Published in final edited form as:

Breast Cancer Res Treat. 2010 April ; 120(3): 539–546. doi:10.1007/s10549-010-0770-x.

Predictors of interobserver agreement in breast imaging using the Breast Imaging Reporting and Data System (BI-RADS)

Anna Liza M Antonio and

UCLA School of Public Health, VA Greater Los Angeles Healthcare System

Catherine M Crespi

UCLA School of Public Health, Department of Biostatistics, and Jonsson Comprehensive Cancer Center, Division of Cancer Prevention and Control Research, University of California, Los Angeles

Abstract

Objective—The Breast Imaging Reporting and Data System (BI-RADS) was introduced in 1993 to standardize the interpretation of mammograms. Though many studies have assessed the validity of the system, fewer have examined its reliability. Our objective was to identify predictors of reliability as measured by the kappa statistic.

Methods—We identified studies conducted between 1993 and 2009 which reported kappa values for interpreting mammograms using any edition of BI-RADS. Bivariate and multivariate multilevel analyses were used to examine associations between potential predictors and kappa values.

Results—We identified ten eligible studies, which yielded 88 kappa values for the analysis. Potential predictors of kappa included: whether or not the study included negative cases, whether single-view or two-view mammograms were used, whether or not mammograms were digital vs. screen-film, whether or not the 4th edition of BI-RADS was utilized, the BI-RADS category being evaluated, whether or not readers were trained, whether or not there was an overlap in readers' professional activities, the number of cases in the study and the country in which the study was conducted. Our best multivariate model identified training, use of two-view mammograms and BI-RADS categories (masses, calcifications and final assessments) as predictors of kappa.

Conclusion—Training, use of two-view mammograms and focusing on mass description may be useful in increasing reliability in mammogram interpretation. Calcification and final assessment descriptors are areas for potential improvement. These findings are important for implementing policies in BI-RADS use before introducing the system in different settings and improving current implementations.

Keywords

Interobserver agreement; Kappa; Mammography; Breast cancer; BI-RADS

INTRODUCTION

According to the American Cancer Society, approximately 40,610 women are expected to be diagnosed with breast cancer in 2009 [1]. Mammographic screenings have proven to be successful in detecting early signs of breast cancer and preventing deaths by prompting early treatment, but their efficacy depends necessarily on the interpretations of their readers.

In order to address the inherent variability in mammographic interpretation, in 1993 the American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS) with the goal of standardizing mammographic reporting and improving communication between clinicians and radiologists. This lexicon provides a dictionary of terms for feature description as well as a list of final assessment categories and recommendations for follow-up. In 1997, the final results of the Mammographic Quality Standards Act of 1992 were published. According to this Act, all mammograms in the United States must be reported using one of the BI-RADS assessment categories. Since 1993, four editions of BI-RADS have been developed.

Many studies have focused on the sensitivity, specificity, and predictive values of the BI-RADS system [2,3,4]. These measures are important for evaluating the validity of the system and its performance in population-based screening. However, fewer studies have evaluated the interobserver variability of the system [5–14]. It is also important to evaluate the system's reliability since if readers cannot agree on an interpretation, its usefulness is jeopardized.

The objectives of this work are (i) to identify studies conducted between 1993 and 2009 that evaluated interobserver agreement in interpreting mammograms by reporting kappa values using any edition of BI-RADS; (ii) to identify potential predictors of kappa; and (iii) to examine associations between these potential predictors and kappa values.

METHODS

Search Strategy

Studies were identified in PubMed and Google Scholar using a search strategy which used the terms 'kappa' and 'BI-RADS' and 'mammography,' and by scanning references of pertinent articles. Search records were limited to those published between January 1, 1993 and April 29, 2009. The start year of 1993 was chosen because the first edition of BI-RADS was published in 1993.

Study Eligibility Criteria

We defined the following criteria for study eligibility: (i) computation of interobserver agreement using kappa statistics (Cohen's, Weighted or Fleiss'); (ii) the use of any of the four editions of BI-RADS to describe and/or assess mammographic lesions; (iii) in English or with an English translation.

