, Cunningham JE, Reynolds JC. Impact of computer-aided detection in a regional screening mammography program. *Am J Roentgenol* 2005;**185**:944-50.

(14) Georgian-Smith D, Moore RH, Halpern E et al. Blinded comparison of computer-aided detection with human second reading in screening mammography. AJR *Am J Roentgenol* 2007;**189**:1135-41.

(15) Gromet M. Comparison of Computer-Aided Detection to Double Reading of Screening Mammograms: Review of 231,221 Mammograms. *AJR Am J Roentgenol* 2008;**190**:1-6.

(16) Anttinen I, Pamilo M, Soiva M, Roiha M. Double reading of mammography screening films--one radiologist or two? *Clin Radiol* 1993;**48**:414-21.

(17) Williams SM, Doyle T, C hartres SC, Richardson AK and Elwood AJ. Impact of independent double reading of mammograms from the inception of a population-based breast cancer screening programme. *Breast* 1995 **4**: 282-288

(18) 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.

(19) Duijm LE, Groenewoud JH, Hendriks JH, de Koning HJ. Independent double reading of screening mammograms in The Netherlands: effect of arbitration following reader disagreements. *Radiology* 2004;**231**:564-70.

(20) Pauli R, Hammond S, Cooke J, Ansell J. Comparison of radiographer/radiologist double film reading with single reading in breast cancer screening. *J Med Screen* 1996;**3**:18-22.

(21) Tonita JM, Hillis JP, Lim CH. Medical radiologic technologist review: effects on a population-based breast cancer screening program. *Radiology* 1999;**211**:529-33.

(22) Renaud R, Schaffer P, Gairard B, et al. Principes et premiers résultats de la campagne européenne de dépistage du cancer du sein dans le Bas-Rhin. *Bull Acad Natl Med* 1991;**175**:129-45.

(23) Liston JC, Dall BJ. Can the NHS Breast Screening Programme afford not to double read screening mammograms? *Clin Radiol* 2003;**58**:474-7.

(24) Leivo T, Salminen T, Sintonen H, et al. Incremental cost-effectiveness of double-reading mammograms. *Breast Cancer Res Treat* 1999;**54**:261-7.

(25) Ciatto S, Ambrogetti D, Bonardi R, 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.

(26) Ciatto S, Del Turco MR, Morrone D, et al. Independent double reading of screening mammograms. *J Med Screen* 1995;**2**:99-101.

(27) Harvey SC, Geller B, Oppenheimer RG, Pinet M, Riddell L, Garra B. Increase in cancer detection and recall rates with independent double interpretation of screening mammography. *Am J Roentgenol* 2003;**180**:1461-7.

(28) Deans HE, Everington D, Cordiner C, Kirkpatrick and AE. Scottish experience of double reading in the National Breast Screening Programme. *Breast* 1998 **7**:75–79.

(29) Anderson ED, Muir BB, Walsh JS, Kirkpatrick AE. The efficacy of double reading mammograms in breast screening. *Clin Radiol* 1994;**49**:248-51.

(30) Seradour B, Wait S, Jacquemier J, Dubuc M, Piana L. Modalités de lecture des mammographies de dépistage du programme des Bouches-du-Rhône. *J Radiol* 1997;**78**:49-54.

(31) Agbaje OF, Astley S, Gillan M, et al.. Mammography reading with computer-aided detection (CAD): Single view vs two views. In Astley S, Brady M, Rose C and Zwiggelaar R (eds.) *Proceedings of the 8th International Workshop on Digital Mammography 2006* Berlin: Springer; 2006, p.125-129.

(32) Bandodkar P, Birdwell RL, Ikeda DM. Computer aided detection (CAD) with screening mammography in an academic institution: Preliminary findings. *Radiology* 2002;**225**:458.

(33) Destounis S. Computer-Aided Detection and Second Reading Utility and Implementation in a High-Volume Breast Clinic. *Appl Radiol* 2004;**33**:8-15.

(34) Morton MJ, Whaley DH, Brandt KR, Amrami KK. The Effects of Computer-aided Detection (CAD) on a Local/Regional Screening Mammography Program: Prospective Evaluation of 12,646 Patients. *Radiology* 2002;**225**:459.

(35) Destounis SV, DiNitto P, Logan-Young W, Bonaccio E, Zuley ML, Willison KM. Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience. *Radiology* 2004;**232**:578-84.

