Factors affecting recall rate and false positive fraction in breast cancer screening with breast tomosynthesis - A statistical approach.

Rosso, Aldana; Lång, Kristina; Petersson, Ingemar; Zackrisson, Sophia

Published in:
Breast

DOI:
10.1016/j.breast.2015.08.007

2015

Document Version:
Peer reviewed version (aka post-print)

Factors affecting recall rate and false positive fraction in breast cancer screening with breast tomosynthesis – a statistical approach

Aldana Rosso[a], Kristina Lång[c], Ingemar F Petersson[a,b], Sophia Zackrisson[c]

a Epidemiology and Register Centre South, Skåne University Hospital, Lund, Sweden

b Orthopaedics, Department of Clinical Sciences Lund, Lund University, Lund, Sweden

c Department of Diagnostic Radiology, Translational Medicine Malmö, Lund University, Malmö, Sweden

Corresponding author: Aldana Rosso, aldana.rosso@skane.se, +46 70 21 27 140

Abstract

In this study, we investigate which factors affect the false positive fraction (FPF) for digital breast tomosynthesis (DBT) compared to digital mammography (DM) in a screening population by using classification and regression trees (C&RT) and binary marginal generalized linear models.

The data was obtained from the Malmö Breast Tomosynthesis Screening Trial, which aimed to compare the performance of DBT to DM in breast cancer screening. By using data from the first half of the study population (7 500 women), a tree with the recall probability for different groups was calculated. The effect of age and breast density on the FPF was estimated using a binary marginal generalized linear model.

Our results show that breast density and breast cancer were the main factors influencing recall. The FPF is mainly affected by breast density and increases with breast density for DBT and DM.

In conclusion, the results obtained with C&RT are easy to interpret and similar to those obtained using binary marginal generalized linear models. The FPF is approximately 40% higher for DBT compared to DM for all breast density categories.

Keywords: digital mammography; breast cancer screening; digital breast tomosynthesis; binary marginal generalized linear models; classification and regression trees; generalized estimating equations

Introduction

Breast cancer screening programs are believed to improve the early detection of breast cancer and thus they may help to reduce breast cancer mortality [1]. However there are negative aspects associated with screening, such as overdiagnosis and false positive cases [1]. Digital mammography (DM) is the standard technique for breast cancer screening. However, it has limitations due to the fact that DM is a two dimensional technique that depicts a three dimensional organ. Hence, cancer detection can be hampered due to overlapping tissue in the images. Laming et al. [2] has estimated that around 15% to 30% of cancer cases may not be detected when screening with DM. Digital breast tomosynthesis (DBT) is a three-dimensional imaging technique that may address some of the limitations that DM has, in particular problems related to overlapping tissue. Several recent studies have shown that the combination of DBT and DM improves the cancer detection rate [3–9].

The Malmö Breast Tomosynthesis Screening Trial (MBTST) was designed to compare the performance of one-view DBT as a single screening modality to two-view DM. The study population consisted of a random sample of 15 000 women invited to participate in the breast cancer screening program in the city of Malmö, Sweden. Women accepting to participate in the study were offered a DBT examination in addition to the DM examination at the screening visit. The first results of the screening trial, obtained after half of the study population was enrolled, were recently presented by Lång et al. [10]. The cancer detection rate for DBT was superior to that for DM, and that the overall recall rate for DBT was higher than that for DM [10].

One of the main concerns of breast cancer screening programs is the significant amount of healthy women that are recalled for further examination and then found free of breast cancer (false positive screening) [1,6,11]. It has been calculated that the cumulative risk of a false-positive screening result in women aged 50–69 undergoing 10 biennial screening tests is around 20 % [12]. The purpose of this article is to quantify the probability of a false positive screening using the MBTST data for the first half of the study population using different statistical methods. The probability of a false positive screening is also called false positive fraction. In some context it is also referred as false positive rate. We will use interchangeably the terms false positive fraction, probability of a false positive screening as well as recall probability and recall rate.

Binary marginal *g*eneralized *l*inear *m*odels (GLM) can be used to estimate how different factors would affect the recall probability for groups of women that share similar characteristics such as breast density and age. A more recently developed non-parametric tool suitable for this type of problems is called Classification and Regression Tree (C&RT) [13] . This technique is employed in clinical research with the aim to obtain a simple pattern to classify subjects between ill and healthy, and to get information about which groups of individuals could benefit more from targeted interventions [14–18]. One of the main advantages of C&RT is that the result of the analysis is a classification tree, which is easier to interpret in clinical practice [14]. However, due to the hierarchical nature of C&RT, it is not possible to estimate the effect of a single variable on the probability of recall. Therefore, we complemented the results obtained with C&RT with regression analysis. We applied C&RT to study which characteristics the recalled women have in common for both imaging methods and to present this information with a classification tree. In order to further analyse how these factors affect the probability of false positive screening we used a binary marginal GLM [19].