Operational Definitions

We used the following definitions. Masses, calcifications, and final assessment were defined to be 'categories.' Each category contained characteristics of interest which we defined as 'descriptors.' Each descriptor was defined by a single word chosen from a set of words or 'terms.' To illustrate, we could describe the characteristic shape of a mass by using the term 'round.' Moreover, we could describe the final assessment and recommendation of a mass lesion to be a benign finding (BI-RADS category 2) with a recommendation to continue routine screening.

Kappa statistics are calculated as a difference between the observed agreement between readers and the expected agreement due to chance. For our study, we used an operational definition of kappa as a measure beyond chance of the agreement between readers on BI-RADS terminology use and/or assignment of final assessment categories. Calculations of kappa vary according to the manner in which expected agreement between readers is calculated or according to the distribution of categories. Cohen's kappa is strictly a measure between two readers, while Fleiss' kappa can extend to more than two readers. These two measures assume equality

between classes of agreement, but if classes are unequal in some respect a weighted kappa can be calculated. For our study, in addition to studies which reported Cohen's kappa and Fleiss' kappa we included studies which reported weighted kappa values since the magnitude of weighted kappa is interpreted in the same manner as unweighted kappa according to Fleiss [15].

Kappa Eligibility Criteria

The BI-RADS descriptor categories for which agreement was measured varied across studies. For some studies [10,12,13], kappa was calculated for each characteristic term, as well as at the descriptor level. In order to have kappa values consistent with our operational definition of kappa, we excluded values of kappa calculated at levels other than at the descriptor-level. In addition, some studies included results on categories not included in others. For example, agreement for 'architectural distortion' was measured only in one study [14]. To ensure comparability among kappas, we further refined our eligibility criteria to include only kappa values calculated on descriptors under the masses, calcifications, and final assessment categories.

Possible Predictors of Kappa

From the eligible studies, we abstracted potential predictors of kappa that were identified by study authors as either evidenced by their study or proposed as possible reasons for extreme values of kappa.

Analyses

Since the kappa values reported within each study were determined by the same set of observers, these values are correlated and thus we used a multilevel analysis with random intercept for study to account for the correlations and model the association between kappa and the potential explanatory variables. Bivariate regression was used to model the relationship between kappa and each explanatory factor singly. Variables with p-values less than 0.25 in bivariate analyses were examined in multivariate analyses. A set of multivariate models that included all possible combinations of these variables were compared using the AIC and BIC criteria; these model fit statistics can be used for nested as well as non-nested models. Lower AIC and BIC in absolute value indicates a better model. All analyses were performed using SAS 9.0 (SAS Institute, Cary, North Carolina).

RESULTS

Literature Search

The PubMed search yielded 23 citations, three of which were immediately excluded because interobserver agreement was evaluated on either ultrasound or magnetic resonance imaging data [16,19,21]. Two studies were written in German with no English translation [24,25] and thus excluded as well. Two studies evaluated only intraobserver agreement [26,27], one study compared observer responses to a study-defined gold standard [18], one reported only mean kappas [30], one measured agreement between craniocaudal and mediolateral oblique view ratings [22] and three measured agreement between film screen and digital mammogram ratings [17,28,29]. Lastly, we excluded two additional studies because agreement was not measured on BI-RADS terminology and/or final assessment categories. This resulted in a sample of eight papers. A search of Google Scholar yielded two additional papers [6,11] and examination of the reference list of each paper confirmed our final sample of ten papers. A description of the papers can be found in Table 1. From these ten papers, we abstracted 88 kappa values and corresponding covariates.

Potential Covariates

A list of variables identified from the literature as potentially affecting interobserver agreement is provided in Table 2. Ooms [12] and Berg [7] both suggest that training radiologists in BI-RADS use prior to breast lesion assessment might improve interobserver agreement. Ciatto et al. [11] further support this idea, especially when introducing the BI-RADS lexicon internationally or in new settings. Baker [13] and Berg [9] suggest that an overlap in professional activities among readers may be responsible for higher values of kappa. Thus, we considered whether or not training in BI-RADS was provided and whether or not the observers worked together at the same institution or any point in time.

