Published in final edited form as:

Acad Radiol. 2014 April ; 21(4): 445–449. doi:10.1016/j.acra.2013.10.015.

Impact of and interaction between the availability of prior examinations and DBT on the interpretation of negative and benign mammograms

Christiane M. Hakim, MD¹, Marie I. Anello, DO¹, Cathy S. Cohen, MD¹, Marie A. Ganott, MD¹, Amy H. Lu, MD¹, Ronald L. Perrin, MD¹, Ratan Shah, MD¹, Marion Lee Spangler, MD¹, Andriy I. Bandos, PhD², and David Gur, ScD³

¹ University of Pittsburgh, Magee-Womens Hospital UPMC, Department of Radiology, 300 Halket Street, Pittsburgh, PA 15213

² University of Pittsburgh, Graduate School of Public Health, Department of Biostatistics, Parran Hall, 130 DeSoto Street, Pittsburgh PA 15261; phone: 412-383-5738; fax: 412-624-2183

³ University of Pittsburgh, Radiology Imaging Research, Department of Radiology, 3362 Fifth Avenue, Pittsburgh, PA 15213

Abstract

Rationale and Objective—To assess the interaction between the availability of prior examinations and digital breast tomosynthesis (DBT) in decisions to recall a woman during interpretation of mammograms.

Materials and Methods—Eight radiologists independently interpreted twice 36 mammography examinations, each of which had current and prior full field digital mammography images (FFDM) and DBT under a HIPPA compliant IRB approved protocol (written consent waived). During the first reading three sequential ratings were provided using FFDM only followed by FFDM+ DBT followed by FFDM+DBT+priors. The second reading included FFDM only, then FFDM+priors, then FFDM+priors+DBT. Twenty two benign cases clinically recalled, 12 negative/benign examinations (not recalled) and two verified cancer cases were included. Recall recommendations and interaction between the effect of priors and DBT on decisions were assessed ($p=0.05$ significance level) using generalized linear model (proc glimmix, SAS. V.9.3., Cary, NC) accounting for case and reader variability.

Results—Average recall rates in non-cancer cases were significantly reduced (51%; $p<0.001$) with the addition of DBT and with addition of priors (23%; $p=0.01$). In absolute terms, the

© 2013 The Association of University Radiologists. Published by Elsevier Inc. All rights reserved.

Corresponding Author: Christiane M. Hakim, MD University of Pittsburgh Magee-Womens Hospital UPMC Department of Radiology 300 Halket Street Pittsburgh, PA 15213 Phone: 412.641.1965 Fax: 412.641.3452 chakim@mail.magee.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conclusion: The availability of prior examinations and DBT during the interpretations of screening mammograms, are independent, primary, contributing factors leading to a reduction in the frequency of recommendations to recall a woman for a diagnostic work up.

addition of DBT to FFDM reduced the recall rates from 0.67 to 0.42 and from 0.54 to 0.27 when DBT was available before and after priors, respectively. Recall reductions were from 0.64 to 0.54 and from 0.42 to 0.33 when priors were available before and after DBT, respectively. Regardless of the sequence in presentation, there were no statistically significant interactions between the effect of availability of DBT and priors ($p=0.80$).

Conclusion—Availability of both priors and DBT are independent, primary factors in reducing recall recommendations during mammographic interpretations.

Keywords

breast cancer; recall rates; tomosynthesis; comparison with priors

Introduction

Digital breast tomosynthesis (DBT) has been approved for clinical use with the expectation of improving the efficacy of screening and diagnostic mammography by increasing the conspicuity of imaged abnormalities [1]. With the advent of DBT, recent publications have substantiated the possible impact of this technology on the interpretation of screening mammograms in terms of reducing recall rates while simultaneously increasing cancer detection rates, in particular that of invasive cancers [2-7]. However, many studies did not include the availability of prior examinations during the interpretation; hence, potentially enhancing the effect of DBT by stressing the reference interpretation. Studies that included prior examinations when available did not report separately on the impact of DBT in the two groups, namely “with” and “without” the availability of priors. In the clinical environment the vast majority of women have been screened previously, allowing prior examinations to be available for comparison. It is well known that the availability of prior examinations during the interpretation of screening mammograms reduces recall rates substantially by enhancing the ability of the observer to assess change over time, if any, while at the same time discard depicted abnormalities that clearly represent a variety of benign (and stable) findings [8-11]. Recent prospective studies indicate that even with the availability of priors, recall rates decreased substantially when DBT is included in the examinations being interpreted [5]. However, there are no studies we are aware of aimed at assessing the relative contribution of priors and DBT on the decision whether to recall women for diagnostic workup of a suspicious findings, or not, or on the interaction, if any, between these primary factors affecting decisions to recall a woman (or not) for a diagnostic workup.