Materials and methods

Study population and image reading

The MBTST was a clinical trial performed at the Mammographic Clinic at the Skåne University Hospital, in the city of Malmö (Clinical Trial number NCT01091545). The study was approved by the Regional Ethical Review Board at Lund University (Dnr 2009/770) and the local Radiation Safety Board at the Skåne University Hospital in Malmö. Participating women gave written informed consent. The main characteristics of the study are discussed here. A thorough description of the study and the evaluation of the results from the analysis of the first half of the study population are presented elsewhere [10].

The Swedish Board of Health and Welfare recommends breast cancer screening with DM for women aged 40-74 at 18-24 month intervals (20). The participants of the MBTST were randomly selected from the screening population in Malmö. The women accepting to take part in the study were offered a DBT examination in addition to the DM examination at the screening visit.

Six readers with at least 8 years of breast imaging experience participated in the study. The readers had experience of DBT reading from previous studies [21–23]. Two blinded readers evaluated the DBT reading sequence independently from the two blinded readers of the DM reading sequence. The DBT sequence of images consisted of an initial presentation of a one-view DBT alone, followed by the addition of a one-view DM and finally previous two-view DM was shown if available. The DM sequence consisted of a two-view DM and then an addition of a prior two-view DM if available. The images were evaluated and scored at each step before moving to the next step according to a 5-point scale: 1. normal, 2. benign findings, 3. non-specific finding with low probability of malignancy, 4. findings suspicious of malignancy, 5. findings highly suspicious of malignancy.

If any of the readers at any step of a sequence scored at least 3 points for the case, it was discussed at an arbitration meeting, where at least two readers re-evaluated the images and decided whether to recall the woman or not [10]. Furthermore, a woman could be recalled if she reported symptoms from the breasts at the examination in spite of negative image findings.

Recalled women were assessed in accordance with ordinary screening routine [10]. The cancer cases were verified with record linkage with the South Swedish Cancer Register. For all women in the study there was at least one-year follow-up.

The breast density was also evaluated at the final step of the DM reading sequence using the 4th edition of the American College of Radiology's Breast Imaging Reporting and Data System (BI-RADS) scale for breast composition [24]: 1. The breast is almost entirely fat, 2. There are scattered fibroglandular densities, 3. The breast tissue is heterogeneously dense, 4. The breast tissue is extremely dense.

The first 7 500 women participating in the trial were examined in January 2010 - December 2012. In this population, 352 women were recalled for further examination (282 recalled in the DBT sequence and 197 in the DM sequence) [10]. The total number of screening detected cancer cases was 68 (67 cases detected in the BT sequence and 47 in the DM sequence) [10]. In this sample, 6 640 women

had a density evaluation. Those without density evaluation were not included in this analysis. The group of women without density evaluation had similar age distribution to the studied population and were neither recalled nor had cancer. The population characteristics were discussed in a previous publication [10]. The most important parameters of the sample for the analysis are listed in Table 1.

Total number of women		6 640
Number of recalled women	Total	352
	Recalled in DBT reading sequence	282
	Recalled in DM reading sequence	197
Cancer cases	Total	68
	Detected in DM reading sequence	47
	Detected in DBT reading sequence	67
Age	Median	54.3
	Min	39.7
	Max	75.9
	38-49	35.0 %
	50-59	28.2 %
	60-76	36.8 %
Breast Density (BI-RADS)	1	19.8 %
	2	37.8 %
	3	34.0%
	4	8.5%

Table 1: Main characteristics of the study population [10]. All women in the study had at least one-year follow-up.

Classification and regression tree

Classification and Regression Tree is a non-parametric technique that splits the data into different groups by searching which variables separate the data the most with respect to the response variable [14]. The separations performed in C&RT are binary. A brief introduction to this method is presented in the Appendix A. The aim of the analysis was to provide a clear visualization of which groups of women were recalled in the DM and DBT reading sequences. Furthermore, the tree also provided an estimate of the predicted probability of recall for the different groups.

