The Detection of Change in Mammographic Density

J. Stone, A. Gunasekara, L. J. Martin, M. Yaffe, S. Minkin, and N. F. Boyd


Abstract

Mammographic density is associated with risk of breast cancer, and factors that change density may also change risk. There has, however, been little research into how change in serial mammograms is best detected. The purpose of the work described here was to examine the effects of different reading conditions on the detection of change in mammographic features. Mammograms were selected from women who had participated in a randomized controlled trial of screening for breast cancer. We selected two age-matched groups of subjects, one who had undergone menopause after entry ($n = 202$) and another who had not ($n = 202$). Serial mammograms from these subjects were then measured four times using a computer-assisted method under different conditions: (a) films were randomized; (b) subjects were randomized (i.e., pairs of films from individuals were read one after the other), but the order of films was random and unknown to the reader; (c) subjects were randomized, and the order of films was sequential and known to the reader; and (d) subjects were randomized, and the order of films was random and unknown to the reader, but both films in each pair were read simultaneously on separate computer screens. The mean effect of the menopause on change in the mammographic measures of total, dense and nondense areas, percent density, and the associated variances were then compared. With one exception, all of the randomization and viewing methods confirmed a change in all mammographic measures at menopause and produced very similar overall results, suggesting that mammographic density is a robust measure. Compared with randomization of all films, the method in which subjects were randomized and paired films read one after the other in random and unknown order was associated with a slightly smaller mean difference and achieved a substantial reduction in variability, suggesting that it is the most sensitive method of randomization and viewing for the detection of change.

Introduction

The appearance of the female breast on mammography varies among women of the same age because of differences in tissue composition (1). Fat is radiologically lucent and appears dark on a mammogram, whereas connective and epithelial tissues are radiologically dense and appear light, an appearance that we refer to as mammographic density. Wolfe (2, 3) first described an association between a qualitative classification of dense mammographic patterns and an increased risk of breast cancer, and several other cohort studies have confirmed this association (4–11). Ten studies [6 case-control (12–17) and 4 cohorts (18–21)], containing 4747 cases of breast cancer, have assessed mammographic density quantitatively, and all found a risk of breast cancer of between 1.8- and 6-fold in the most extensive category of density compared with the least, and in 8, the gradient in risk was at least 4-fold. The risk of breast cancer associated with mammographic density is larger than almost all other risk factors for the disease and persists for at least a decade after the date of the mammogram used to classify density (19). In the two largest cohort studies (the Breast Cancer Detection and Demonstration Project and the Canadian NBSS2), it was found that about a third of breast cancer could be attributed to density in $>50\%$ of the breast (19, 22).

Mammographic density differs from other risk factors for breast cancer in a number of respects. The relative and attributable risk of breast cancer is stronger than almost all other risk factors, density reflects events in the organ at risk itself, and mammographic density can be changed. Mammographic density is reduced by increasing age, parity, menopause, by greater body weight (see Refs. 22, 23 for reviews), by Tamoxifen (24, 25), a gonadotrophin-releasing hormone agonist (26), and is increased by hormone replacement therapy (27).

The detection of change in mammographic density is the subject of several research projects now in progress. There has, however, been little systematic examination of how change in serial mammograms is best detected. The purpose of the work described here was to examine the effects of different reading conditions on the detection of change in mammographic features, using a computer-assisted method of measurement, in serial mammograms that have been the subject of a previously published study and in which change has been described previously (28).

Materials and Methods

Background

In previous work, we examined longitudinally the effect of the menopause on mammographic densities in women in the Canadian NBSS. The results showed that the menopause has effects on characteristics of the mammogram that are distinct from the effects of age. These effects were a reduction in the...
area of radiologically dense tissue, an increase in the area of non-dense tissue, and a decrease in the percent density. The mammograms used in this previous study were also used in this study because we were confident that there was a change in mammographic measures at menopause and that the changes were large. Furthermore, the study compared all four mammographic measures (total breast area, area of dense tissue, area of non-dense tissue, and percent density) over time, but also between two groups of women (see below), one group experienced menopause and the other did not.

**General Method**

The general method used was to identify women within the Canadian NBSS, a multicenter-randomized controlled trial of screening for breast cancer, and who had been allocated to the mammography arm of the study in which they received annual mammography for 5 years. We selected subjects who were premenopausal at entry and had undergone menopause after entry. We compared the mammogram obtained at entry with the mammogram that most closely followed menopause, using a computer-assisted method of measurement. The changes seen in the mammograms of these subjects were compared with those in an individually age-matched group of women who were also premenopausal at entry, had been followed for the same length of time, and had not experienced menopause. The first group we refer to as the prepostmenopausal group and the second the premenopausal group.