In this preliminary (pilot) study we performed a two reading mode observer study to preliminarily assess the interaction between the two primary contributing components to the reduction in recall rates in women who had undergone mammography examinations previously. By design the study used an enhanced set of cases (a stress test) for this purpose, and represents a pilot to a pivotal study we are in the process of preparing for.

Materials and Methods

A fully de-identified digital mammography image set used in this study consisted of prior and current full field digital mammography (FFDM) images and digital breast tomosynthesis

(DBT) images. The current FFDM and DBT images were acquired at our facility between 2008 and 2010 under institutional review board (IRB) approved research protocols in which consenting participants (written consent required) underwent both conventional FFDM, as part of routine screening, and DBT. All DBT images were acquired on a research system (Dimensions, Hologic Inc, Bedford MA) using standard tomosynthesis acquisition technique which generated 15 low-dose projection images, or frames, acquired for reconstruction of the 3D DBT image sets. All prior examinations were acquired on a clinical system (Selenia, Hologic Inc, Bedford MA) as a part of the standard care during previous screening examinations. After acquisition, the data from the low dose frames were used to reconstruct 50 to 90 parallel slices 1 millimeter thick, the number of which depended on the thickness of the compressed breast.

Study population

FFDM and DBT examinations performed on 36 women ranging in age from 39 to 77 years (mean age 53.5 years $\pm$ 8.5) were selected for this pilot study from a sequentially ascertained set of a total of 189 examinations acquired between September, 2008 and November, 2010. Selection was based on the availability of a prior FFDM examination, between one to three years prior to the “current” FFDM examination of interest, and a predefined set of findings as a result of the final clinical interpretation plus a follow up status verified at least 1 year post the examination in question for all negative and benign cases. The remaining cases in the set (153 cases) were sequestered for use in a future pivotal study to follow. This pilot study was specifically powered statistically to demonstrate a 20 percent difference in recall rates by either reading modes in question or the provision of either of the additional information provided during interpretation. The prior FFDM images were collected under an IRB approved research protocol (written consent waived) and ranged from 12 to 32 months prior to the current FFDM examination (mean 23.8 months $\pm$ 4.6). Clinical outcome was verified on each of the 36 assembled examinations. To stress test the experiment in terms of obtaining higher than clinically expected recall rates, twenty-four (67%) of the examinations (including the two cancer cases) were actually recalled for a diagnostic workup during the clinical practice and the remaining 12 (33%) examinations were rated as negative or benign (BI-RADS 1 or 2), hence not recalled during the original clinical interpretations. Of the 24 recalled examinations, five were biopsied and 19 were determined benign during diagnostic workup and imaging based follow-up. Two cancer cases were included to assure that readers did not rate all cases as “negative” or “benign” (i.e. screening BIRADS 1 or 2) by expectation, and radiologists were informed that not “all cases in the study are negative for the presence of cancer”. One cancer case was recalled for asymmetry and initial biopsy revealed invasive ductal carcinoma (IDC); however, after excision, two masses were described as 0.9 cm with IDC and 0.3 cm with invasive lobular carcinoma (ILC). The other cancer case was recalled for an indeterminate mass and at later verified by excisional biopsy as a 1.5 cm ILC tumor. Therefore, the dataset included two (6%) verified cancer cases, 22 (61%) verified benign cases recalled in the clinical practice, and 12 (33%) negative/benign cases not recalled. The subjectively rated breast density distribution for these cases was 1/36 (3%), 8/36 (22%), 27/36 (75%), and 0/36 (0%) for tissue density almost entirely fat, scattered fibro-glandular density, heterogeneously dense, and extremely dense, respectively. As we do not believe current DBT is optimal for examinations of women with density

BIRADS 4 that constitute but a small fraction of women in the clinical arena and in this pilot study we wanted to evaluate changes, if any, in particular in the important density BIRADS 3, no cases with BIRADS density 4 were included in the study.

Participating Radiologists

Eight board-certified, MQSA qualified radiologists (including MQSA qualification for DBT interpretations) with breast imaging experience ranging from 2 to 32 years volunteered to participate as readers in the study. The radiologists were unaware of the specific aims of the study and received an "Instruction to Observers" document defining the task at hand to review before beginning the study. The document informed the readers that this was a pilot study to assess if and how the availability of different types of imaging based information may affect decisions in a screening environment. The document also defined the type of examinations used in the study, gave the general set up and the protocol for reviewing and rating each of the examinations. It also informed readers that computer aided detection (CAD) would not be provided, and requested that if at all possible, readers should attempt to complete the readings in two sessions. No prevalence information was provided.