With regards to the selection of cases, five studies [5,6,8,11,13] reported a higher number of abnormal cases or the lack of negative cases as factors affecting kappa. Although a set of mammograms including only abnormal cases may not be representative of mammograms typically seen in a clinical setting it may still be a potentially informative study factor. A small number of cases has been cited by Kerlikowske [14] as a factor affecting interobserver variability. Based on these concerns, we included the selection of cases and the number of cases as potential covariates.

In three studies [10,11,12], the view type of a mammogram was noted as a possible factor affecting kappa. These variables included the use of a two-view mammogram as opposed to a single-view mammogram and the use of full-field digital mammograms. Two-view mammograms usually consist of two orthogonal views, which help to eliminate the problem of superimposition. According to Ciatto [11], the improvement in visual contrast and clarity provided by digital mammograms may improve the level of interobserver agreement.

Only one article noted the country in which the study was conducted as a source of variability in kappa [11]. Since the BI-RADS lexicon originated in the United States, U.S. users of the lexicon may be more familiar with its use. Thus we included examined country as a potential predictor. Due to our small sample size of ten papers, we dichotomized country of study as within the U.S. versus outside the U.S.

We included whether or not a study used the fourth edition of the BI-RADS lexicon. Because the fourth edition includes a guidance chapter with examples of lesions, its use may improve agreement since radiologists have a point of reference for description.

We included the BI-RADS descriptors for masses, calcifications, and final assessment categories as potential covariates because differences in kappa values may exist within these groups [5,6,7]. This was coded as a categorical variable, using dummy variables for masses and calcification and with final assessment as the reference category.

Table 2 provides a list of covariates considered and the coding scheme used for analysis. The data were abstracted from each of the ten studies using the potential covariates and coding scheme defined in Table 2. Except for number of cases, all potential predictors were coded as binary.

Model Selection

Table 3 provides the results of the bivariate analyses. Four of the nine possible predictors had p-values < 0.25. These included: the BI-RADS categories, masses (p-value=0.036), calcifications (p-value=0.164), the lack of negative cases (p-value=0.055), training in BI-RADS use (p-value=0.039) and the use of two-view mammograms (p-value=0.147). These covariates were considered for the multivariate analyses.

Table 4 provides the results of the multivariate analyses. In multivariate analysis, we compared models that included all possible combinations of the four selected predictors, which yielded $2^4 = 16$ possible models. The AIC and BIC criteria suggested that the best model was the model including masses/calcifications/final assessment category, training and two-view mammogram. From this model, for studies measuring kappa on the masses category compared to the final assessment category, the mean predicted kappa was 0.09 higher controlling for other factors. Similarly, for studies measuring kappa on the calcifications category compared to the final assessment category, the mean predicted kappa was 0.04 higher controlling for other factors. Training in BI-RADS use and using two-view mammograms displayed a similar positive effect. Studies in which observers are trained to use the BI-RADS lexicon had a predicted mean value of kappa 0.11 higher compared to studies which do not train their observers, holding all other factors constant. Finally, studies which utilized two-view mammograms had a predicted mean value of kappa 0.17 higher compared to studies which use single-view mammograms, holding all other factors constant.

The highest predicted mean value of kappa ($\kappa = 0.58$) for the best model occurred for studies measuring agreement on masses category descriptors, in which observers are trained in BI-RADS use and in which two-view mammograms are used. The smallest predicted mean value of kappa ($\kappa = 0.21$) occurred for studies measuring agreement on final assessment, in which observers are not trained in BI-RADS use, and in which single view mammograms are used.

DISCUSSION

Mammography is a valuable tool in the early detection of breast cancer, but radiologists can present considerably different interpretations and recommendations for follow-up. Efforts to reduce variability in interpretations may help to increase the efficacy of mammographic screenings using BI-RADS. This paper describes the first study to identify potential covariates of kappa as measure of agreement on BI-RADS terminology use /or assignment of final categories/recommendations for follow-up. While small, this study can be regarded as a pilot study that can serve to inform future investigations.

