eviCore Opinion Statement

**Digital Breast Tomosynthesis for use in Breast Cancer Screening**

**Introduction**

Despite uncertainties with the use of digital breast tomosynthesis (DBT), also called 3D mammography, clinical interest is increasing. Some researchers and patient advocacy groups are promoting its ability to improve breast cancer screenings. Preliminary studies suggest that tomosynthesis has increased specificity and sensitivity over 2D digital mammography (DM), and therefore reduced recall and biopsy rates (Lourenco et al., 2015; Citatto et al., 2013; Friedwald et al., 2014; Skaane et al., 2013). Facilities are installing units with DBT capabilities, even though standards have not been developed for this imaging modality and the FDA-approved bodies are not prepared to accredit 3D tomography (FDA, 2014).

Multiple recent events have also fueled the increased interest in DBT. In May 2013, the FDA approved Hologic’s C-View 2D imaging software for mammography (Hologic, 2013). The C-View images are generated by the 3D tomosynthesis data and bypass the need for additional 2D imaging, which results in less radiation exposure. In November 2014, the Centers for Medicare and Medicaid Services (CMS) assigned billing codes and reimbursement values to DBT. Also in November 2014, the American College of Radiology (ACR) released a position statement declaring that tomosynthesis is no longer investigational, while encouraging more studies to clarify the clinical role(s) of tomosynthesis and its long-term outcomes (ACR, 2014). At the request of the ACR, the American Medical Association (AMA) approved three new CPT® codes specific to DBT (77061, 77062, and 77063) that were introduced in January 2015. Lastly, a recent modeling study bode favorably for possible net cost saving with DBT by avoiding the use of what would have been follow-up services (other breast imaging, biopsies, etc.) (Bonafede et al., 2015).

Notwithstanding encouraging results of the recent DBT studies, shortcomings have been noted cautioning for additional research to answer critical remaining questions. This can best be illustrated in some of the responses to the Friedewald et al study, which was reported in the Journal of the American Medical Association (JAMA) June 2014 issue. Breast imaging authorities weighing in on the study noted that, “The nonrandomized design of the study by Friedewald et al precludes drawing causal inferences about the results, and the lack of long-term follow-up information limits the ability to provide definitive estimates of false-negative result rates, diagnostic accuracy, interval cancer rates, or over diagnosis.” These experts went on to state that, “The continuing controversy surrounding the most effective strategy for deploying the various available technologies continues unabated, and clear consensus is lacking on when to screen, how often, and with what tools, or even which screen-detected cancers could be managed more conservatively; only an appropriately powered multisite clinical trial of modern technology can answer the remaining questions definitively” (Pisano and Yaffe, 2014). In answering another
commentary regarding their study, Dr. Friedewald and two of the other authors acknowledged that, “It is clear that there is a learning curve associated with interpreting all mammography and for tomosynthesis, this curve will be steeper since the reader must learn to confidently ignore findings that would ordinarily be recalled on standard mammography but that do not persist on the tomosynthesis images.” In addition, the authors stated that, “More research investigating patient-level, physician-level, and site-level characteristics would be useful to further delineate how best to use this technology to maximize outcomes and personalize breast cancer screening” (Friedwald, Rafferty & Conant, 2014).

There are additional significant issues which will need to be addressed before DBT can be considered fully ready for mainstream use. For example, the results of a survey for physician members of the Society of Breast Imaging were recently reported in the Journal of the American College of Radiology (JACR) June 2014 issue. The conclusion was that, “DBT is becoming more common but remains a limited resource. Clinical guidelines would assist practices in deciding whether to adopt DBT and in standardizing which patients should receive DBT” (Hardesty et al., 2014).

Evidentiary gaps still exist regarding which subgroups of women would benefit from this technology (e.g. breast density and patient age); the optimal frequency of DBT screening, as well as the long-term impact on clinical outcomes. The ACR acknowledged these same research gaps in their November 2014 position statement, which also emphasized the need for additional studies to evaluate these remaining questions of clinical application and long-term impact (ACR, 2014). Moreover, these additional studies would ideally be prospective, randomized trials funded by sources other than the conflicted manufacturers of DBT equipment. The Lung Cancer Screening Trial taught us that patience and persistent in demanding data showing survival-based outcomes are not only attainable but valuable in confident screening decisions (The National Lung Screening Trial, 2011). This soon followed with a U.S. Preventive Services Task Force (USPSTF) Recommendation for low-dose CT screening for lung cancer (Moyer, 2013). So taking this final step, in DBT breast screening, should not be discounted simply based on favorable mid-term results. In fact, the same USPSTF has placed DBT within their current review toward the final research plan in a revised breast cancer screening recommendation (USPSTF, 2014).

