Performance differences across the Atlantic when UK and USA radiologists read the same set of test screening cases

Yan Chen*[a], Alastair G. Gale[a], Michael Evanoff[b]
[a]Applied Vision Research Centre, Loughborough University, Loughborough, UK;
[b] American Board of Radiology, Tucson, USA

ABSTRACT

Two groups of experienced radiologists from the UK and the USA read the same set of 40 recent FFDM screening cases to examine the effects of mammography experience, volume of cases read per year, screening practice and monitor resolution on performance,. Sixteen American radiologists reported these cases using twin DICOM calibrated monitors which were half the resolution of the clinical mammographic workstations used by 16 UK radiologists. In terms of effects of volume of cases read per year, then when the group of American radiologists were split into high and low volume readers (using 5,000 cases p.a. as a criterion) no difference in any performance measure was found. This may be partly explained by the fact that they were all were very experienced which may have counteracted any case volume effect here. Comparing the two groups of radiologists from both countries, then the UK group performed better in terms of the number of cancers detected although the American group recalled more cases, despite having poorer monitors. This reflects differences in clinical screening practice between the countries, however differences simply due to the reporting monitors used cannot be ruled out. Data from the study were also compared to that from all UK screeners who had read these cases as either soft copy or as mammographic film.

Keywords: breast screening, FFDM, performance

1. INTRODUCTION

Breast cancer is a disease which affects one in eight women in the UK and the USA at some point in their lives[1,2]. The best way to minimize the effects of this cancer is to detect it as early as possible and consequently many countries have instituted nationwide screening programmes. These operate differently in each country. The UK has a centrally organized national breast cancer screening programme (NHSBSP) under the auspices of the National Health Service which has run successfully for some 23 years. Currently the programme screens over 2 million women every three years aged between 45 and 74 years[3]. This programme was originally established using mammographic film and has for many years imaged each woman using two mammographic views (the Medio-Lateral Oblique and Cranio-Caudal) but is currently very rapidly changing to employing similar two-view Full Field Digital Mammography (FFDM). All breast screening radiologists and advanced practitioners (specially trained technologists who also read and interpret screening mammographic cases) have to read a minimum of 5,000 cases a year in order to take part in the national breast screening programme[4]. Despite the headline figure of breast cancer affecting 1 in 8 women the incidence in the screened population is actually very low per 1,000 women screened; consequently this annual high number of cases read ensures that every individual experiences a high number and range of normal as well as abnormal cases. The number of women recalled after screening for subsequent examinations in the UK is kept low, circa 4.2% nationally[3].

In the USA the organization of breast screening is somewhat different. American breast screening radiologists typically read a much lower annual volume of cases; the American College of Radiology specifies as part of the mammography accreditation program that an interpreting physician should interpret 960 mammographic examinations over a two year

*y.chen@lboro.ac.uk

Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment,
edited by Craig K. Abbey, Claudia R. Mello-Thoms, Proc. of SPIE Vol. 8318, 831811
© 2012 SPIE · CCC code: 1605-7422/12/$18 · doi: 10.1117/12.913609

period as part of continuing experience[5]. Additionally, in the USA more women are recalled as compared to in the UK, with one study, reporting on real life data, figures of 13.3% as compared to 7.2%[6] in the UK in this study[6]. This reflects differences between the two countries in their approach to screening practice.

If the volume of cases read relates to the ability to more readily determine whether a case should be recalled or not then performance differences between the two countries would be expected in any experimental comparison between radiologists from these countries. As part of the UK screening programme all screeners undertake the annual PERFORMS scheme which has been well described previously[7,8]. This is an educational exercise where screeners examine sets of recent difficult cases containing a range of mammographic features, receiving immediate feedback as well as more detailed feedback some time later when their performance data are compared anonymously to that of their peers.

This present study investigated what happens when experienced breast screening radiologists from both countries examined the same FFDM case set from the PERFORMS scheme, albeit using different resolution displays, as a first step in a more detailed study using comparable displays. Breast screening experience and the volume of screening cases read per year are known key predictors of an individual's screening performance. Additionally, high resolution mammography workstations are a clinical requirement for soft copy reporting of FFDM images. Notwithstanding this, our previous work[9] has demonstrated that good performance in identifying abnormalities on screening mammograms can be achieved using a single office monitor with suitable image manipulation software as compared to a clinical mammography workstation when the same UK screeners read the same cases on different occasions. In the present study the effects of differences in experience, case volume per annum and monitor resolution were investigated. Comparative data involving UK and USA radiologists interpreting the same test case set of digital screening mammograms which examines screening practices, screening experience, case volume and reporting monitor resolutions have not previously been investigated.