In our analysis of ten studies, nine possible predictors of kappa were identified. Multivariate analyses showed a positive association between kappa and training, use of two-view mammograms, and focusing on describing mass lesions. This may be due to the common experience of the readers, and increased visual clarity of the lesion. Although seemingly obvious observations, Ciatto et al. [11] reported that even well practiced and proficient radiologists may have limited consistency in reporting. Thus, interpretation of mammograms cannot solely rely on years of past experience.

Our finding of lower values of kappa for calcifications and final assessments as compared to masses suggests that these are areas for potential improvement. This result supports the finding of Berg et al. (2002) where cases with greatest variation compared to masses included calcifications [7]. Implementation of measures to increase interrater agreement in these areas, for example, more targeted training, may be important to consider. In supplemental analyses, we attempted to determine whether training had a differential effect on kappa for masses, calcifications and final assessments. However, our sample size was not sufficient to support these analyses. This may be a fruitful area for future research.

We found that studies that lacked negative cases had a trend towards lower kappa values in bivariate analyses. This agrees with a finding by Taplin et al., in which BI-RADS assessment of negative cases and benign lesions was more consistent than assessment of abnormal cases [31]. Since a set of mammograms that includes only abnormal cases is not typical of a clinical setting, this finding does not have direct relevance to predicting kappa in clinical settings.

However, we note that this factor was absent from the four best multivariate models. Another potential concern is that without negative cases, we exclude the possibility of a missed diagnosis, where the reader fails to detect an abnormality. Thus variability in lesion detection cannot be assessed. However, this has little impact on our findings since we are interested in determining reader variability in lesion description rather than lesion detection.

We examined digital versus film-screen mammography as a predictor of interobserver agreement. Since receiving FDA approval in 2000, digital mammography has gained attention as a potentially better screening tool than film-screen mammography. Developed to address the limitations of film-screen mammography, digital mammography provides easier acquisition, storage and retrieval of images. In addition, because images are digital in form, they are easily manipulated to improve clarity [32,33]. Despite these advantages, the diagnostic accuracy of digital versus film-screening mammography has not been completely resolved. For example, Skaane et al. reported in the Oslo II follow-up study that digital mammography demonstrated a significantly higher cancer detection rate [34]. On the other hand, the results of the Digital Mammographic Screening Trial showed no advantage of one technique over the other for the population as a whole but concluded that digital mammography was significantly better in detecting breast cancer in young women, pre-menopausal and peri-menopausal women and women with dense breasts [35]. Adding to this body of research, our results did not show a difference in interobserver agreement for digital versus film-screen mammograms. However, our results are aggregated over the study populations. In future research, it may be worthwhile to compare kappa values for different subgroups, a topic that we were unable to pursue in this research since our source studies did not provide the necessary information.

According to Rastogi [36], global cancer mortality is expected to increase by 104% by 2020, but it is also estimated that one-third of all cancers are preventable and potentially curable provided that detection is made early in the course of the cancer [37]. According to Kanavos [38], much of the disparity in cancer mortality in developing countries is due to the lack of early detection and prevention. Given this concern, it is encouraging that we did not find a significant difference in interrater agreement between studies conducted within and outside the U.S. This suggests that the BI-RADS system may be used outside the U.S. with similar interrater reliability. However, our study included studies from only a limited number of non-U.S. locations, which limits our ability to generalize on this point.

Our study has several additional limitations. We may not have fully identified all possible predictors from the literature. In addition, we were not able to abstract all potential predictors from all papers. For example, information regarding the length of time of reader interpretations could not be consistently abstracted because it was not always reported.

Another limitation of our study is sample size. Since we only had a sample of ten studies, the variation between studies may not be estimated well; however, despite our small sample size, the predictive value of our model is not diminished according to Gelman [39], and we were able to identify significant predictors.