Observer Study

The reader study was conducted on a dual monitor Hologic SecurView DX workstation that was remotely controlled by a laptop with an in-house developed study management application. This application (Study Manager) communicated with the Hologic SecurView system using the SecurView AppSync programming interface. Study Manager was responsible for driving the study in terms of case sequencing (worklist) and was also used to record all scoring. However, all case viewing was performed on the SecurView workstation which includes two 5 megapixel LCD displays that allowed for the viewing of one, two or four images per display for each monitor (i.e., up to eight images could be displayed simultaneously). Each examination for this study consisted of two conventional views of each breast (cranio-caudal (CC) and medio-lateral oblique (MLO) projections).

The study involved two separate sequential readings where the radiologists, who are familiar with reading DBT, interpreted retrospectively a selected set of 36 examinations. For the first reading the images were displayed in the order of the "current" four view FFDM (FFDM only), followed by enabling the viewing of the DBT (FFDM+DBT), followed by enabling the viewing of priors (FFDM+DBT+priors). For the second reading the priors were offered before DBT; therefore, the second reading sequence was FFDM only, FFDM+priors, and FFDM+priors+DBT. The mode sequencing was presented in the same manner for all readers. During each reading the radiologist was asked to score/rate the case after each set of images were presented within the sequence, therefore, each reader sequentially scored each case three times under each of the two readings. Radiologists reported/scored their breast based interpretations and recommendations using a screening Breast Imaging-Reporting and Data System (BI-RADS) rating scale. The reader was prompted to provide a breast based probability of malignancy (POM) rating (0 -100) only if the breast in question was scored as a BI-BRADS 0 (recall). There was a minimum of a seven week interval (time delay) between each of the reading modes and most readers were able to complete the readings

under each mode (36 examinations) in one session. Due to clinical scheduling/coverage issues, two readers required two reading sessions to complete the reading sessions.

Data Analysis

Recall recommendations were evaluated for each reader and for different subsets of cases (i.e., verified benign recalled in the clinic, not recalled, cancer, and all non-cancer cases combined). A case was considered a recall if either breast received a BI-RADS score of 0. Between mode recall rates for the 34 cases without verified cancer were compared. Analysis of recall rates and possible interactions between prior examinations and DBT was performed using the generalized linear mixed model for binary data (PROC GLIMMIX, SAS v.9.3, SAS Institute, Cary, NC) adjusting for sources of variability due to cases and readers. Effects of availability of DBT, and priors, adjusted for possible differences between the reading sessions (order of presentation), were evaluated using Type III tests at a 0.05 significance level. Our primary analysis used case as a diagnostic unit as we see the screening interpretation as a triage segmenting women into “negative” (both breast are rated screening BIRADS 1 or 2) and “suspicious” needing a diagnostic workup (at least one breast rated screening BIRADS 0). We also repeated the analyses using breast as a diagnostic unit (rather than a woman or a case) to verify robustness of the study conclusions.

Results

During the first reading, all readers had the highest recall rate for FFDM only and had the lowest recall rate for FFDM+DBT+priors, with the exception of one reader who had their lowest recall rate with FFDM+DBT. During the second reading, all readers had the lowest recall rate with FFDM+priors+DBT and all readers had their highest recall rate with FFDM only, with the exception of one reader who had their highest recall rate with FFDM+priors (Table 1).

The combination of FFDM + DBT,+ priors had the lowest average recall rate of non-cancer cases for both readings (0.33 and 0.27, respectively, Table 1). Adjusting for availability of priors and reading session, the addition of the DBT led to a 51% reduction in the recall rate of non-cancer cases ($p<0.001$). During the first reading, the absolute unadjusted non-cancer recall rate was reduced from 0.67 (FFDM only) to 0.42 (FFDM+DBT). Similarly, during the second reading when DBT was made available after priors, the recall rate for non-cancer cases was reduced from 0.54 (FFDM+priors) to 0.27 (FFDM+priors+ DBT). Adjusting for availability of DBT and reading session, availability of priors also led to a 23% reduction in the recall rates for non-cancer cases ($p=0.01$). Recall rates were reduced from 0.42 (FFDM +DBT) to 0.33 (FFDM+DBT+Priors) when priors were made available after DBT in the first reading and from 0.64 (FFDM only) to 0.54 (FFDM+Priors) when priors were made available first (before DBT). There was no significant difference in recall rates of the non-cancer cases between the comparable modes in the first and second reading sessions ($p=0.17$). Most important, there was no statistically significant interaction between the effect of availability of DBT and availability of priors on decisions to recall a case for a diagnostic workup ($p=0.80$).