2. METHOD

2.1 USA group

The opportunity arose to demonstrate the PERFORMS scheme to a number of leading American radiologists involved in breast screening who were examiners at the American Board of Radiology board certification examinations in Louisville (2011). During these board examination the examiners had breaks in their schedules where they were then timetabled to take part and experience aspects of the PERFORMS scheme. In the UK radiologists undertaking PERFORMS will read two sets of 60 difficult cases per annum and on each occasion they can spend a variable amount of time in doing this. In Louisville, as the available time to participate was limited for each examiner, then we deliberately restricted the number of cases that they examined to 40. Efforts to acquire clinical mammographic workstations for these participants in this investigation, unfortunately proved to be unsuccessful and consequently a test room was set up with four workstations, each comprising a PC running dual 20" DICOM calibrated monitors. The test room lighting was reduced and measured to reflect typical screening room light levels for digital mammography reporting. Each radiologist interpreted 40 difficult FFDM cases from a recent set of PERFORMS test cases. Due to the limits of the monitor resolution these cases were carefully selected so as to exclude any small micro-calcifications which would be expected not to have visualized well on these monitors.

For each of the four workstations the case set of 40 images was loaded and a dedicated PERFORMS tablet computer set up. This device ran the PERFORMS reporting software which the user first logs into. It then presents a matrix of cases to select and examine, starting with practice cases. The user selects a case on the tablet and the corresponding FFDM images on the workstation. They then examine the workstation images and make responses on the tablet using a graphical user interface. For each case the user decides whether various mammographic features are present or not and locates these on small images of the mammographic cases on the tablet computer. They also decide how they would classify each breast; ranging from normal, benign to malignant and whether or not they would recall each breast. In doing this a rating scale is used which is broadly similar to BIRADS[10] but which follows the UK classification approach[11]. Because of this difference in scoring method each participant was first led through the practice cases and how to utilize the reporting process by one of the experimenters in order to familiarize the participants with the process.

Once they were comfortable with the use of the DICOM image viewer employed here as well as the reporting software then the experimental study began and they read the reported the 40 cases in their own time and at their own rate. After participating, each individual completed a short questionnaire concerning their usual screening practice including such questions as how many cases a year they read and for how many years they had been involved in mammography.

2.2 UK group

For comparison purposes the anonymous data were used of 16 experienced UK breast radiologists out of over several hundred who had previously read the same cases as part of a recent round of the PERFORMS scheme using their usual clinical mammographic workstations. These individuals were randomly selected from all those who had recently read the latest PERFORMS test set of 60 cases as soft copy images. For each radiologist their anonymous data were extracted for the 40 cases which were used here in this study.

Secondly, data from the study were compared to the anonymous data of all UK participants in the scheme for these 40 cases who had read the cases as soft copy. This included data for both breast screening radiologists and advanced practitioners who are specially trained technologists who read screening cases in the UK.

3. RESULTS

3.1 USA group

It is well documented that the volume of cases read typically affects screening performance. Out of the 16 American radiologists a number reported that they annually read over 5,000 cases and the rest read less than this. As this is the same criterion volume of cases that UK radiologists must read in order to participate in breast screening in the UK then the American group were first split according to this reported number into two sub-groups of low (<5,000) and high annual volume (=>5,000) of screening cases read in order to examine whether reported volume was related here to their performance on this test set. All of these radiologists also reported that they had been involved with mammography for over 15 years. In terms of the study experience they all reported finding the PERFORMS case set a very interesting educational experience.

In the PERFORMS scheme there are several performance metrics which are used. Here, the three key performance measures of malignancies correctly identified, coupled with the decisions of whether or not a case should be recalled (correct recall) or returned to routine screening (correct return to screen) were examined. These data were calculated and compared between these two sub-groups. No significant differences were found (t=0.23, *p*=n.s) and consequently, the performance data of these two sub-groups were combined in the subsequent analyses.