At this time eviCore is reluctant to remove the experimental and investigational status for DBT despite the promising mid-term findings, until additional evidence addresses the important concerns outlined above. eviCore will diligently monitor the peer-reviewed literature for research that further defines the best use of DBT in the clinical practice setting and will amend these recommendations accordingly.
Literature Review

Two-dimensional digital mammography (DM) has surpassed film mammography as the standard practice in breast cancer screening, and is considered to be equivalent with film mammography in accuracy (Pisano, 2005; ACR 2013). Although DM screening has a specificity of 90%, a significant limitation is its inability to fully visualize the breast where there is tissue overlap. Tissue overlap can obscure and even mimic abnormalities, and is considered to be a primary source of screening recalls (Smith, 2005). The National Cancer Institute (NCI), reports that approximately 20% of cancers are missed at mammography screening (National Cancer Institute, 2012). Average recall rates are around 10%, and the average cancer detection rate is 4.7 per 1,000 screenings (Rosenberg et al., 2006). Inaccurate or equivocal results may lead to biopsies, emotional stress, or a delay in diagnosis. Brodersen and Siersma (2013) reported that false-positive mammography screening has long-term psychosocial consequences. To overcome these limitations with DM, digital breast tomosynthesis (DBT) was developed to improve the accuracy of mammography by using 3-dimensional images to clarify the overlapping tissue (Blue Cross Blue Shield, 2014). Digital breast tomosynthesis builds on the current 2-dimensional digital mammography capabilities. While conventional mammography captures one image of overlapping tissue, digital breast DBT captures multiple low-dose images per view, along an arc of the breast which further clarifies the overlapping breast tissue. Proponents of DBT believe that this technique is superior to DM at visualizing dense breast tissue as well as detecting smaller lesions in non-dense breasts (Helvie, 2010).

The following literature review is not intended to be a comprehensive review of all published studies regarding DBT. Instead, the intent is to provide a glance at some of the most recent studies that have examined its cancer detection and recall rates in screening mammography as well as its impact on diagnostic procedures.

In a recent retrospective analysis (Lourenco et al., 2015) comparing DM to DBT screening examinations, the authors found a decrease in overall recall rates among DBT (6.4%) compared to DM (9.3%). When examined further, the recall rate was lower with DM for masses, distortions, and calcifications, whereas DBT had lower recall rates for asymmetries and focal asymmetries. The use of ultrasonography without additional mammographic views in diagnostic evaluations increased with DBT. The biopsy positive predictive value (PPV) and cancer detection rate (CDR) did not differ significantly between DM and DBT in this study. Interestingly, the authors noted that most of the previous studies which have reported an increase in the CDRs were from Europe, where screening is performed on a biennial basis. They stated that, “One would thus expect more cancers within a population screened every 2 years as compared with a population screened every year, as is common practice in the United States.” There were limitations such as the retrospective design, which prevents causal inference from being drawn from their results. In addition, it was noted that potential differences in the patient populations could have affected the results and an accurate assessment of false-negative findings was not performed since many patients had not yet returned for subsequent imaging.
Friedwald and colleagues (2014) conducted a retrospective analysis of 454,850 mammography screenings from 13 breast centers with 139 interpreting radiologists. The study compared the screening performance of DM alone for one year, ending on the date of tomosynthesis introduction at each institution, to that of screening DM plus DBT (DM/DBT) after its introduction. Recall rates were lower with DM/DBT than with DM alone, 91 versus 107 per 1000 screens. The biopsy rate was higher for DM/DBT than with DM alone, 19.3 versus 18.1 per 1000 screens. The CDR was higher for DM/DBT than with DM alone, 5.4 versus 4.2 per 1000 screens. When examined by histology, the detection of invasive cancer was greater with DM/DBT than with DM alone, 4.1 versus 2.9 per 1000 screens. There was no difference in the detection of ductal carcinoma in situ (DCIS) cancers between the screening modalities. The study has several limitations, such as the lack of a randomized trial design, which raises the possibility that results were not purely due to the addition of DBT, and the lack of data regarding potential variations in patient characteristics. They were unable to evaluate the number of repeat examinations and follow-up data were not available that would allow evaluation of false-negative result rates. The authors conclude with the statement that, “Further studies are needed to assess the relationship to clinical outcomes.”