Since kappa is bounded on the interval $[-1,1]$, we may discover in future studies that predicted values of kappa based on fitted models lie outside of the range so an appropriate transformation may be necessary to ensure interpretability [40]. Because our kappa values appeared normal in distribution and our final model yielded interpretable parameter estimates, we made the decision not to apply a transformation to the data.

In conclusion, training, using two-view mammograms, and focusing on the description of mass lesions may be useful in reducing interobserver variability in BI-RADS use. These findings are important for implementing policies to ensure the reliability of mammogram interpretation

using the BI-RADS lexicon prior to its introduction in a new setting or country, and to increase reliability in current settings.

Acknowledgments

Crespi was supported by NIH CA 16042.

References

1. American Cancer Society. Cancer facts and figures 2009. [Accessed 1 May 2009]. <http://www.cancer.org/downloads/STT/500809web.pdf>
2. Wiratkapun C, Lertsithichai P, Wibulpholprasert B. Positive predictive value of breast cancer in the lesions categorized as BI-RADS category 5. *J Med Assoc Thai* 2006 Aug;89(8):1253–9. [PubMed: 17048437]
3. Masroor I. Prediction of benignity or malignancy of a lesion using BI-RADS. *J Coll Physicians Surg Pak* 2005 Nov;15(11):686–8. [PubMed: 16300702]
4. Resende LM, Matias MA, Oliveira GM, Salles MA, Melo FH, Gobbi H. Evaluation of breast microcalcifications according to Breast Imaging Reporting and Data System (BI-RADS) and Le Gal's classifications. *Rev Bras Ginecol Obstet* 2008 Feb;30(2):75–9. [PubMed: 19142479]
5. Lazarus E, Mainiero MB, Schepps B, Koelliker SL, Livingston LS. BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. *Radiology* 2006 May;239(2):385–91. Epub 2006 Mar 28. [PubMed: 16569780]
6. Coşar ZS, Cetin M, Tepe TK, Cetin R, Zarali AC. Concordance of mammographic classifications of microcalcifications in breast cancer diagnosis: Utility of the Breast Imaging Reporting and Data System (fourth edition). *Clin Imaging* 2005 Nov–Dec;29(6):389–95. [PubMed: 16274891]
7. Berg WA, D'Orsi CJ, Jackson VP, Bassett LW, Beam CA, Lewis RS, Crewson PE. Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography? *Radiology* 2002 Sep;224(3):871–80. [PubMed: 12202727]
8. Gülsün M, Demirkazık FB, Ariyürek M. Evaluation of breast microcalcifications according to Breast Imaging Reporting and Data System criteria and Le Gal's classification. *Eur J Radiol* 2003 Sep;47(3):227–31. [PubMed: 12927667]
9. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment. *AJR Am J Roentgenol* 2000 Jun;174(6):1769–77. [PubMed: 10845521]
10. Ciatto S, Houssami N, Apruzzese A, Bassetti E, Brancato B, Carozzi F, Catarzi S, Lamberini MP, Marcelli G, Pellizzoni R, Pesce B, Risso G, Russo F, Scorsolini A. Categorizing breast mammographic density: intra- and interobserver reproducibility of BI-RADS density categories. *Breast* 2005 Aug;14(4):269–75. [PubMed: 16085233]
11. Ciatto S, Houssami N, Apruzzese A, Bassetti E, Brancato B, Carozzi F, Catarzi S, Lamberini MP, Marcelli G, Pellizzoni R, Pesce B, Risso G, Russo F, Scorsolini A. Reader variability in reporting breast imaging according to BI-RADS assessment categories (the Florence experience). *Breast* 2006 Feb;15(1):44–51. Epub 2005 Aug 1. [PubMed: 16076556]
12. Ooms EA, Zonderland HM, Eijkemans MJ, Kriege M, Mahdavian Delavary B, Burger CW, Ansink AC. Mammography: interobserver variability in breast density assessment. *Breast* 2007 Dec;16(6):568–76. [PubMed: 18035541]
13. Baker JA, Kornguth PJ, Floyd CE Jr. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. *AJR Am J Roentgenol* 1996 Apr;166(4):773–8. [PubMed: 8610547]
14. Kerlikowske K, Grady D, Barclay J, Frankel SD, Ominsky SH, Sickles EA, Ernster V. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. *J Natl Cancer Inst* 1998 Dec 2;90(23):1801–9. [PubMed: 9839520]