3.2 Comparison of USA and UK performances

The data of the 16 American radiologists were then compared to that of the selected 16 UK radiologists on the screening metrics. There was no significant difference (figure 1) between these two groups in correct recall (CR) decisions (UK: *M* = 90.7%. *SE* = 1.05, *p* = n.s.; USA: *M* = 92.3%, *SE* = 1.93, *p* = n.s.). Also, there was no significant difference between these two groups in Negative Predictive Value (NPV) decisions (UK: *M* = 95.7%. *SE* = 0.45; USA: *M* = 96.9%, SE = 0.74 *p* = n.s). Furthermore, there was no significant difference (figure 2) between these two groups in A_z scores (UK: M = 0.97 SE = 0.01; USA: M = 0.95, SE =0.01 *p* = n.s).

However, there were significant differences (figure 1) in correct return to screening (CS) decisions (UK: M = 89.8%, SE = 1.70; USA: M = 81.1%, SE = 1.94; $p<0.05$) and in the percentages of malignancies (CD) detected (UK: M = 99.7%, SE = 0.28; USA: 92.3%, SE = 1.94; $p<0.05$) and in the Positive Predictive Value (PPV) percentages (UK: M = 81.3%, SE = 2.56; USA: M = 65.0%, SE=2.43; $p<0.05$).

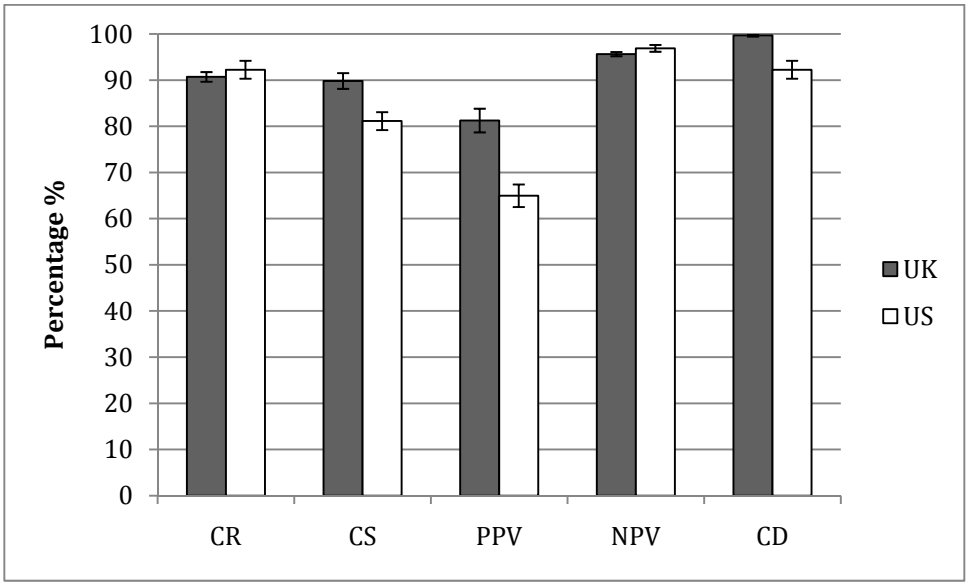

Figure 1. Performance values comparing 16 USA radiologists and 16 UK radiologists