(36) Khoo LA, Taylor P, Given-Wilson RM. Computer-aided detection in the United Kingdom National Breast Screening Programme: prospective study. *Radiology* 2005;**237**:444-9.

(37) Gilbert FJ, Astley SM, Mcgee MA, et al. Single reading with computer-aided detection and double reading of screening mammograms in the United Kingdom National Breast Screening Program. *Radiology* 2006;**241**:47-53.

(38) Garvican L, Field S. A pilot evaluation of the R2 image checker system and users' response in the detection of interval breast cancers on previous screening films. *Clin Radiol* 2001;**56**:833-7.

(39) Menna S, Di Virgilio MR, Burke P, et al. Diagnostic accuracy of commercial system for computer-assisted detection (CADx) as an adjunct to interpretation of mammograms. *Radiologia Medica* 2005;**110**:334-40.

(40) Sittek H, Perlet C, Helmberger R, Linsmeier E, Kessler M, Reiser M. Computerassistierte Analyse von Mammographien in der klinischen Routinediagnostik. *Radiologe* 1998;**38**:848-52.

(41) Skaane P, Kshirsagar A, Stapleton S, Young K, Castellino RA. Effect of computer-aided detection on independent double reading of paired screen-film and full-field digital screening mammograms. *Am J Roentgenol* 2007;**188**:377-84.

(42) Fenton JJ, Barlow WE, Elmore JG. Computer-aided screening mammography. *N Engl J Med* 2007;**357**:85.

(43) Ciatto S, Ambrogetti D, Risso G, et al. The role of arbitration of discordant reports at double reading of screening mammograms. *J Med Screen* 2005;**12**:125-7.

(44) Thurfjell E. Mammography screening. One versus two views and independent double reading. *Acta Radiol* 1994;**35**:345-50.

(45) Thurfjell E. Mammography screening methods and diagnostic results. *Acta Radiol Suppl* 1995;**395**:1-22.

(46) Warren RM, Duffy SW. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. *Br J Radiol* 1995;**68**:958-62.

(47) Wivell G, Denton ER, Eve CB, Inglis JC, Harvey I. Can radiographers read screening mammograms? *Clin Radiol* 2003;**58**:63-7.

(48) Denton ER, Field S. Just how valuable is double reporting in screening mammography? *Clin Radiol* 1997;**52**:466-8.

(49) Mucci B, Athey G, Scarisbrick G. Double reading of screening mammograms: the use of a third reader to arbitrate on disagreements. *Breast* 1999 **8:**37-39

(50) Pacher B, Tscherney R, Litmann-Rowenta B, Liskutin J, Mazewski I, Leitner H et al. Konsensuelles Zweitbefunden von Mammographien in der Praxis. *Rofo* 2004;**176**:1766-9.

(51) Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. *Radiology* 1994;**191**:241-4.

(52) Van der Valk, P, Beijerinck D, Deurenberg JJ. Cost-effectiveness of consensus double reading of screening mammograms. *Radiology* 1998;**209**:588-9.

(53) Vizcaino I, Salas D, Vilar JS, Ruiz-Perales F, Herranz C, Ibanez J. Breast cancer screening: first round in the population-based program in Valencia, Spain. Collaborative Group of Readers of the Breast Cancer Screening Program of the Valencia Community. *Radiology* 1998;**206**:253-60.

(54) Haiart DC, Henderson J. A comparison of interpretation of screening mammograms by a radiographer, a doctor and a radiologist: results and implications. *Br J Clin Pract* 1991;**45**:43-5.

(55) 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**:95-201.

(56) Duijm LE, Groenewoud JH, Fracheboud J, de Koning HJ. Additional double reading of screening mammograms by radiologic technologists: impact on screening performance parameters. *J Natl Cancer Inst* 2007; **99**:1162-70.

(57) 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.

(58) Doutriaux-Dumoulin I, Allioux A, Campion L, Meingan P, Molina L. Cancers detectes par le deuxieme lecteur: analyse des donnees de la campagne de depistage du cancer du sein en Loire-Atlantique, 2003-2005 (nouveau cahier des charges). *J Radiol* 2007;**88**:1873-80.

(59) Smith-Bindman R, Chu PW, Miglioretti DL, et al. Comparison of screening mammography in the United States and the United kingdom. *JAMA* 2003; **290**:2129-37.

(60) Gur D, Sumkin JH, Hardesty LA, et al. Recall and detection rates in screening mammography. *Cancer* 2004;**100**:1590-4.