The analysis was performed using The Salford Predictive Modeller Software Suite, version 7. The Gini impurity index was used as splitting criteria and the obtained trees were validated using 10-fold cross validation (see Appendix A for further discussion). The response variable was whether the woman was recalled or not. We calculated separate trees for the DM and DBT reading sequences. The variables included in the models were breast density, cancer status (whether the women had breast cancer or not) and age at examination. The variables breast density and cancer status were included as categorical variables and the variable age at examination was included as a continuous variable. The ratio between the obtained proportions of recalled women for different groups and the 95 % confidence interval were estimated using the McNemar test. These calculations were performed in Stata version 13.1 using the command *mcc*.

Binary marginal generalized linear models

The goal of the analysis was the estimation of the false positive fractions for DBT and DM separately, meaning the estimation of the marginal probabilities of recalling a woman without breast cancer for

each method. In order to estimate how each covariate affected the probability of a false positive screening a marginal generalized linear model for binary outcomes was used [19]. Generalized estimating equations (GEE) were employed to fit the model [19]. Since the same women were examined with both methods (paired design), we took into account the intragroup correlation in the model. Paired designs are more efficient than unpaired designs with the same amount of subjects. The most frequently used link functions are the logit-link, the probit-link and the log-link. The logit and probit functions behave well numerically, however the interpretation of the results is not straightforward. On the other hand, the log-link has the advantage that it allows interpreting the model coefficients directly in terms of the relative FPF. When the log link is used, the fitted probabilities may exceed 1, although this is rarely observed in practice (19). In order to facilitate the interpretation of the results, we applied the log-link function in the model.

Since we calculated the FPF, only data for breast cancer free women was included in the model (n = 6572). We expected that the screening method, the breast density and the age at examination would influence the FPF. The variables breast density and cancer status were included as categorical variables and the variable age at examination was included as a continuous variable.

The modelling process was performed in several steps. Firstly, a model for the FPF including the main effects screening method, density, age, and the interaction term between screening method and density was assessed. This interaction term was chosen since the DBT images generally show more features than the DM images, especially in dense breasts. The main coefficients for method and density were statistically significant and the coefficients for age and for the interaction terms were not statistically significant. Secondly, we fitted a reduced model with only the main effects age, method and breast density. The term age was not statistically significant. Finally, we fitted a model containing only the covariates method and density, which were statistically significant.

All the calculations were performed Stata version 13.1. The model was fitted using the *glm* command and the expected FPF for different density levels were calculated using the command *margins*. The statistical significance of the model coefficients was assessed using the Wald test. The area under the receiver operating characteristic curve was calculated using the command *roctab*. The statistical significance level was 5 % for all the calculations.

Results

Classification and regression tree

The classification tree for both reading sequences DM and DBT is shown in Figure 1. The number of women in the sample at each classification step is also indicated in Figure 1. As expected, the most important factor determining whether the women were recalled was the breast cancer status (breast cancer yes or no). Breast density was the main splitting factor for breast cancer free women. In order to evaluate the performance of the classification tree to discriminate which women were recalled we calculated the area under the receiver operating characteristic curve (AUC ROC). The AUCs ROC for the DBT and for the DM trees were 0.70, indicating that the classification structure was good. In the case of DM, we also found a tree in which age was a classification variable for women with density 3 or 4. This model had an AUC ROC= 0.72. The objective of this part of the analysis is to find a pedagogical way to present information about which groups of women that are recalled. Therefore,

we preferred to use the simpler tree with fewer classification variables since the improvement in the AUC given by the inclusion of age in the model was negligible.