**Reading Methods**

All measurements were made on one cranio-caudal view for each subject. The same side (left or right) was used for both images that were compared within each subject and for both members of the matched pairs of subjects. Mammograms were measured using the computer-assisted method shown in Fig. 1. An operator placed a threshold on the edge of the breast (white line in Fig. 1) and on the edge of dense breast tissue (black line in Fig. 1). The computer then calculated the areas defined by these thresholds. The mammograms were repeatedly measured by the same reader (N. F. B.) using the same computer-assisted method, however, using four different methods to randomize and present the images to the reader. The methods used were used in the same order as they are described below, and their order was not determined randomly. Each randomization method involved six sets of films each containing two mammograms for each of the 404 subjects (202 pairs) read over the course of ~1 week, each set in ~1 h. Average reliability for measuring percent density was assessed by rereading a 10% random sample of images, within and between each session.
The four reading methods of randomization and viewing were spread out over the course of 1 year.

**Randomization of Films.** The first method to remeasure the mammograms was designed to mimic and/or recreate the original effect. As with the original, each set contained both films for an individual but in random order, *i.e.*, the unit of randomization was the film.

**Randomization of Subjects, Films within Subject Viewed One after the Other, Order Unknown.** The second method involved reading the films of each subject in pairs, one after the other, but in random order and unknown to the reader, *i.e.*, the unit of randomization was the subject.

**Randomization of Subjects, Films within Subject Viewed One after the Other, Order Known.** The third method was the same as the second, except the order of the films within a subject was sequential (not random) and known to the reader, *i.e.*, the films were read in succession, in the order in which they had been taken.

**Randomization of Subjects, Films within Subject Viewed Simultaneously, Order Unknown.** The fourth and final method involved simultaneous viewing of both films for each subject on two separate computer monitors in random and unknown order.

Examples of each of the methods of randomization and viewing are given in Fig. 2.

**Statistical Methods**

Pearson’s correlation coefficient was used to test the reliability of each of the mammographic measures within and between reading sets.

A series of paired *t* tests was used to compare the differences in the means of the four mammographic measures. The first set of paired *t* tests compared the change over time (before and after menopause or equivalent) within a subject for each of the prepostmenopausal and premenopausal groups. The second set of paired *t* tests compared the differences of this change between groups. A RCBD was used to test the hypothesis that there was no difference between these differences across the methods of randomization and viewing.

Jackknifing was used to compare the variation in reading methods rather than the differences in the means. Jackknifing repeatedly calculated the log variance while deleting one observation at a time from the observed data. The log of the sample variance of the between-group differences (pre versus postmenopausal) in the within-subject change (before versus after) of each mammographic measure is jackknifed for each method of randomization and viewing. A RCBD was applied to the new jackknifed estimates to test the hypothesis that there was no difference between the methods (29).

**Results**

Reliability for the computer-assisted measurements within sets was assessed by repeated measurements of 84 images. Reliability between sets was assessed by repeated measurements of 14 images read in sets 1, 3, and 6 (interset repeats were accidentally omitted from method 1). The correlation coefficients are given for each of the mammographic measures within and between each of the randomization methods in Table 1. The within-set correlation coefficients were all > 0.84, indicating a high degree of reliability. Because the percentage of dense area measurement is a ratio of two of the other measurements (dense area/total area), the between-set correlation coefficients are lowest for percent density. However, there does not appear to be a trend in the correlation coefficient’s between the methods of randomization and viewing for any of the measures.

The original published result found a mean difference of

---

**Table 1**

<table>
<thead>
<tr>
<th>Randomization and viewing method</th>
<th>Mammographic measure</th>
<th>Within-set correlation coefficient</th>
<th>Between-set correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Percent dense area</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Dense area</td>
<td>0.91</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Nondense area</td>
<td>0.98</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>Total area</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>Percent dense area</td>
<td>0.92</td>
<td>0.59–0.88</td>
</tr>
<tr>
<td></td>
<td>Dense area</td>
<td>0.84</td>
<td>0.93–0.99</td>
</tr>
<tr>
<td></td>
<td>Nondense area</td>
<td>0.96</td>
<td>0.81–0.96</td>
</tr>
<tr>
<td></td>
<td>Total area</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>3</td>
<td>Percent dense area</td>
<td>0.91</td>
<td>0.67–0.90</td>
</tr>
<tr>
<td></td>
<td>Dense area</td>
<td>0.87</td>
<td>0.92–0.99</td>
</tr>
<tr>
<td></td>
<td>Nondense area</td>
<td>0.96</td>
<td>0.89–0.98</td>
</tr>
<tr>
<td></td>
<td>Total area</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>4</td>
<td>Percent dense area</td>
<td>0.95</td>
<td>0.58–0.95</td>
</tr>
<tr>
<td></td>
<td>Dense area</td>
<td>0.93</td>
<td>0.92–0.99</td>
</tr>
<tr>
<td></td>
<td>Nondense area</td>
<td>0.98</td>
<td>0.88–0.98</td>
</tr>
<tr>
<td></td>
<td>Total area</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
the change of percent density over time between the two groups of 3.26% ($P = 0.004$). The corresponding estimates for dense area, nondense area, and total area of the breast were 3.39 cm² ($P = 0.007$), −4.37 cm² ($P = 0.01$), and −0.97 cm² ($P = 0.44$), respectively.