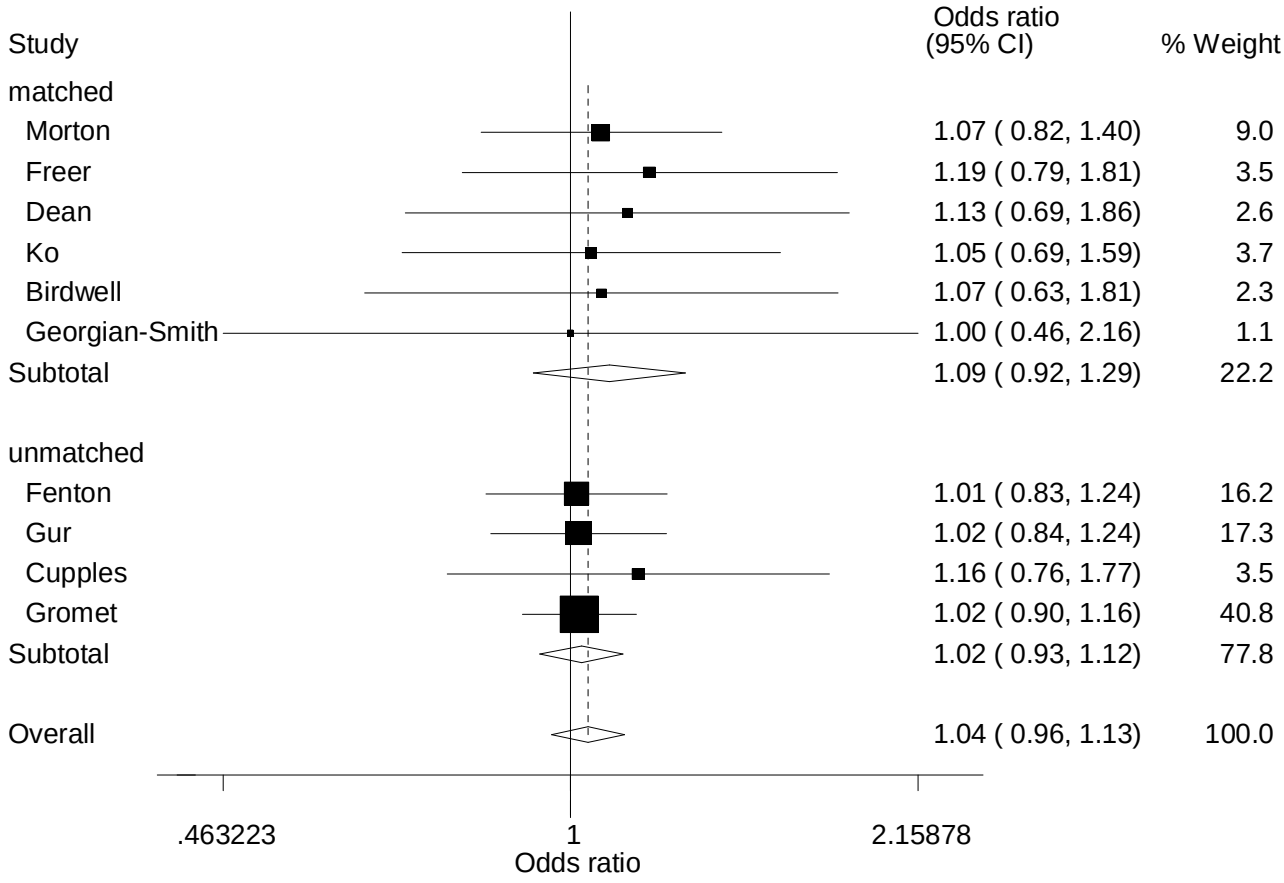

Figure 1a: the impact of CAD on cancer detection rate.