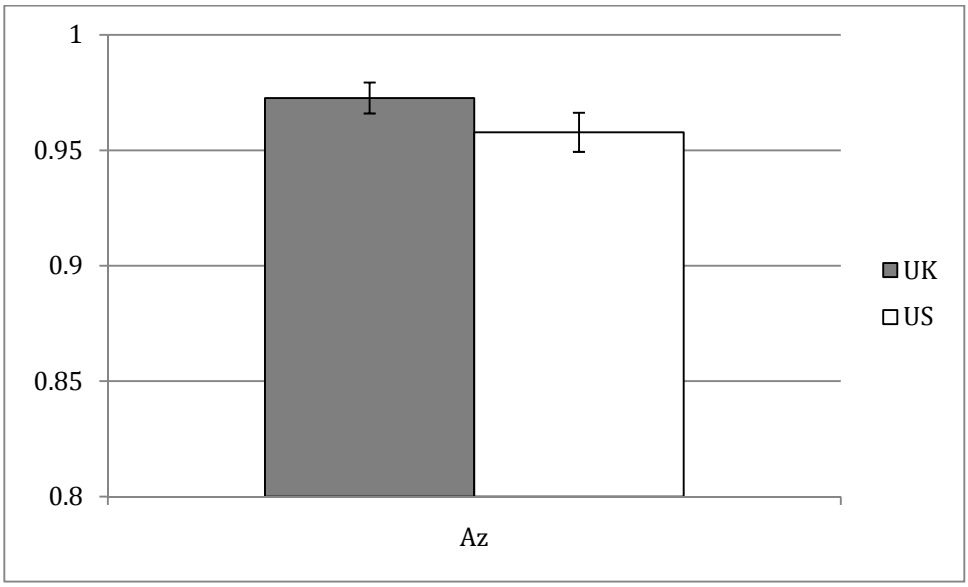

Figure 2. A_z values comparing 16 USA radiologists and 16 UK radiologists

3.3 Further UK – USA comparisons

The data of the 16 American radiologists were then compared on the screening metrics to that of all UK radiologists who had read these cases as soft copy images. There was no significant difference (figure 3) between these two groups in correct recall (CR) decisions (UK: M = 92.2%. SE = 0.47; USA: M = 92.3%, SE = 1.93, p = n.s.). Also, there was no significant difference between these two groups in NPV decisions (UK: M = 97.2%, SE = 0.16, USA: M = 96.9%, SE = 0.74, p = n.s). No significant difference was also between these two groups in cancer detection (CD) percentages (UK: M = 92.2%, SE = 0.004; USA: M = 92.3%, SE = 0.19, p = n.s).

However, there were significant differences in correct return to screening (CS) decisions (UK: $M = 88.5\%$, $SE = 0.47$; USA: $M = 81.1\%$, $SE = 1.94$; $p<.05$) and in PPV percentages (UK: $M = 76.3\%$, $SE = 0.67$; USA: $M = 64.9\%$, $SE = 2.43$, $p<.05$) and in the A_z scores (figure 4) (UK: $M = 0.97$, $SE = 0.001$; USA: $M = 0.95$, $SE = 0.008$, $p<.05$).