15. Fleiss, JL. Statistical methods for rates and proportions. New York: John Wiley & Sons; 1973. p. 598-626.
16. Wenkel E, Heckmann M, Heinrich M, Schwab SA, Uder M, Schulz-Wendtland R, Bautz WA, Janka R. Automated breast ultrasound: lesion detection and BI-RADS classification--a pilot study. *Rofo* 2008 Sep;180(9):804-8. Epub 2008 Aug 14. [PubMed: 18704878]
17. Skaane P, Diekmann F, Balleyguier C, Diekmann S, Piguat JC, Young K, Abdelnoor M, Niklason L. Observer variability in screen-film mammography versus full-field digital mammography with soft-copy reading. *Eur Radiol* 2008 Jun;18(6):1134-43. Epub 2008 Feb 27. [PubMed: 18301902]
18. Castella C, Kinkel K, Eckstein MP, Sottas PE, Verdun FR, Bochud FO. Semiautomatic mammographic parenchymal patterns classification using multiple statistical features. *Acad Radiol* 2007 Dec;14(12):1486-99. [PubMed: 18035278]
19. Caramella T, Chapellier C, Ettore F, Raoust I, Chamorey E, Balu-Maestro C. Value of MRI in the surgical planning of invasive lobular breast carcinoma: a prospective and a retrospective study of 57 cases: comparison with physical examination, conventional imaging, and histology. *Clin Imaging* 2007 May-Jun;31(3):155-61. [PubMed: 17449375]
20. Thomas A, Kümmel S, Fritzsche F, Warm M, Ebert B, Hamm B, Fischer T. Real-time sonoelastography performed in addition to B-mode ultrasound and mammography: improved differentiation of breast lesions? *Acad Radiol* 2006 Dec;13(12):1496-504. [PubMed: 17138118]
21. Thomas A, Fischer T, Frey H, Ohlinger R, Grunwald S, Blohmer JU, Winzer KJ, Weber S, Kristiansen G, Ebert B, Kümmel S. Real-time elastography--an advanced method of ultrasound: First results in 108 patients with breast lesions. *Ultrasound Obstet Gynecol* 2006 Sep;28(3):335-40. [PubMed: 16909438]
22. Gupta S, Chyn PF, Markey MK. Breast cancer CADx based on BI-RADS descriptors from two mammographic views. *Med Phys* 2006 Jun;33(6):1810-7. [PubMed: 16872088]
23. Martin KE, Helvie MA, Zhou C, Roubidoux MA, Bailey JE, Paramagul C, Blane CE, Klein KA, Sonnad SS, Chan HP. Mammographic density measured with quantitative computer-aided method: comparison with radiologists' estimates and BI-RADS categories. *Radiology* 2006 Sep;240(3):656-65. Epub 2006 Jul 20. [PubMed: 16857974]
24. Teifke A, Vomweg TW, Hlawatsch A, Nasresfahani A, Kern A, Victor A, Schmidt M, Bittinger F, Düber C. Second reading of breast imaging at the hospital department of radiology: reasonable or waste of money? *Rofo* 2006 Mar;178(3):330-6. [PubMed: 16508842]
25. Lorenzen J, Wedel AK, Lisboa BW, Löning T, Adam G. Diagnostic mammography and sonography: concordance of the breast imaging reporting assessments and final clinical outcome. *Rofo* 2005 Nov;177(11):1545-51. [PubMed: 16302136]
26. Yamada T, Saito M, Ishibashi T, Tsuboi M, Matsushashi T, Sato A, Saito H, Takahashi S, Onuki K, Ouchi N. Comparison of screen-film and full-field digital mammography in Japanese population-based screening. *Radiat Med* 2004 Nov-Dec;22(6):408-12. [PubMed: 15648457]
27. Pijnappel RM, Peeters PH, Hendriks JH, Mali WP. Reproducibility of mammographic classifications for non-palpable suspect lesions with microcalcifications. *Br J Radiol* 2004 Apr;77(916):312-4. [PubMed: 15107321]
28. Perisinakis K, Damilakis J, Kontogiannis E, Gourtsoyiannis N. Film-screen magnification versus electronic magnification and enhancement of digitized contact mammograms in the assessment of subtle microcalcifications. *Invest Radiol* 2001 Dec;36(12):726-33. [PubMed: 11753144]
29. Venta LA, Hendrick RE, Adler YT, DeLeon P, Mengoni PM, Scharl AM, Comstock CE, Hansen L, Kay N, Coveler A, Cutter G. Rates and causes of disagreement in interpretation of full-field digital mammography and film-screen mammography in a diagnostic setting. *AJR Am J Roentgenol* 2001 May;176(5):1241-8. [PubMed: 11312188]
30. Baker JA, Kornguth PJ, Lo JY, Floyd CE Jr. Artificial neural network: improving the quality of breast biopsy recommendations. *Radiology* 1996 Jan;198(1):131-5. [PubMed: 8539365]
31. Taplin SH, Ichikawa LE, Kerlikowske K, et al. Concordance of breast imaging reporting and data system assessments and management recommendations in screening mammography. *Radiology* 2002 Feb;222(2):529-35. [PubMed: 11818624]
32. Pisano ED, Yaffe MJ. Digital mammography. *Radiology* 2005 Feb;234(2):353-62. [PubMed: 15670993]