We simply note that for both readings of the two cancer cases, average POM ratings were lowest for the FFDM only mode (48%, 32% and 30%, 27%, respectively) and highest for the combination of FFDM, DBT, and priors (97%, 71% and 77%, 64%, respectively).

Repeated analyses by breast (rather than case) showed similar results and the same conclusion as the case based analyses.

Discussion

Unlike traditional technology and practice assessment studies in which one typically estimates both sensitivity and specificity, as they are inter related, the primary aim of this study was to assess a possible interaction between two primary factors affecting recall. Hence, there was no need for a study design that provided two dimensional summary measures, such as those obtained in a typical receiver operating characteristic (ROC) or free-response ROC (FROC) based investigation, thereby increasing substantially sample size and the effort to execute the study. Our findings of no significant interaction between the two primary factors, namely priors and DBT, suggest that the effect of the two types of information is largely independent of each other. In other words, prior examinations and DBT are independent primary contributing components affecting observers' decisions in negative and/or benign cases. There are no studies that have assessed in this environment the interaction, if any, between the availability of priors and DBT on decisions to recall a woman (or not) for an additional imaging based diagnostic work up. There are also no studies to date assessing the specific impact of the availability of prior examinations on DBT and/or the combination of priors and DBT on recall rates in a FFDM+DBT based screening environment. This type of an investigation will eventually lead to improved understanding as to the specific subsets of screened women who are more likely to benefit from DBT based imaging during their periodic screening (e.g. "baseline" versus "prevalence" when priors are not available, women with heterogeneously dense breast, etc.). We found in this pilot study a significant reduction in recall rates with the addition of DBT, regardless of the availability of prior images, independent of the well known and expected reduction in recall rates with addition of prior images. The effect of the addition of prior images however, was smaller in our study than that of DBT. This is probably due to the fact that in the clinic many of the cases in our dataset had been interpreted with priors; hence, there may have been a selection bias leading to the use of examinations in which priors are not as effective in further reducing recall rates. Indeed, as evident from Table 1, for the cases that were not recalled in the clinic, we observed a larger percent of reduction in the recall rates. We note that even in this set of examinations not originally recalled during clinical interpretations, the frequency of recommendations for recall was quite high. This can be possibly attributed to a combination of several factors, including but not limited to the fact that readers expected many of the cases were "difficult". Also, in reader studies observers tend to be more conservative, as patient case is not affected by their recommendations and the general natural expectation in reader studies of having somewhere close to a 50% prevalence of "actually positive" cases. Interestingly, in concordance with observations made in other studies [5], both cancer cases were rated as having cancer with substantially higher confidence levels with the availability of DBT. Our study is a pilot in preparation for a pivotal study that will include a larger sample size and a more representative set of cases.

This pilot study provides us with estimates of the effects of availability of DBT and prior information as well as between-reader variability of recall rates and the possible expected interaction between the two types of additional information. The size of the effects we observed clearly suggests that it is indeed appropriate to use a “more representative” density distribution and additional “typical” cases in our pivotal study (less of a “stress test”) without prohibitively increasing sample size requirements. This study highlights the importance of performing pilot studies, because it allows for required sample size estimation as well as in providing a better perspective as what to expect when performing a study with datasets that are more representative of the clinical practice.

As this is a pilot study we can list a number of limitations, including but not limited to the use of sequential ratings, possible enhancement of the magnitude of our observed effect due to case selection of primarily cases with density BIRADS 3 and high fraction of actually recalled examination during the clinical interpretation and the suboptimal balancing between the first and second readings. It has been shown and is largely accepted that in these types of sequential ratings based studies the impact of using a sequential reading design on inferences is minimal, if any [2]. The use of a stress test in pilot studies is important, as it generally helps in assessing an upper limit of the effect being investigated. The impact of mode balancing is important, in particular in studies aimed at assessing “between mode performance differences”, but this was not the primary aim of our study. However, these limitations cannot detract from the independent, large impact either priors or DBT, and definitely the combination of both, had on decisions made by the observers. More important perhaps in the context of this study is the consistency between the recall rates for each of the case sets during the two reading sessions for FFDM only (the beginning state) and the FFDM+priors+DBT (end state regardless of the order of presentation of priors and DBT). Together with the consistency in the results in terms of the lack of significant interaction between the impact of the availability of priors or DBT on recall rates, regardless of the order presented (made available) to the observer, our findings support the general validity of our observations.