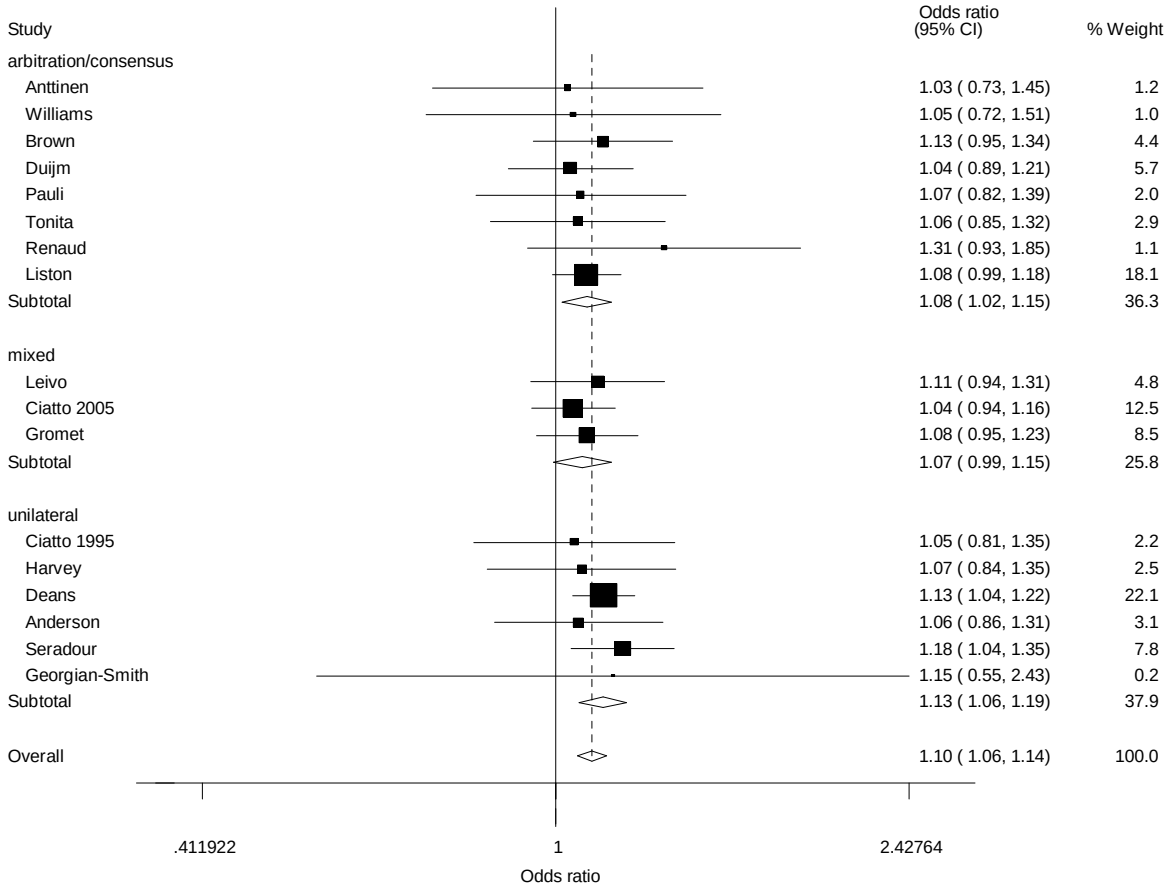

Figure 1b: the impact of double reading on cancer detection rate.