33. Hambly NM, McNicholas MM, Phelan N, Hargaden GC, O'Doherty A, Flanagan FL. Comparison of digital mammography and screen-film mammography in breast cancer screening: a review in the Irish breast screening program. *AJR Am J Roentgenol* 2009 Oct;193(4):1010–8. [PubMed: 19770323]
34. 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;244:708–717. [PubMed: 17709826]
35. Pisano ED, Gatsonis C, Hendrick E, et al. DMIST Investigators Group. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med* 2005;353:1773–1783. Erratum in *N Engl J Med* 2006; 355:1840. [PubMed: 16169887]
36. Rastogi T, Hildesheim A, Sinha R. Opportunities for cancer epidemiology in developing countries. *Nat Rev Cancer* 2004;4:909–917. [PubMed: 15516963]
37. Alwan A. Non-communicable diseases: a major challenge to public health in the region. *East Mediterr Health J* 1997;3:6–16.
38. Kanavos P. The rising burden of cancer in the developing world. *Ann Oncol* 2006 Jun;17(Suppl 8):viii15–viii23. [PubMed: 16801335]
39. Gelman, A.; Hill, J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press; 2006.
40. Lesaffre E, Rizopoulos D, Tsonaka R. The logistic transform for bounded outcome scores. *Biostatistics* 2007 Jan;8(1):72–85. Epub 2006 Apr 5. [PubMed: 16597671]

Table 1

Description of ten studies reporting kappa values for mammogram patients.

Study	Publication year	Country	No. readers	No. mammograms	Lowest reported kappa	Highest reported kappa	Mean kappa value	Standard deviation
Baker [13]	1996	United States	5	60	0.50	0.77	0.66	0.10
Kerlikowske [14]	1998	United States	2	2616	0.23	0.76	0.46	0.17
Berg [9]	2000	United States	5	106	-0.02	0.58	0.28	0.21
Berg [7]	2002	United States	27	54	0.23	0.54	0.40	0.09
Gulsun [8]	2002	Turkey	2	82	0.16	0.45	0.31	0.10
Ciatto [10]	2005	Italy	12	100	0.02	0.77	0.54	0.26
Cosar [6]	2005	Turkey	3	83	0.22	0.74	0.50	0.14
Ciatto [11]	2006	Italy	12	50	0.07	0.44	0.21	0.16
Lazarus [5]	2006	United States	5	94	0.14	0.56	0.34	0.15
Ooms [12]	2007	The Netherlands	4	57	0.65	0.84	0.74	0.06

Table 2

Potential predictors of kappa.