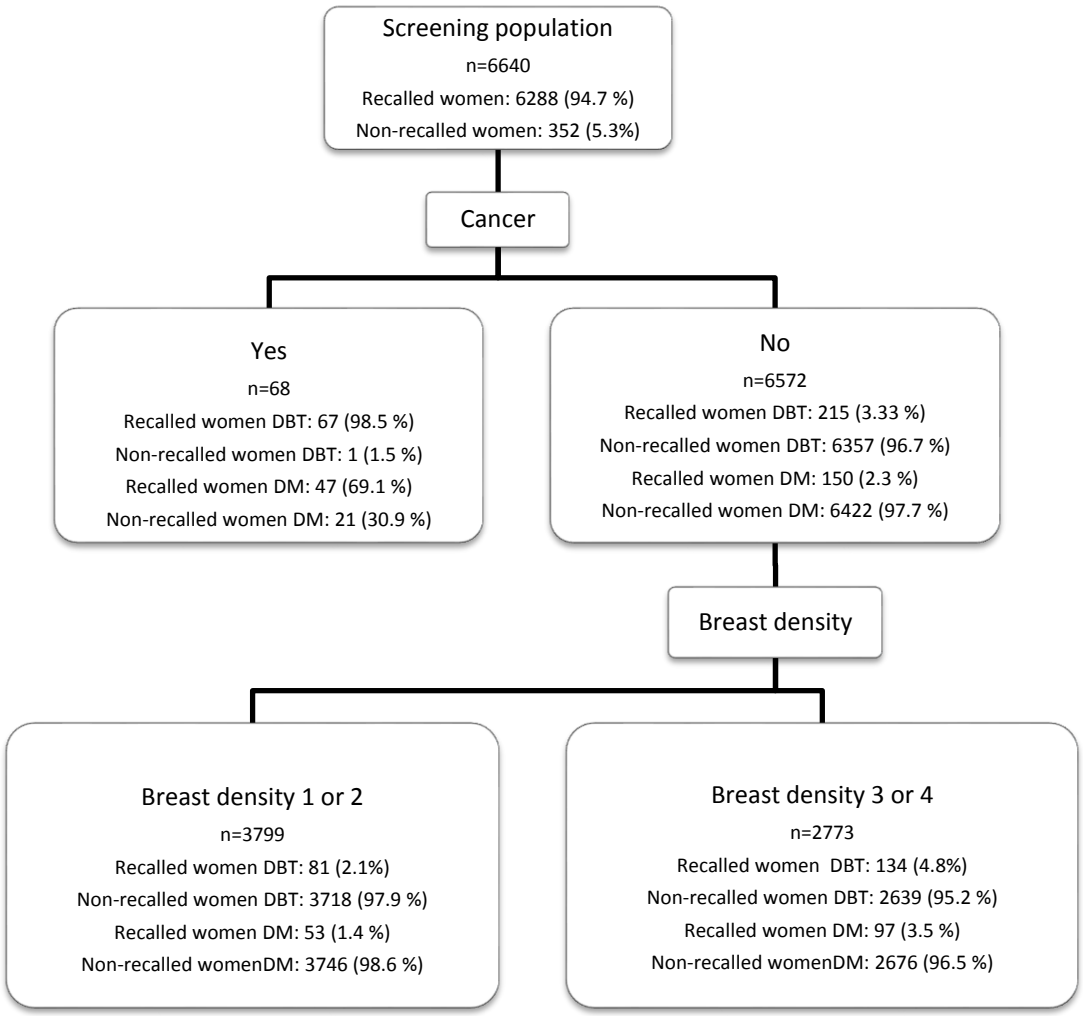

Figure 1: Classification tree for recalled women. **At each classification step, the number of women in the sample and the corresponding percentage are indicated.**

The predicted probability of a woman to be recalled can be estimated from the proportion of women recalled with the same characteristics. The results are listed in Table 2. The ratio between the proportions from the different groups and the 95 % confidence intervals are also shown in Table 2.

Group	n	Recalled proportion (%) with 95 % confidence interval		Ratio DBT/DM with 95 % confidence interval
		DBT	DM	
Women with breast cancer	68	98.5 (92.1, 100.0)	69.1 (56.7, 79.8)	1.4 (1.2 , 1.7)
Breast cancer free women with breast density (BI-RADS) 1 or 2	3799	2.1 (1.7, 2.6)	1.4 (1.0, 1.8)	1.5 (1.2, 2.0)
Breast cancer free women with breast density (BI-RADS) 3 or 4	2773	4.8 (4.1, 5.7)	3.5 (2.8, 4.3)	1.4 (1.1 , 1.7)

Table 2: Proportions of recalled women for different groups of women obtained with the classification tree. The exact 95 % confidence intervals are also shown. The ratio between the proportions for DBT and DM and the 95 % confidence intervals are calculated using the McNemar test.

The probability to recall a woman with breast cancer is approximately 99 % and 70 % for DBT and DM, respectively. The probability of recalling a breast cancer free woman increases with breast density for both methods. The ratio between the recalled proportions using DBT and DM for the different groups is about 1.4 for all cases.

By using the breast distribution and the breast cancer incidence of this study (Table 1) we estimate how many women would be false positive per 1000 screened women. We would expect approximately 12 and 20 false positive screenings with DBT for women with low (BI-RADS 1+2) and high (BI-RADS 3+4) breast density, respectively. By using DM we would expect around 8 and 15 false positive screenings for women with low and high breast density, respectively.