Table 2 shows the mean change in each of the four mammographic measures over time, for each group, and for each randomization and viewing method, with their applicable SDs. It also shows the paired between-group difference of the mammographic measures over time, for each group, and for each randomization and viewing method, with their applicable SDs. This reduced the mean paired differences and increased the between group variation, and as a result, method 3 yielded the highest $P$s for the paired differences of the four methods of randomization and viewing.

As with method 3, method 4 also gave relatively small mean paired differences for all of the mammographic measures, except total breast area (2.68, 2.98, −4.26, and −1.28 for percent density, dense area, nondense area, and total breast area, respectively). However, it was the least variable method. The introduction of two screens and simultaneous viewing of both images in each pair altered the mean change of percent density within a group closer to zero (−0.65 for the pregroup and −3.32 for the prepost group) while reducing the variability. This was also true for all of the other mammographic measures for the prepost group.

The last column of Table 2 gives the $P$s from the RCBD, which tested the null hypothesis that the mean paired differences for each of the methods of randomization and viewing were equal. The null hypothesis could not be rejected for any of the mammographic measures ($P = 0.2194, 0.3146, 0.1408,$ and $0.5102$ for percent density, dense area, nondense area, and total breast area, respectively). Thus, the mean paired differences were not significantly different across the four methods of randomization and viewing.

Table 3 shows the mean of the jackknifed log variance of the four mammographic measures for each randomization and viewing method. For all measures except total breast area, method 1 yielded the largest mean log variances (6.17, 6.94, 7.09, and 6.41 for percent density, dense area, nondense area, and total breast area, respectively), whereas method 4 yielded the smallest (5.12, 5.68, 6.19, and 5.65 for percent density, dense area, nondense area, and total breast area, respectively).

The corresponding $P$s from the RCBD were all $<0.001$, which was evidence to reject the null hypothesis that the mean log variances were equal across the four methods of randomization and viewing.
variances for each of the methods of randomization and viewing were equal. Within a mammographic measure, all multiple comparisons were statistically significant, with the (borderline) exception of total breast area. The mean log variance for total breast area in method 3 (6.01) was not significantly different from method 4 (6.02; $P = 0.06$).

**Discussion**

With the exception of method 3, all of the randomization and viewing methods confirmed the original finding of change in all mammographic measures at menopause and produced overall results very similar to each other, suggesting that mammographic density is a robust measure. The purpose of the different randomization and viewing procedures used in methods 2, 3, and 4 was to eliminate the variation attributable to viewing the films at different times. Although these procedures appeared to reduce the within-subject variability across the methods, they also provided the reader with additional information that introduced various levels of bias to each of the randomization and viewing methods. This bias, in general, moved the mean changes toward zero, perhaps because of the knowledge that the films came from the same individual. Knowledge of the order of the films (method 3) appeared to introduce a different bias, moving the mean changes of both groups toward the same nonzero level. This perhaps reflects expectation of a change over time, although the reader was aware of this possibility. Knowledge of the sequence of the films had the effect of making smaller changes appear larger, whereas larger changes appeared smaller. This may explain why in method 3 the mean changes were larger in the premenopausal group and smaller in the prepostmenopausal group for each of the mammographic measures relative to the other randomization and viewing methods. Method 3 was the only one in which the order of films within a subject was known and illustrates a form of bias that is unfavorable and unique to this method, making it the least desirable method of randomization and viewing.

Despite the various levels of bias in each of the randomization and viewing methods, the mean paired differences were not significantly different for any of the four mammographic measures. A probable explanation is that the introduction of bias only effects the measurement within subjects, not between subjects. The same amount of bias was introduced to both groups (with the exception of method 3 for reasons described above) and was eliminated after taking the difference between groups.

Although the mean paired differences were not different across the methods of randomization and viewing, the variability between the methods was significantly different for all four mammographic measures. Method 4 produced the lowest mean log variance for all measures, except total breast area; however, this method raises certain practical issues that are less than ideal. The first, of course, was the tradeoff between the reduction of variability and the bias of the mean change toward zero. Simultaneous reading without knowledge of sequence (method 4) allowed the reader to adjust the films so that their appearance on the screen was similar in terms of brightness and contrast. The images appeared more alike, which reduced the within-subject variability but also reduced the mean change within subjects. Films belonging to the same women may have been taken on different machines with positioning, compression, and film processing. Adjusting brightness and contrast of the images so that they appear as alike as possible on the viewing screen may not adequately control these sources of variability. The second is the inconvenience of requiring two computers. Both monitors must be of the same make with identical settings and the time required to prepare a set of films for reading is about doubled.

Overall, method 2 appears to be the best method of randomization and viewing. It introduces a small amount of bias, relative to method 1, toward zero in the mean change. Estimates of mean change were slightly smaller than those of method 1 in which no information is provided to the reader, whereas in method 2, pairing of mammograms is known, but the order in which they were taken is not known. The effect of method 2 on mean change, relative to method 1, is additionally reduced after taking the differences between groups. The slight reduction in effect size with method 2 is amply compensated by a substantial reduction in variability, making method 2 the most sensitive method with the lowest $P$s in Table 2 for all mammographic measures except total breast area.

**References**