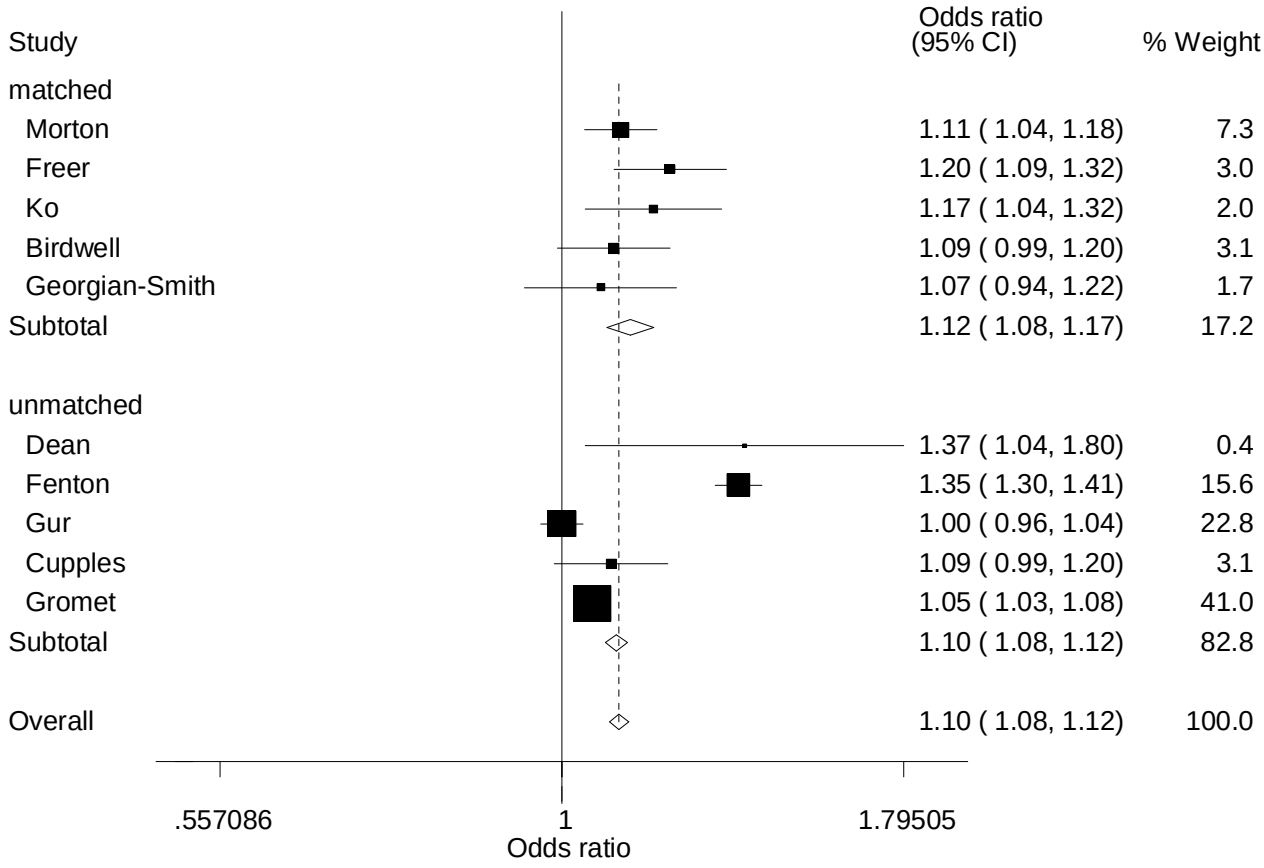

Figure 1c: the impact of CAD on recall rate

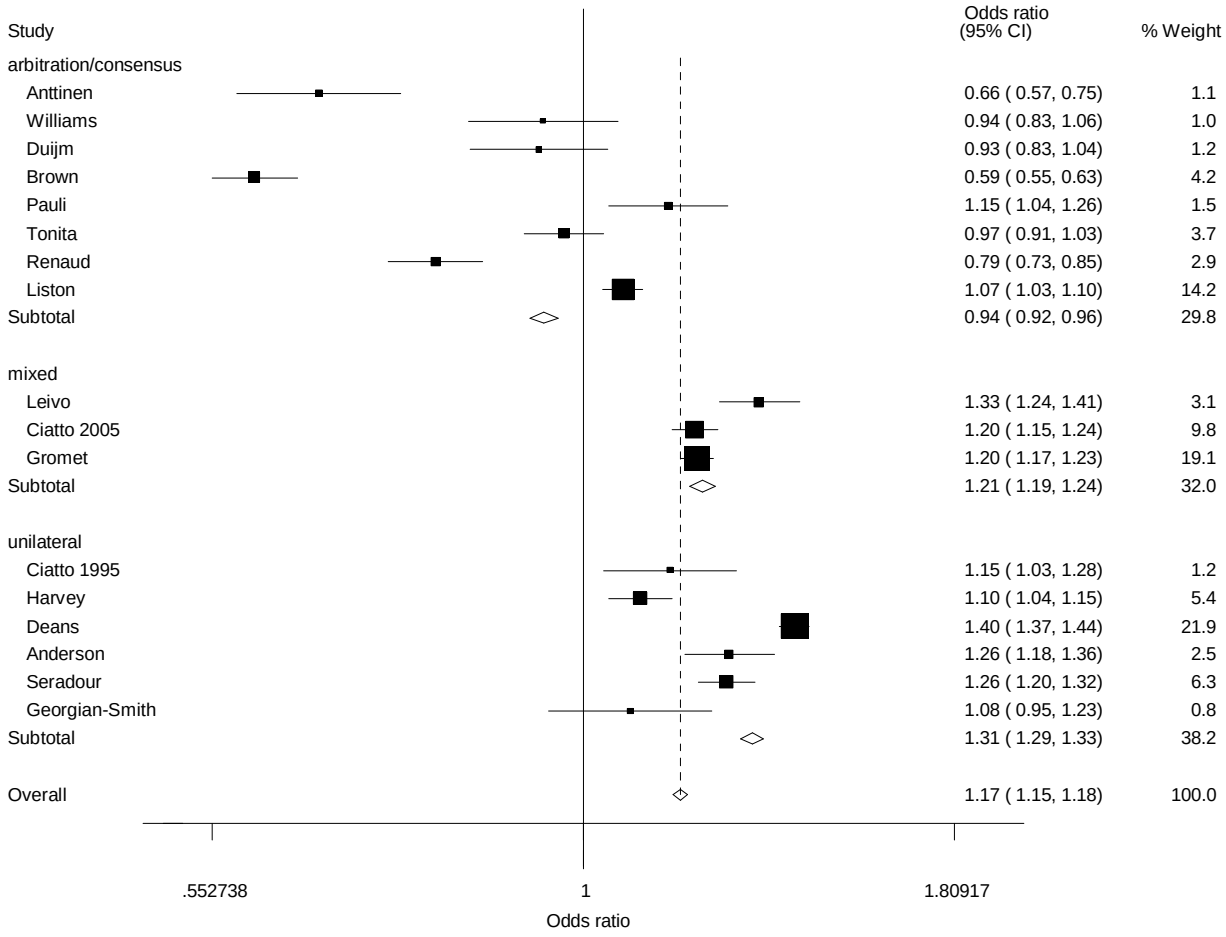

Figure 1d: the impact of double reading on recall rate

Caption for Figure 1: Forest plots of the included studies. Each study in the meta-analysis is shown as a horizontal line. The length of the line indicates the width of the 95% confidence intervals. The position of the midpoint shows the measured effect. The size of the centre square reflects the contribution to the pooled estimates, which is largely determined by sample size. The summary results are shown as diamonds. The centre of the diamond shows the combined estimate of the effect and the distance to the left and right

extremities shows the 95% confidence interval. Where the diamond sits wholly to one side of the mid-line, there is evidence of an effect. Summary results are calculated for the two approaches to studies of CAD and to the different forms of double reading and for the overall total.