Binary marginal generalized linear models

In order to further investigate how breast density and age affect the FPF for DM and DBT we calculated a binary marginal generalized linear model [19]. The fitting procedure was described in detail in the section Material and methods. Briefly a model containing the main effects age, breast density and imaging method and an interaction term between imaging method and breast density was fitted. This model was then reduced by removing the non-statistically significant terms in several steps until a model with only statistically significant terms was achieved. The results for the model for the FPF with covariates imaging method and density are listed in Table **3**. A more detailed table with the coefficients of the model is listed in Appendix B. The FPF of DBT is larger than that of DM. The ratio between FPF for DBT and DM is 1.432 with 95 % confidence interval (1.226, 1.673). In accordance with the results from the C&RT analysis, this model also indicates that the breast density affects the FPF for both methods in a similar way. When the breast density increases, the FPF increases. The FPF approximately doubles for breasts with density BI-RADS = 2 compared to breasts with density BI-RADS = 1 for both DM and DBT. For breasts with density BI-RADS = 4, the FPF is approximately five times higher than for fatty breasts, BI-RADS = 1.

Covariates	Parameter	Estimates with 95 % confidence interval	p-value
Method Density (density BI-RADS =1 used as reference level)	FPF_{DBT}/ FPF_{DM}	1.432 (1.226,1.673)	<0.001
	$FPF_{Density\ BI\text{-}RADS=\ 2}$/ $FPF_{Density\ BI\text{-}RADS=\ 1}$	1.898 (1.180, 3.052)	0.008
	$FPF_{Density\ BI\text{-}RADS=\ 3}$/ $FPF_{Density\ BI\text{-}RADS=\ 1}$	3.367 (2.135, 5.308)	<0.001
	$FPF_{Density\ BI\text{-}RADS=\ 4}$/ $FPF_{Density\ BI\text{-}RADS=\ 1}$	5.304 (3.197, 8.802)	<0.001

Table 3: Estimates for a model with breast density and method as covariates. All the estimates are rounded at the third decimal place. The total number of women included in the analysis is 6572 (density BI-RADS 1 = 1307, density BI-RADS 2 = 2492, density BI-RADS 3 = 2221, density BI-RADS 4 = 552).

The estimated FPFs at different density levels are listed in Table 4. The overall FPF for DM is 0.023 with 95 % confidence interval (0.019, 0.026). The overall FPF for DBT is 0.032 with 95 % confidence interval (0.028, 0.037). Since the interaction term between the variables method (DM and DBT) and breast density is not statistically significant, this model indicates that the breast density affects the FPF in the same manner for DM and DBT.

We also calculated the expected amount of recalled women per 1000 screened women as in the C&RT analysis. The results are presented in Table 4. For fatty breast (BI-RADS categories 1 and 2), we would expect around 9 false positive screening cases per 1000 examined women for both methods. For dense breasts (BI-RADS categories 3 and 4), we would expect 19 false positive cases for DBT and 14 for DM per 1000 examined women.

A measure of the performance of the model to correctly classify which women would have a false positive screening is the area under the receiver operating characteristic curve (AUC ROC). For DBT the AUC ROC was 0.62 and for DM was 0.64, indicating that the discriminant capacity of the model is good.

Density (BI-RADS)	FPF_{DBT} Estimate with 95 % CI	Number of recalled women per 1000 screening for DBT	FPF_{DM} Estimate with 95 % CI	Number of recalled women per 1000 screening for DM
1	0.013 (0.007, 0.019)	2	0.009 (0.005, 0.013)	2
2	0.025 (0.019, 0.031)	7	0.017 (0.013, 0.022)	7
3	0.044 (0.036, 0.052)	13	0.031 (0.024, 0.037)	10
4	0.069 (0.049, 0.089)	6	0.048 (0.034, 0.063)	4
Overall	0.032 (0.028,0.037)	28	0.023 (0.019,0.026)	23

Table 4 : FPF for different density values for a model with breast density and method as covariates. All the estimates are rounded at the third decimal place. The number of recalled women per 1000 screened women is calculated using the density population parameters observed in this study and listed in Table 1. The results for the number of women are rounded to the first integer.

Discussion

In this study we present a statistical analysis of the first part of the data collected in the MBTST with focus on recall rate and false positive screening. By using classification and regression trees we calculate a simple diagram with information about which groups of women that are recalled. We conclude that the main factor affecting whether a breast cancer free woman was recalled for further work-up was the breast density for both DBT and DM. In order to estimate how the FPF of DM and DBT is affected by breast density and age, we apply a binary marginal generalized linear model. The overall FPF is lower for DM than for DBT. The FPFs for both methods increase with breast density in a similar way.

In some cases, it has been shown that the predictive accuracy of C&RT is somewhat lower than for logistic models [18]. In our case, the results obtained using C&RT and the regression model are comparable. Both models have good discriminatory performance and give similar estimates for the amount of recalled women per 1000 screened women. The regression tree may be a more pedagogical alternative to the regression coefficient table in those situations where the objective is to communicate the results to clinicians in a simple way.