Greenberg and colleagues (2014) retrospectively compared the performance metrics of screening with DM alone versus that of DM combined with DBT (DM/DBT) in a community radiology practice. The results revealed recall rates of 16.2% for DM alone and 13.6% for DM/DBT. The overall biopsy rate of 26.3 per 1,000 patients screened for the DM/DBT cohort was higher compared to 21.6 per 1,000 in the DM alone cohort. The CDR for DM alone was 4.9/1,000 and 6.3/1,000 for DM/DBT. However, there was no significant difference in the PPV for biopsy. Both overall and non-cancer biopsy rates were higher in the DM/DBT group. The study was limited by its retrospective and nonrandomized design that could introduce selection bias. For example, patients with known risk factors for developing breast cancer self-selecting DM/DBT screenings could have contributed to a higher PPV for recalls from screening (PPV1) and CDR in this cohort. It should be noted that the upper end of the recommended recall rate is 14% and that this groups’ DM recall rate of 16.2% is higher and their DM/DBT 13.6% recall rate is just below that benchmark. The authors point out that additional research regarding which patient subgroups will benefit most from DBT screening and the effect of repeat DBT screening on interval CDRs may be useful.

Haas and colleagues (2013) also compared the screening recall rates and CDRs of DM alone and DM plus DBT (DM/DBT). All women who presented for mammography screening in a one-year period at four different clinical sites were asked to participate in this study; one site was a mobile mammography clinic. Those who presented to sites with DBT capabilities were offered screening with DM/DBT, while those who were screened at sites without DBT underwent DM alone. Excluded were patients with breast implants or large breasts requiring tilted images. A total 13,158 women presented for screening. Of these women, 6100 underwent DM/DBT and 7058 underwent DM alone. Patients receiving DM/DBT were more likely to have a personal
history of breast cancer or family history of breast cancer. The overall recall rate was lower for the DM/DBT group (8.4%) than the group who received DM alone (12%). In addition, recall rates were significantly reduced with the use of DBT for the almost all breast density classifications: scattered fibroglandular density, heterogeneously dense breasts, and extremely dense breasts. The authors stated that the number of patients needed to prevent one recall for each of these density groups was 37.9, 15.2, and 11.2, respectively. Similarly, reductions in recall rates for DM/DBT compared to DM alone were found across all age groups. Regression analyses revealed that patients younger than 40 had the greatest reduction in odds of recall from the use of DM/DBT. Despite the significant reductions in recall rates, the CDRs were comparable between the two screening modalities (DM/DBT=35 cancers, 5.7 per 1,000 CDR; DM=37 cancers, 5.2 per 1,000 CDR). The study was limited by its retrospective data collection method and non-randomized patient selection. It should also be noted that two of the study’s authors are paid consultants for Hologic.

Skaane et al. (2013) provided preliminary results from two study arms of the Oslo Tomosynthesis Screening Trial, which is a larger four-arm population-based screening program. This sub-study aimed to compare CDRs, false-positive rates before arbitration, and PPV between DM alone and DM combined with DBT (DM/DBT). A total of 12,621 of the 29,652 fifty to sixty nine year old women, who were invited to participate in biennial screening mammography consented to do so and were included in the final analysis. Excluded were women who were unable to stand during the screening and women who had breast implants. The standard practice was independent double reading with any mammogram judged to be suspicious for malignancy by at least one reader being discussed at a consensus based arbitration meeting. False-positive rates before arbitration were 61.1 per 1,000 examinations with DM alone and 53.1 per 1,000 examinations with DM/DBT. After arbitration, PPVs for recalled patients with cancers verified later were comparable between the two cohorts. The cancer detection in the DM/DBT group at 8.0 per 1,000 was greater than the 6.1 per 1,000 in the DM alone group. Twenty-five additional invasive cancers were found with DBT, but there was no improvement in the detection of DCIS. The study had significant limitations such as the inclusion of only 50–69 year old women who were screened biennially which will probably result in a higher CDR than that expected during the annual screening which is the usual practice in the United States. The consensus-based arbitration process is also a variable that is not part of the usual U.S. practice. Since they were not allowed to record the reasons why some women declined to participate, there may have been some self-selection bias. In addition, the interpretation load for each reader could not be balanced and the analyses were adjusted for individualized performance levels. Equipment and financial support were provided by Hologic, the manufacturer of tomosynthesis equipment. The institution received financial support for additional overtime readings from Hologic.