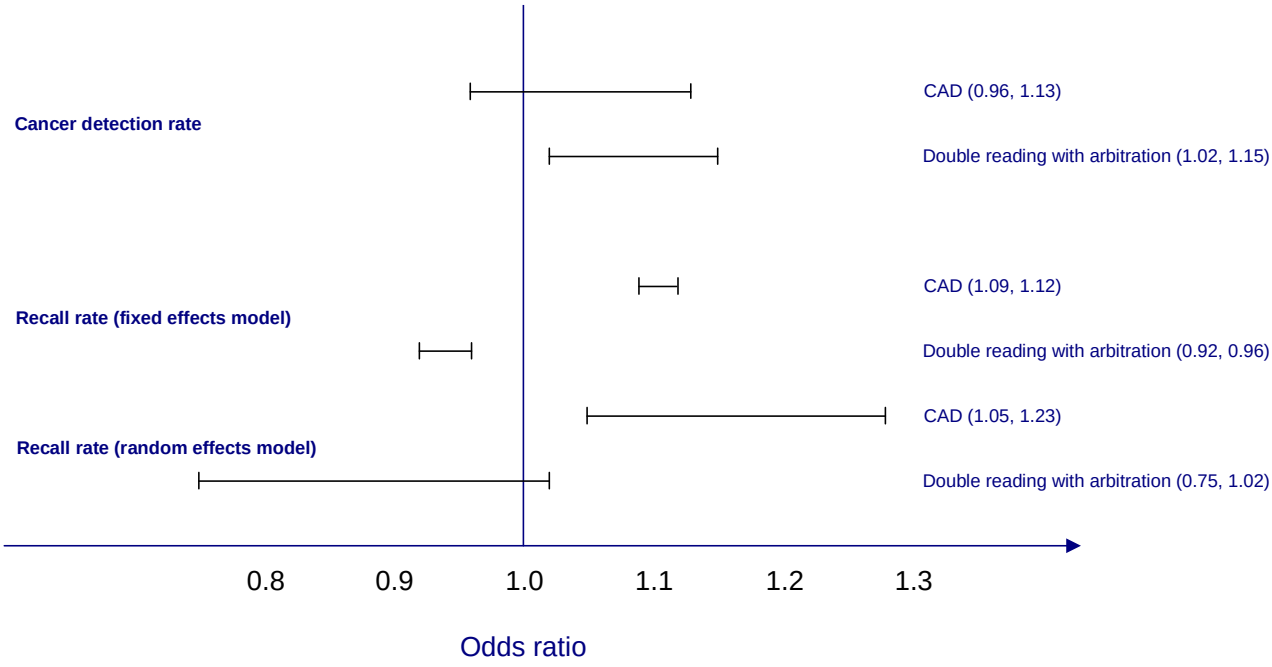

Figure 2: 95% confidence intervals for the pooled estimates of the effect sizes for the impact on cancer detection rate and recall rate of double reading with arbitration and of single reading with CAD.

Study	Year	Type	Sample size	Age of women in sample	Country	Number of readers	Experience in years	Study duration in months	CAD software used	Proportional contribution to CDR	Proportional contribution to recall rate
Freer[7]	2001	Matched	12,860	49	USA	2		15	R2 2.0	0.20	0.19
Birdwell[10]	2005	Matched	8682	54	USA	7	10,30	19	R2 2.2	0.07	0.08
Dean[8]	2006	Matched/ Unmatched	9520		USA	1	15	27	CADx 3.2	0.13	0.34
Ko[9]	2006	Matched	5016		USA	2	15	26	CADx 3.2	0.05	0.15
Morton[6]	2006	Matched	21349	60	USA	12	12	12	R2 v2.2	0.08	0.09
Cupples[13]	2004	Unmatched	27274	54.5	USA	4		24	R2 2.0	0.16	0.08
Gur[12]	2004	Unmatched	115571	50	USA	24		18	R2	0.02	0.00
Fenton[11]	2007	Unmatched	116086	55	USA	38	10,19	2,25	R2	0.01	0.31
Georgian-Smith[14]	2007	Matched	6381	Unknown	USA	8	14	22	R2	0.00	0.06
Gromet[15]	2008	Unmatched	118808	53	USA	9	15	48	R2 3.2	0.02	0.04

Table 1. Summary of included studies comparing single reading to single reading with CAD. The proportional impact on CDR is $(CDR_{CAD} – CDR_{control}) / CDR_{control}$. The proportional impact is on recall rate is $(RR_{CAD} – RR_{control}) / RR_{control}$.

Study	Year	Form of double reading	Sample size	Screening age range	Country	No. of readers	Reader's experience (years)	Study duration (months)	Proportional impact of double reading on CDR	Proportional impact of double reading on recall rate
Renaud[22]	1991	arbitration	17228	50-65	France			12	0.31	-0.19
Pauli[20]	1996	arbitration	17202	50-64	UK				0.06	0.14
Tonita[21]	1999	arbitration	27863	50-69	Canada	8		14	0.06	-0.03
Liston[23]	2003	arbitration	177167	50-64	UK	5		84	0.08	0.06
Duijm[19]	2004	arbitration	65779	59	Netherlands	8	31 months	30	0.04	-0.07
Anttinen[16]	1993	consensus	15457	50-59	Finland	4		15	0.03	-0.34
Williams[17]	1995	consensus	5659	50-64	NZ	2		18	0.04	-0.06
Brown[18]	1996	consensus	33734	50-64	UK	6		41	0.13	-0.39
Leivo[24]	1999	mixed	95423	50-59	Finland			60	0.11	0.32
Ciatto[25]	2005	mixed	177631	50-69	Italy	11		66	0.04	0.19
Anderson[29]	1994	unilateral	31146	50-64	UK	3	3,14	16	0.06	0.25
Ciatto[26]	1995	unilateral	18817	50-69	Italy				0.042	0.14
Seradour[30]	1997	unilateral	95967	50-69	France	126		24	0.18	0.25
Deans[28]	1998	unilateral	257212	50-64	UK	18		48	0.13	0.38
Harvey[27]	2003	unilateral	25369	>40	USA	7	3,18	18	0.070	0.08
Georgian-Smith[14]	2007	unilateral	6381	unknown	USA	8	3,26	22	0.15	0.18
Gromet[15]	2008	mixed	112,413	54	USA	9	15	48	0.08	0.07

Table 2. Summary of included studies comparing single reading to double reading. The proportional impact on CDR is ($CDR_{doublereading}$ – $CDR_{control}$) / $CDR_{control}$. The proportional impact is on recall rate is ($RR_{doublereading}$ – $RR_{control}$) / $RR_{control}$.