The comparison of our results with previous studies was not straightforward since the MBTST was the first trial with focus on the evaluation of the performance of one-view DBT relative to two-view DM in a standard screening program environment [10]. The overall FPF for DM is in agreement with some older reported values [25]. Two recent population-based screening trials have focused on the comparison of DM to the combination DM and DBT [3,7]. Ciatto and colleagues reported that the DM recall rates were approximately 4 % and 7 % for women with low breast density (BI-RADS 1 and 2) and high breast density (BI-RADS 3 and 4) [7], respectively. The values obtained in the present study are somewhat lower and differences in the study population and design as well as in the reading and

recall procedures may explain the discrepancy. Ciatto and colleagues also reported the recall rate for the combination of DBT and DM to be around 3 % and 6 % for low breast density (BI-RADS 1 and 2) and high breast density (BI-RADS 3 and 4) [7], respectively. Skaane and colleagues reported that the recall rate for the combination of DBT and DM was around 5 % [3]. These numbers are comparable with those obtained here using one-view DBT alone. Teertstra et al. [26] performed a study to evaluate the potential value of DBT in a population of women with abnormal screening mammogram or with clinical symptoms. The authors reported that the FPF for DBT was approximately 0.16 and that the FPFs of DBT and DM were similar [26]. The discrepancies between the previously reported values and those obtained here are probably due differences in the study design and population.

This study has several limitations both related to the clinical study and to the statistical methods. The limitations related to the clinical study are discussed by Lång et al [10]. Briefly, the MBTST limitations are mostly related to the fact that the study was performed in a Swedish population, with limited amount of readers and with only one type of tomosynthesis equipment [10]. Furthermore, DBT was used for the first time in this population, i.e. it should be regarded as a prevalence screening where a higher recall rate and more findings (both cancers and non-cancers) not visible at earlier DM screenings are observed.

Regarding the statistical methods, the accuracy of the model is dependent on the influence of the covariates included in the models on the recall. In this case, we had information about the breast density and age of the women. However, there may be other factors that may affect the breast composition and its appearance in the images that were not included in the models. Furthermore, we had a homogenous group of radiologists with several years of experience. Wallis et al. [27] compared the diagnostic accuracy of DM to DBT in an observer study involving two institutions and 130 cases. The study showed that two-view DBT outperforms DM but only for readers with the least experience [27]. No difference in the diagnostic accuracy of DM compared to one-view DBT were observed [27]. A similar observer study was performed at our institution involving eight breast radiologists and 185 cases to compare the ability to detect breast cancers using one-view DBT relative to two-view DM [28]. Our data showed that the diagnostic accuracy was better for DBT compared to DM for experienced readers. In the case of inexperience readers no significant difference was observed [28]. Different study settings and readers may explain the differences and further studies are needed as DBT gains more acceptance. Unfortunately the MBTST does not provide information about how one-view DBT would perform when it is used by less experienced radiologists.

The sample size is another limiting factor. Even though there were 7 500 participating women, only a small fraction of them (352) were selected for additional follow-up. In order to obtain reliable estimates for the different variables in the model, a moderate number of women in each category are needed. Finally, the performances of both models were assessed using the same set of data. In order to assess the prediction capability of the models, new data would be needed.

Conclusions

The aim of this study was to provide information about which women that are recalled for further work-up due to inconclusive results from breast cancer screening using one-view DBT compared to DM. We analysed the data using traditional parametric methods and recently developed non-parametric tools. In situations where the objective is to communicate the results to clinicians a

classification tree is recommended. The main conclusion is that both imaging modalities have limitations for women with dense breasts. The results presented here provide an important piece of information to be considered in the discussion about implementing DBT in breast cancer screening.

Contributions

Aldana Rosso, Kristina Lång, Ingemar F Petersson and Sophia Zackrisson were involved in study design, interpreted the results and wrote the manuscript. Sophia Zackrisson is the principal investigator of the MBTST. Kristina Lång and Sophia Zackrisson participated in the data collection. Aldana Rosso performed the statistical analysis.

Ethical approval

The MBTST was approved by the Regional Ethical Review Board at Lund University (Dnr 2009/770) and the local Radiation Safety Board at the Skåne University Hospital in Malmö. Participating women gave written informed consent.

Funding

This project has received funding from the Skåne University Hospital.

Conflict of interest statement

The authors declare that they have no conflicts of interest.