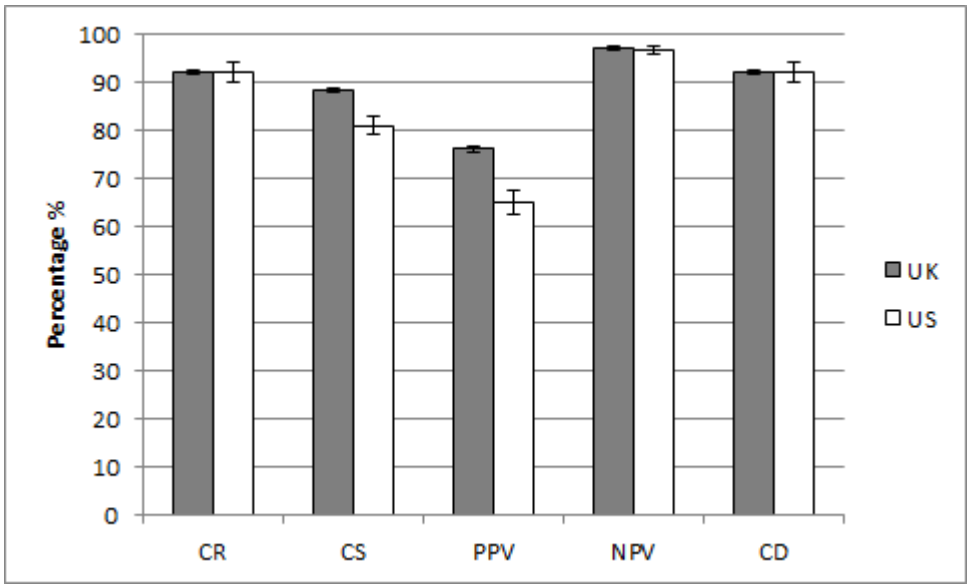

Figure 3. Performance values comparing 16 USA radiologists and UK screeners

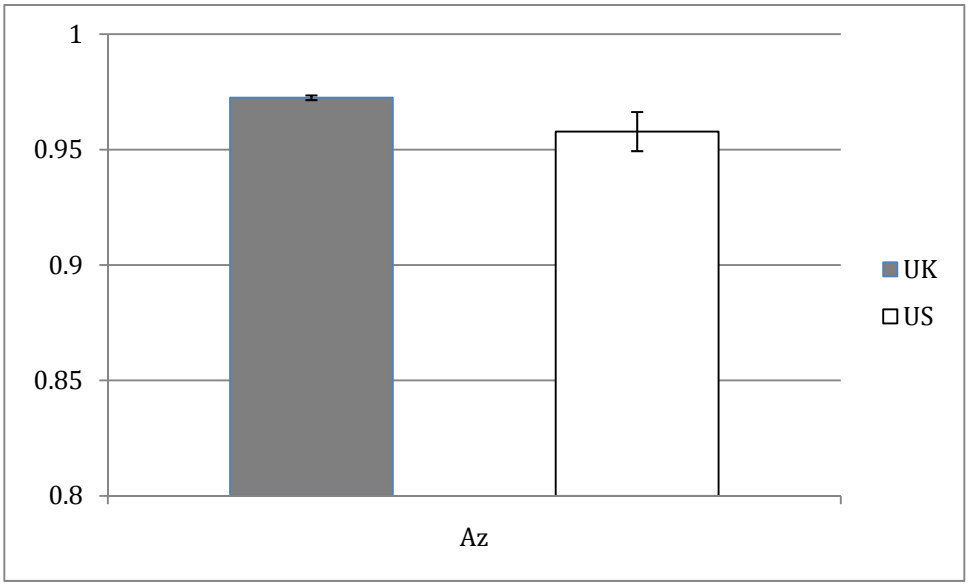

Figure 4. A_z values comparing 16 USA radiologists and UK screeners

4. DISCUSSION

This study is the first step in investigating how experienced breast screening radiologists from the UK and USA fare when examining the same FFDM cases and using the same reporting software. Previously we have carried out a somewhat similar exercise in comparing how two groups from these countries have read a PERFORMS scheme test set of mammographic films using multi-viewers[12]. In the present study the American group unfortunately only had access to monitors which had approximately half the resolution of a mammographic workstation and therefore would have had difficulty in visualizing small micro-calcifications. To counter this to some extent, the cases which were used were carefully selected so as not to have significant micro-calcifications. We have previously shown that acceptable performance can be obtained when the same group of screeners examined a case set using either their clinical mammographic workstations or a single office monitor as long as suitable interaction software is utilised[9]. The reporting software was new to the American group, however it is a transparent reporting system which has been well researched to be very user friendly. The software does not use the BIRADS categorization but a close approximation and in practice none of the American group had difficulty in using it.

Somewhat in contrast, comparative data were used here from 16 randomly selected experienced UK radiologists to form the UK group which had several advantages – they had utilized their routine clinical workstations to examine the cases on and were familiar with the reporting software which has been in use in the UK annually for circa seven years.

Examining the data from Louisville firstly split into high and low volume of cases read per annum yielded no significant differences. On the face of it this is somewhat surprising as usually better performance in breast screening relates to the volume of cases an individual reads. However, here the radiologists were all very experienced in mammography and thus even the low volume readers had had considerable years of expertise in examining mammograms. It is always difficult to tease out experience and volume of cases read as independent factors as inevitably the two are intrinsically related. We have previously reported that volume of cases read is important[12, 13] when examining performance on the scheme. Furthermore, when we earlier examined the data of 450 participants on the PERFORMS scheme and related this to their real life volume of cases read per annum and years of screening experience it was found that years of experience was much more important than volume of cases read[14]. Somewhat relatedly, Beam, Conant and Sickles[15] have reported in an American study that current reading volume was not significantly related to accuracy with expertise reflecting 'a complex multifactorial process'.

Comparing how these 16 radiologists fared to 16 experienced UK radiologists demonstrated that both groups somewhat similarly correctly recalled those cases which should be recalled, based on known case pathology and actual screening outcome. Significant differences were found in decisions of correctly returning cases to screening (i.e. judging that a case was normal or benign and not worthy of further investigation at that time), percentages of malignancies detected and in PPV. This is as predicted based on differences between the two countries in routine screening practices where the American group over-read cases as compared to the UK group. That the UK group detected significantly more malignancies probably reflects the differences here in the workstation monitors employed.