In a follow-up publication to the study described above, Skaane and colleagues (2014) compared the performance of full field digital mammography (FFDM) plus DBT with two versions of synthesized 2D images plus DBT. This was a two-year study conducted as part of the Oslo
breast cancer screening program. The initial version of reconstructed 2D images was used in period 1 and the current version in period 2 with both being compared to FFDM plus DBT. Excluded were women who were unable to stand during the screening and women who had breast implants. Eight radiologists, with an average of 16 years of experience in mammography, interpreted the images. The study found that the CDRs for the current version of 2D reconstructions used in period 2 was 7.7/1,000 compared to the FFDM plus DBT which had a rate of 7.8/1,000. The false-positive rates in period 2 of the trial were comparable for the FFDM plus DBT group at 4.6% and the current synthesized 2D images plus DBT at 4.5%. No “substantial differences” were found between the two modalities in breast density. There were no significant differences in CDRs between modalities, nor did the PPV of recalls differ significantly. The authors concluded that current synthesized 2D images with DBT performed comparably to that of FFDM plus DBT. The study had limitations associated with the consensus/arbitration step which could have preferentially decreased the recall rates on the basis of only one of the two reading modes being compared and later dismissed during arbitration. Since the reference standard for positive cases was a cancer detected in any of the study arms, absolute sensitivity and specificity cannot be calculated until the completion of a 2-year follow-up. There were also limitations due it being a single-institution, single-vendor study with vendor-specific software. Note that one of the authors has received travel costs and a fee for one lecture from Hologic.

Rafferty and colleagues (2013) examined radiologists’ diagnostic accuracy and recall rates with DM alone and DM with DBT (DM/DBT). In this study, 997 participants from across five sites underwent two sequential imaging sets at their screening examinations, one with DM followed by another with DBT. It should be noted that women presenting for screening mammography or for breast biopsy were invited to participate. After each screen, a breast imager read the DM images, and a second imager (trained in tomosynthesis) read the DBT images. The breast imagers were blinded to the results of the other readings. Radiologists initially scored each study without any prior imaging or clinical history on each patient and then after integrating this information. Recall of patients for additional evaluation was based on the recommendation of either reader. Retrospective comparisons of DM to DM/DBT to assess overall diagnostic accuracy and recall rates of non-cancer cases for the two methods were performed. They found that the DM/DBT diagnostic accuracy and the reduction in recall rates for non-cancer cases were superior to that of DM alone. Almost nearly all of the gain in reader performance was attributable to non-calcification cases with a non-significant increase in diagnostic accuracy for calcification cases by using DM/DBT. However, it was noted that circumscribed lobulated masses, were being inappropriately dismissed by some readers. This study had several limitations, such as its highly enriched nature with the inclusion of biopsy lesions, both malignant and nonmalignant, as well as cases identified for recall by the accruing sites. The results of these enriched reader studies may be different from those of a true screening population. Therefore the magnitude of reduction in recall rate may not be as substantial in
clinical practice. Note that one author is an employee of Hologic, another is a consultant and statistician for Hologic and a third owns stock in Hologic.

Ciatto and colleagues (2013) reported initial results from the Screening with Tomosynthesis or Standard Mammography (STORM) study. STORM is a prospective study that compared DM plus DBT (DM/DBT) with DM alone. The study population included 7,292 women who were 48–71 years old. Patients underwent both DM and DBT at their screening. The radiologists interpreted the DM images first, followed by interpretation of the integrated DM/DBT images. Initial STORM results found that the CDR was higher for DM/DBT (8.1 per 1,000 screens) compared to DM alone (5.3 cancers 1,000 screens). The low number of women with high density breasts limited the analysis by breast density. The STORM protocol involved double-reading by experienced breast radiologists, and as such their results might not apply to other screening settings. It appears that they also use a standard 2 year interval between screening rounds which is less frequent than the annual standard in the United States. The researchers did not assess the effect of repeat (incident) screening with integrated DM/DBT on cancer detection, which might have a smaller impact on cancer detection rates. While two patients from the study developed interval cancers (follow-up range 8–16 months), data for interval cancers were otherwise not available. It was noted that studies of interval cancer rates after integrated DM/DBT would need to be randomized, controlled trials with a very large sample size. Interestingly the authors stated that, “Our results do not warrant an immediate change to breast-screening practice, instead, they show the urgent need for randomized controlled trials of integrated 2D and 3D versus 2D mammography, and for further translational research in breast tomosynthesis. We envisage that future screening trials investigating this issue will include measures of breast cancer detection, and will be designed to assess interval cancer rates as a surrogate endpoint for screening efficacy.” It should be noted that 10 of the 12 authors received assistance from Hologic, including tomosynthesis technology and technical support for the duration of the study as well as travel support to attend collaborators’ meetings. An eleventh author received travel support from Hologic to attend a collaborators’ meeting.