Appendix A: Classification and regression tree

Classification and regression tree (C&RT) is a non-parametric method applied to find associations between several variables and an outcome. The method was first developed by Breiman and colleagues [13] in the 1980s. Classification and regression tree classifies the observations into mutually exclusively groups by selecting those variables that most separate the data. The result of the analysis is a classification tree, in which all the observations are classified into groups. The separations performed are binary. The goal of the method is to achieve a tree in which all the elements in the leaves (nodes) belong to the same category. In very few cases it is possible to achieve a perfect classification and therefore there are several alternative numerical rules about how to split the data in the most efficient way. These rules (splitting criteria) attempt to minimize the impurity of the classification, meaning that most observations should share the same characteristic at each leaf (node) of the tree. All splitting criteria compare the impurity in the parent node with the impurity that would be achieved by splitting the data into two child nodes. When building the tree, a stopping rule is needed in order to decide at which point adding predictor variables does not significantly improve the performance of the model. One possible approach to find the optimal tree is to build several trees until all predictor variables are used. Then, the optimal tree can be selected by cross validation [29]. This method divides the sample into V number smaller samples and is called V fold cross validation. All possible trees are fitted until the maximum tree size is achieved using V-1 samples, and the remaining sample is used to evaluate the rate at which the cases are misclassified by the trees. This procedure is repeated using another sample as the remaining sample, until all samples are used. The misclassification cost of all the trees are then combined and applied to the tree obtained with the entire sample. The best tree is the tree with the lowest misclassification cost.

The goal of the analysis presented here was to find which variables were most important in determining the recall of a woman by using digital mammography or breast tomosynthesis. We applied one of the most frequently used splitting criterion: the Gini impurity index [14,29]. We calculated a separated tree for each screening method. The explanatory variables in the dataset were the age at examination, the breast density and the cancer status (whether the women had breast cancer). The result of the analysis is a classification tree in which cancer status and breast density were the splitting variables. These variables were selected since they gave the minimum amount of impurities. For example, women with cancer that were not recalled were the impurities in the cancer node (see Figure 1). In order to find the optimal tree, we applied a 10-fold cross validation method.

Appendix B: Results from the regression model

The coefficient of the final regression model for the false positive fraction with covariates breast density and method is listed in Table 5.

Covariates	Parameter	Estimates with 95 % confidence interval	Standard Error	p-value
Method (DM used as reference level) Density (density BI-RADS =1 used as reference level)	Method DBT	0.359 (0.204, 0.515)	0.079	<0.001
	Density Level 2	0.641 (0.166, 0.116)	0.242	0.008
	Density Level 3	1.214 (0.759, 1.669)	0.232	<0.001
	Density Level 4	1.669 (1.162, 2.175)	0.258	<0.001
	Constant	-4.696 (-5.121, -4.272)	0.217	<0.001

Table 5: Coefficients for the model for the false positive fraction with breast density and method as covariates. All the estimates are rounded at the third decimal place. The total number of women included in the analysis is 6572 (density BI-RADS 1 = 1307, density BI-RADS 2 = 2492, density BI-RADS 3 = 2221, density BI-RADS 4 = 552).

References

1. The benefits and harms of breast cancer screening: an independent review. The Lancet. 2012 Nov;380(9855):1778–86.

2. Laming D, Warren R. Improving the detection of cancer in the screening of mammograms. J Med Screen. 2000;7(1):24–30.

3. Skaane P, Bandos AI, Gullien R, Eben EB, Ekseth U, Haakenaasen U, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. Radiology. 2013 Apr;267(1):47–56.

4. Haas BM, Kalra V, Geisel J, Raghu M, Durand M, Philpotts LE. Comparison of tomosynthesis plus digital mammography and digital mammography alone for breast cancer screening. Radiology. 2013 Dec;269(3):694–700.

5. Roth RG, Maidment ADA, Weinstein SP, Roth SO, Conant EF. Digital breast tomosynthesis: lessons learned from early clinical implementation. Radiographics. 2014 Aug;34(4):E89–102.

6. Hofvind S, Ponti A, Patnick J, Ascunce N, Njor S, Broeders M, et al. False-positive results in mammographic screening for breast cancer in Europe: a literature review and survey of service screening programmes. J Med Screen. 2012 Sep 1;19(suppl 1):57–66.

7. Ciatto S, Houssami N, Bernardi D, Caumo F, Pellegrini M, Brunelli S, et al. Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study. Lancet Oncol. 2013 Jun;14(7):583–9.

8. Friedewald SM, Rafferty EA, Rose SL, Durand MA, Plecha DM, Greenberg JS, et al. Breast cancer screening using tomosynthesis in combination with digital mammography. JAMA. 2014 Jun 25;311(24):2499–507.