Variable	Coding	References
Training in BI-RADS use	1 = observers were instructed or trained in BI-RADS use prior to participation in study 0 = no training was provided	[7], [11], [12]
Overlap in observer professional activities	1 = any overlap in professional activities at any point in time 0 = no overlap	[9], [13]
No negative cases	1 = selection of cases did not include negative mammogram cases 0 = negative mammogram cases were included	[5], [6], [8], [11], [13]
Mammogram View Type	1 = two view mammogram 0 = single view mammogram	[10], [11], [12]
Mammogram Film Type	1 = full field digital mammogram 0 = traditional film-screen mammogram	[11]
Number of cases	Continuous variable	[14]
Use of BI-RADS 4th ed.	1 = study used BI-RADS 4th edition 0 = study did not use BI-RADS 4th edition	[5], [6], [11]
Country	1 = study was conducted in the United States 0 = study was conducted outside of the United States	[11]
BI-RADS descriptor category		
Masses	1 = kappa belongs to masses descriptor category 0 = kappa does not belong to masses descriptor category	[5], [6], [7]
Calcifications	1 = kappa belongs to calcifications descriptor category 0 = kappa does not belong to calcifications descriptor category	
Final Assessments (Reference category)		

Table 3

Bivariate analyses of predictors of kappa.

Variable	Estimate	Standard Error	p-value
Masses (vs. final assessments)	.12	.06	.036
Calcifications (vs. final assessments)	.07	.05	.164
Training in BI-RADS use (vs. no BI-RADS training)	.15	.07	.039
Overlap in professional activities (vs. no overlap)	-.05	.11	.665
No negative cases (vs. negative cases included)	-.21	.09	.055
Two view mammogram (vs. single view)	.26	.16	.147
Digital mammogram (vs. film-screen)	.08	.14	.570
Number of cases (one-case increase)	.000006	.00007	.940
Use of BI-RADS 4 th ed (vs. other edition)	-.01	.10	.902
Country (U.S. vs. non-U.S.)	-.04	.11	.736

Analyses conducted using multilevel regression with random intercepts for study.

Table 4

Multilevel multiple regression models for kappa

Model	Regression coefficient (SE)						AIC	BIC
	Intercept	Masses	Calcifications	Training	No negative cases	Two-view mammogram		
1	.21(.12)	.09(.06)	.04(.05)	.11(.07)		.17(.13)	-33.4	-32.8
2	.35(.05)	.10(.06)	.06(.05)	.12(.07)			-34.0	-33.3
3	.21(.13)	.11(.06)	.06(.05)			.20(.14)	-34.8	-34.2
4	.37(.05)	.12(.06)	.07(.05)				-35.0	-34.4
5	.37(.14)	.10(.06)	.06(.05)		-.16(.08)	.14(.12)	-35.3	-34.7
6	.50(.07)	.12(.10)	.08(.05)		-.18(.08)		-36.4	-35.8
7	.45(.05)						-39.3	-38.7
8	.39(.05)			.15(.07)			-39.3	-38.7
9	.21(.13)			.14(.07)		.21(.13)	-39.5	-38.9
10	.21(.15)					.25(.16)	-40.0	-39.4
11	.39(.16)				-.18(.09)	.20(.14)	-40.6	-40.0
12	.58(.08)				-.20(.09)		-40.8	-40.2
13	.39(.08)	.06(.05)	.05(.05)	.18(.05)	-.18(.04)	.09(.07)	-40.7	-40.4
14	.48(.04)	.07(.05)	.06(.05)	.18(.05)	-.19(.04)		-42.5	-42.2
15	.41(.07)			.20(.04)	-.20(.04)	.00(.07)	-48.0	-47.7
16	.52(.03)			.22(.04)	-.21(.04)		-48.5	-48.2

Analyses conducted using multilevel regression with random intercepts for study.