More recently, Houssami and colleagues (2014) provided additional findings from STORM. The aim of this study was to examine differences in the number of detected and non-detected (including interval) cancers, CDR, incremental CDR and false positive rate between single-reading and double-reading using conventional DM and DM/DBT. Using the same STORM study population described above (Ciatto et al., 2013), there were 65 breast cancers including six interval cancers. For single-reading, 35 cancers were detected with both DM and DM/DBT, 20 cancers were detected only with DM/DBT compared with none with DM alone and 10 cancers were not detected. For double-reading, 39 cancers were detected with both DM and DM/DBT, 20 were detected only with DM/DBT compared with none detected with DM alone and six cancers were not detected. The incremental CDR attributable to DM/DBTs of 2.7 per 1,000 versus DM alone was evident for both single and for double-reading. Incremental CDR attributable to double-reading versus single-reading of 0.55/1000 screens was evident for both
DM alone and DM/DBT. Moreover, the false positive to true positive ratios (FP/TP) showed that DM/DBT had higher sensitivity and better FP to TP trade-off than DM alone for single and double-reading methods. This study has limitations which were detailed in the initial STORM study above. It was noted that the sequential screen-reading may have over-estimated DM/DBT cancer detection. The authors recommended that these findings be used to assist in designing randomized controlled trials to assess initial outcomes and the impact on interval cancer rates as well as whether the additional cancer detection from DM/DBT will confer incremental screening benefit. The centers and investigators participating in this trial received assistance from Hologic in the form of tomosynthesis technology and technical support for the duration of the study as well as travel support to attend collaborators’ meetings.

In addition to screening mammography, some researchers have studied the accuracy of tomosynthesis for the purpose of diagnostic mammography. One such study (Zuley et al., 2013) compared the diagnostic accuracy of supplemental conventional mammography views and DBT in 182 patients with 217 verified breast lesions. These lesions were identified initially with standard two-view mammography, ultrasonography, magnetic resonance imaging (MRI), or clinical examination. Each lesion was interpreted twice by every reader; once with DBT and the standard mediolateral oblique (MLO) and craniocaudal (CC) mammography views, and once with the standard MLO and CC mammography views as well as the standard mammographic supplemental views. Afterwards, the radiologists provided a number from 1 to 100 for a probability-of-malignancy rating, as well as a Breast Imaging Reporting and Data System (BI-RADS) assessment for each lesion. The findings indicated that DBT improves diagnostic accuracy for breast soft-tissue findings such as masses, asymmetries, and architectural distortion compared with the standard supplemental diagnostic mammographic views. The false-positive rate decreased from 85% to 74% for cases that were rated as BI-RADS 3 or higher, and from 57% to 48% for cases that were classified as BI-RADS 4 or 5. There was no associated decrease in specificity. This study was retrospective and as such the results may not completely translate to clinical practice. The age distribution in their study population, women aged 60 years or younger, is not representative of the entire screened population. In this classification study, the lesion locations were marked for the readers so this study cannot be used to demonstrate that DBT will increase cancer detection. Only non-calcified lesions were assessed, therefore the results cannot be extrapolated to the assessment of calcified lesions. Due to these limitations, the study’s results may not be applicable to all diagnostic cases. Note that multiple authors of this study received grants from Hologic with at least one having travel expenses paid by Hologic to attend a scientific advisory meeting.
Conclusion

Digital breast tomosynthesis is being lauded for its potential benefits with improved screening and diagnostic sensitivity, recall rates, lesion detection and characterization (Helvie, 2010). To date, single and multiple center studies have reported improvements in sensitivity and reductions in recall rates as well as improved cancer detection rates with the use of DBT. Although these performance metrics are promising, there are limitations in these studies as noted above. Among these limitations is a lack of evidence demonstrating DBT’s impact on long-term outcomes. In addition, there are remaining questions regarding the appropriate clinical role(s) of DBT as well as which subgroup(s) of women might benefit from these examinations. Therefore, we do not recommend the coverage of DBT at this time. eviCore will continue to closely monitor the peer-reviewed literature for research that further defines the best use of DBT in clinical practice and hopefully demonstrates favorable long-term outcomes. As these issues are addressed and clarified, we will amend our recommendations accordingly.

References