9. Houssami N, Skaane P. Overview of the evidence on digital breast tomosynthesis in breast cancer detection. Breast Edinb Scotl. 2013 Apr;22(2):101–8.

10. Lång K, Andersson I, Rosso A, Tingberg A, Timberg P, Zackrisson S. Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the Malmö Breast Tomosynthesis Screening Trial, a population based study. Eur J Radiol. 2015;doi: 10.1007/s00330–015 – 3803–3.

11. Elmore JG, Barton MB, Moceri VM, Polk S, Arena PJ, Fletcher SW. Ten-year risk of false positive screening mammograms and clinical breast examinations. N Engl J Med. 1998 Apr 16;338(16):1089–96.

12. Hofvind S, Ponti A, Patnick J, Ascunce N, Njor S, Broeders M, et al. False-positive results in mammographic screening for breast cancer in Europe: a literature review and survey of service screening programmes. J Med Screen. 2012;19 Suppl 1:57–66.

13. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. 1 edition. New York, N.Y.: Chapman and Hall/CRC; 1984. 368 p.

14. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Ann Behav Med Publ Soc Behav Med. 2003 Dec;26(3):172–81.

15. Mohktar MS, Redmond SJ, Antoniades NC, Rochford PD, Pretto JJ, Basilakis J, et al. Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data. Artif Intell Med. 2015 Jan;63(1):51–9.

16. Nilsson J, Ohlsson M, Höglund P, Ekmehag B, Koul B, Andersson B. The International Heart Transplant Survival Algorithm (IHTSA): A New Model to Improve Organ Sharing and Survival. PloS One. 2015;10(3):e0118644.

17. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J Clin Epidemiol. 2013 Apr;66(4):398–407.

18. Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. Stat Med. 2007 Jul 10;26(15):2937–57.

19. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. 1 edition. Oxford University Press; 2004. 318 p.

20. The National Board of Health and Welfare. Nationella riktlinjer för bröst-, prostata-, tjocktarms- och ändtarmscancervård [Internet]. 2014. Available from: http://www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/19383/2014-4-2.pdf

21. Svahn TM, Chakraborty DP, Ikeda D, Zackrisson S, Do Y, Mattsson S, et al. Breast tomosynthesis and digital mammography: a comparison of diagnostic accuracy. Br J Radiol. 2012 Nov;85(1019):e1074–82.

22. Svahn T, Andersson I, Chakraborty D, Svensson S, Ikeda D, Förnvik D, et al. The diagnostic accuracy of dual-view digital mammography, single-view breast tomosynthesis and a dual-view combination of breast tomosynthesis and digital mammography in a free-response observer performance study. Radiat Prot Dosimetry. 2010 May;139(1-3):113–7.

23. Andersson I, Ikeda DM, Zackrisson S, Ruschin M, Svahn T, Timberg P, et al. Breast tomosynthesis and digital mammography: a comparison of breast cancer visibility and BIRADS classification in a population of cancers with subtle mammographic findings. Eur Radiol. 2008 Dec;18(12):2817–25.

24. Sickles E, D´Orsi C, Basselt L, et al. ACR BI-RADS® Mammography. In: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. [Internet]. 5th ed. American College of Radiology; Available from: http://www.acr.org/Quality-Safety/Resources/BIRADS

25. Carney PA, Miglioretti DL, Yankaskas BC, Kerlikowske K, Rosenberg R, Rutter CM, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Ann Intern Med. 2003 Feb 4;138(3):168–75.

26. Teertstra HJ, Loo CE, van den Bosch MAAJ, van Tinteren H, Rutgers EJT, Muller SH, et al. Breast tomosynthesis in clinical practice: initial results. Eur Radiol. 2010 Jan;20(1):16–24.

27. Wallis MG, Moa E, Zanca F, Leifland K, Danielsson M. Two-View and Single-View Tomosynthesis versus Full-Field Digital Mammography: High-Resolution X-Ray Imaging Observer Study. Radiology. 2012;262(3):788–96.

28. Svahn T, Lång K, Andersson I, Zackrisson S. Differences in Radiologists' Experiences and Performance in Breast Tomosynthesis. In: Maidment AA, Bakic P, Gavenonis S, editors. Breast Imaging [Internet]. Springer Berlin Heidelberg; 2012. p. 377–85. Available from: http://dx.doi.org/10.1007/978-3-642-31271-7_49

29. Zhang H, Singer B. Recursive Partitioning and Applications. 2nd ed. 2010 edition. New York: Springer; 2010. 262 p.