Acknowledgments

Funding: This work is supported in parts by Grant CA163300 and CA143019 from the National Cancer Institute, National Institutes of Health to the University of Pittsburgh.

References

1. US Food and Drug Administration. Selenia Dimensions 3D Systems – P080003. Available at <http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/DeviceApprovalsandClearances/Recently-ApprovedDevices/ucm246400.htm>
2. Gur D, Abrams GS, Chough DM, et al. Digital breast tomosynthesis: Observer performance study. *Am J Roentgenol.* 2009; 193:586–91. [PubMed: 19620460]
3. Poplack SP, Tosteson TD, Kogel CA, et al. Digital breast tomosynthesis: Initial experience in 98 women with abnormal digital screening mammography. *Am J Roentgenol.* 2007; 189:616–623. [PubMed: 17715109]
4. Skaane P, Gullien R, Bjørndal H, et al. Digital breast tomosynthesis (DBT): initial experience in a clinical setting. *Acta Radiol.* Jun 1; 2012 53(5):524–9. [PubMed: 22593120]
5. Skaane P, Bandos AI, Gullien R, Eben EB, Ekseth U, Haakenaasen U, Izadi M, Jebsen IN, Jahr G, Krager M, Niklason LT, Hofvind S, Gur D. Comparison of Digital Mammography Alone and

- Digital Mammography Plus Tomosynthesis in a Population-based Screening Program. *Radiology*. 2013; 267(1):47–56. [PubMed: 23297332]
6. Rafferty EA, Park JM, Philpotts LE, Poplack SP, Sumkin JH, Halpern EF, Niklason LT. Assessing radiologist performance using combined digital mammography and breast tomosynthesis compared with digital mammography alone: results of a multicenter, multireader trial. *Radiology*. 2013; 266(1):104–13. [PubMed: 23169790]
 7. Bernardi D, Ciatto S, Pellegrini M, Anesi V, Burlon S, Cauli E, Depaoli M, Larentis L, Malesani V, Targa L, Baldo P, Houssami N. Application of breast tomosynthesis in screening: incremental effect on mammography acquisition and reading time. *Br J Radiol*. Dec; 2012 85(1020):e1174–8. [PubMed: 23175484]
 8. Sumkin JH, Holbert BL, Herrmann JS, Hakim CA, Ganott MA, Poller WR, Shah R, Hardesty LA, Gur D. Optimal reference mammography: a comparison of mammograms obtained 1 and 2 years before the present examination. *AJR Am J Roentgenol*. Feb; 2003 180(2):343–6. [PubMed: 12540430]
 9. Thurfjell MG, Vitak B, Azavedo E, Svane G, Thurfjell E. Effect on sensitivity and specificity of mammography screening with or without comparison of old mammograms. *Acta Radiol*. Jan; 2000 41(1):52–6. [PubMed: 10665871]
 10. Roelofs AA, Karssemeijer N, Wedekind N, Beck C, van Woudenberg S, Snoeren PR, Hendriks JH, Rosselli del Turco M, Bjurstam N, Junkermann H, Beijerinck D, Séradour B, Evertsz CJ. Importance of comparison of current and prior mammograms in breast cancer screening. *Radiology*. Jan; 2007 242(1):70–7. [PubMed: 17185661]
 11. Varela C, Karssemeijer N, Hendriks JH, Holland R. Use of prior mammograms in the classification of benign and malignant masses. *Eur J Radiol*. Nov; 2005 56(2):248–55. [PubMed: 15890483]

Average recall rates for different subset of cases for first and second readings.

subset	n*8	First Reading						Second Reading											
		FFDM only			FFDM+DBT			FFDM+DBT+Priors			FFDM only			FFDM+Priors			FFDM+Priors+DBT		
		Average rate	#	Recalls	Average rate	#	Recalls	Average rate	#	Recalls	Average rate	#	Recalls	Average rate	#	Recalls	Average rate	#	Recalls
Cancer	16	1.00	16	1.00	16	0.88	14	1.00	16	1.00	16	1.00	16	1.00	16	1.00	16	1.00	16
Recalled benign	176	0.70	123	0.46	81	0.39	69	0.66	117	0.60	106	0.31	55						
Not recalled	96	0.63	60	0.35	34	0.23	22	0.59	57	0.44	42	0.20	19						
All non-cancer	272	0.67	183	0.42	115	0.33	91	0.64	174	0.54	148	0.27	74						