The American group, notwithstanding the poorer monitors used here, were then compared to the whole of the UK screening programme who had participated in reading these same cases because this would then encompass a wider spectrum in the UK of screening behaviours. Again, broadly similar results were found in that the American group over-read the case set as compared to all the UK participants.

5. CONCLUSIONS

The purpose of the study was primarily to examine what happens when two groups of experienced breast screeners from different routine clinical screening practices examine the same set of difficult cases. Unfortunately it was not possible here for both groups to utilize clinical mammographic workstations. Despite this, the use of lower resolution monitors by the American group was clearly offset by their experience in mammography (all > 15 years) such that even the very experienced but low reported volume readers performed well with there being no significant differences between the two American sub-groups, split by reported volume of cases read per annum.

Not surprisingly, the comparative selected UK group of 16 radiologists overall performed better in reporting these test cases as they were using high resolution mammographic workstations. However, the American group still recalled more, reflecting their real life screening practice. Examining how the American group did as compared to all UK screening radiologists and advanced practitioners who had read the same cases as soft copy demonstrated that the performance data for the American radiologists were broadly comparable to those of UK participants.

The second part of this study will take place in 2012 at Louisville with an American group of experienced radiologists using clinical mammographic workstations and it is hoped that this will give yet more insight into the factors underlying screening performance in these two countries. The PERFORMS scheme was specifically designed for individual screeners in the UK to enable them to gain insights into their performance in identifying early signs of cancer as well as how well they perform as compared to colleagues. This study demonstrates that screeners from outside the UK can use the scheme equally well and report finding it a useful educational aid.

ACKNOWLEDGEMENTS

This work is partly supported by the UK National Health Service Breast Screening Programme.

REFERENCES

[1] http://www.breastcancer.org/symptoms/understand_bc/statistics.jsp
[2] http://www.nhs.uk/news/2011/02February/Pages/breast-cancer-rates-rise-to-one-in-eight.aspx
[3] Patnick J (ed.) NHSBSP Annual Review 2011, NHS Cancer Screening Programmes 2011
[4] Royal College of Radiologists: Quality Assurance Guidelines for Radiologists, 1990
[5] http://www.acr.org/accreditation/mammography/overview/overview.aspx
[6] Smith-Bindman R, Ballard-Barbash R, Miglioretti DL, Patnick J, Kerlikowske K; Comparing the performance of mammography screening in the USA and the UK, J Med Screen 2005;12:50–54
[7] Gale A.G., "PERFORMS – a self assessment scheme for radiologists in breast screening," Seminars in Breast Disease: Improving and monitoring mammographic interpretative skills, 6(3), 148-152, (2003)
[8] Gale A.G., "Maintaining quality in the UK breast screening program", In D.J. Manning & C. Abbey (Eds.) Proc. SPIE Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment. 7627, 1-11 (2010).
[9] Chen Y., Turnbull A., James J., Gale A.G., Scott H."Breast Screening: visual search as an aid for digital mammographic interpretation training" . SPIE Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment. 7627, (2010).
[10] American College of Radiology, "ACR breast imaging reporting and data system atlas". Reston, VA: American College of Radiology (2003)
[11] Maxwell A.J., Ridley N.T.,Rubin G., Wallis M.G., Gilbert F.J., Michel M.J. "The Royal College of Radiologists Breast Group breast imaging classification". Clinical Radiology, Vol. 64, Issue 6, , Pages 624–627
[12] Esserman L., Cowley H., Eberle C., Kirkpatrick A., Chang S., Berbaum K., & Gale A.G.: Improving the Accuracy of Mammography: Volume and Outcome Relationships. Journal of the National Cancer Institute, 2002, Vol. 94, No. 5, 369-375, March 6
[13] Scott H.J., Gale A.G., & Wooding D.S. : Breast Screening Technologists: does real-life case volume affect performance? In: Image Perception, Observer Performance, and Technology Assessment, D.P. Chakraborty & M.P. Eckstein (eds.) Proceedings of SPIE Vol. 5372, 2004
[14] Scott H.J. & Gale A.G.: How much is enough: factors affecting the optimal interpretation of breast screening mammograms. In Image Perception, Observer Performance, and Technology Assessment. Y Jiang and B Sahiner (Eds.) Proceedings of SPIE 2007.
[15] Beam C.A., Conant E.F., Sickles E.A. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation J Natl Cancer Inst (2003) 95 (4): 282-290.