Radiology: Imaging Cancer ORIGINAL RESEARCH

AI-enhanced Mammography With Digital Breast Tomosynthesis for Breast Cancer Detection: Clinical Value and Comparison With Human Performance

Daphne Resch, MD • Roberto Lo Gullo, MD* • Jonas Teuwen, PhD • Friedrich Semturs, PhD • Johann Hummel, PhD • Alexandra Resch, MD, PhD** • Katja Pinker, MD, PhD***

From the Department of Biomedical Imaging and Image-guided Therapy, Division of Molecular and Gender Imaging, Medical University of Vienna, Austria (D.R.); Department of Radiology, Breast Imaging Service, Memorial Sloan-Kettering Cancer Center, New York, NY (R.L.G., J.T.); Center for Medical Physics and Biomedical Engineering, Medical University Vienna, Vienna, Austria (F.S., J.H.); St Francis Hospital Vienna, Vienna, Austria (A.R.); Sigmund Freud University Medical School, Vienna, Austria (A.R.); and Department of Radiology, Division of Breast Imaging, Columbia University Irving Medical Center, 161 Fort Washington Ave, New York, NY 10032 (K.P.). Received August 25, 2023; revision requested October 22; final revision received April 23, 2024; accepted May 30. **Address correspondence to** K.P. (email: *kp3124@cumc.columbia.edu*).

Supported in part by the National Institutes of Health/National Cancer Institute Cancer Center Support Grant (no. P30 CA008748).

* D.R. and R.L.G. contributed equally to this work.

** A.R. and K.P. are co–senior authors.

Conflicts of interest are listed at the end of this article.

See also commentary by Kataoka and Uematsu in this issue.

Radiology: Imaging Cancer 2024; 6(4):e230149 • **https://doi.org/10.1148/rycan.230149** • **Content codes:** BR AI

Purpose: To compare two deep learning–based commercially available artificial intelligence (AI) systems for mammography with digital breast tomosynthesis (DBT) and benchmark them against the performance of radiologists.

Materials and Methods: This retrospective study included consecutive asymptomatic patients who underwent mammography with DBT (2019–2020). Two AI systems (Transpara 1.7.0 and ProFound AI 3.0) were used to evaluate the DBT examinations. The systems were compared using receiver operating characteristic (ROC) analysis to calculate the area under the ROC curve (AUC) for detecting malignancy overall and within subgroups based on mammographic breast density. Breast Imaging Reporting and Data System results obtained from standard-of-care human double-reading were compared against AI results with use of the DeLong test.

Results: Of 419 female patients (median age, 60 years [IQR, 52–70 years]) included, 58 had histologically proven breast cancer. The AUC was 0.86 (95% CI: 0.85, 0.91), 0.93 (95% CI: 0.90, 0.95), and 0.98 (95% CI: 0.96, 0.99) for Transpara, ProFound AI, and human double-reading, respectively. For Transpara, a rule-out criterion of score 7 or lower yielded 100% (95% CI: 94.2, 100.0) sensitivity and 60.9% (95% CI: 55.7, 66.0) specificity. The rule-in criterion of higher than score 9 yielded 96.6% sensitivity (95% CI: 88.1, 99.6) and 78.1% specificity (95% CI: 73.8, 82.5). For ProFound AI, a rule-out criterion of lower than score 51 yielded 100% sensitivity (95% CI: 93.8, 100) and 67.0% specificity (95% CI: 62.2, 72.1). The rule-in criterion of higher than score 69 yielded 93.1% (95% CI: 83.3, 98.1) sensitivity and 82.0% (95% CI: 77.9, 86.1) specificity.

Conclusion: Both AI systems showed high performance in breast cancer detection but lower performance compared with human double-reading.

©RSNA, 2024

Breast cancer is a major public health concern and is the leading cause of cancer in women worldwide. Presently, the estimated lifetime risk of developing breast cancer is 12.8%, affecting approximately one in eight women (1,2). In health care, the implementation of artificial intelligence (AI) is gaining momentum, promising optimized workflow efficiency, increased productivity, and improved patient outcomes. There is particular interest in using AI within the field of breast imaging, which is facing increased pressure due to the high prevalence of breast cancer and exponential growth in requested breast imaging services, alongside a concurrent workforce shortage (3).

Several institutions have attempted to implement conventional radiomics or machine learning–based computer-aided detection (CAD) systems; however, these systems were not satisfactory, as they yielded a high number of false-positive findings (4–6). More recently (7), deep learning (DL)–based CAD systems in place of conventional CAD systems have demonstrated improved diagnostic accuracy for breast cancer. Additionally, diagnostic performance improved among radiologists who used DL-based AI CAD systems (8,9), supporting their widespread adoption. For both breast screening and diagnostic breast imaging, digital breast tomosynthesis (DBT) (often referred to as three-dimensional mammography) is now widely implemented. The implementation of DL-based AI for DBT, which has larger data volumes compared with traditional full-field digital mammography (often referred to as two-dimensional [2D] mammography), has been shown to improve both diagnostic accuracy and the clinical workflow (10–13).

Despite the demonstrated potential of AI-enhanced mammography, including AI-enhanced DBT, there is a lack of studies directly comparing different commercially available AI algorithms for DBT. Such studies are

This copy is for personal use only. To order printed copies, contact reprints@rsna.org

Abbreviations

ACR = American College of Radiology, AI = artificial intelligence, AUC = area under the ROC curve, BI-RADS = Breast Imaging Reporting and Data System, CAD = computer-aided detection, DBT = digital breast tomosynthesis, DL = deep learning, ROC = receiver operating characteristic, 2D = two-dimensional

Summary

Two artificial intelligence systems for mammography with digital breast tomosynthesis demonstrated high performance in detecting malignancies, although performance was lower when compared against human double-reading.

Key Points

- In asymptomatic patients undergoing mammography with digital breast tomosynthesis, two artificial intelligence (AI) systems, Transpara and ProFound AI, had overall areas under the receiver operating characteristic curve (AUCs) of 0.86 (95% CI: 0.85, 0.91) and 0.93 (95% CI: 0.90, 0.95), respectively, for detecting malignancies, while human double-reading had an overall AUC of 0.98 (95% CI: 0.96, 0.99).
- ProFound AI performed significantly better than Transpara for breast cancer detection, particularly in patients with nondense breasts (AUC, 0.93; $P < .001$).

Keywords

Mammography, Breast, Oncology, Artificial Intelligence, Deep Learning, Digital Breast Tomosynthesis

urgently needed to shed light on the advantages and disadvantages of implementation of these algorithms in clinical practice. Thus, the present exploratory study aimed to compare the intraindividual diagnostic performance of two DL-based commercially available AI CAD systems for mammography with DBT and benchmark them against the performance of radiologists. We also sought to determine the differences between the two different DL systems and their benefits based on our results and in conjunction with the literature.

Materials and Methods

Study Sample

This study was approved by the institutional review board and performed in accordance with the Declaration of Helsinki. The need for informed consent was waived by the institutional review board.

The database of the Radiologicum Margareten (Department of Radiology) at St Francis Hospital, Vienna, Austria, was searched for consecutive asymptomatic patients who underwent breast imaging in 2019 and 2020 for either screening or diagnostic purposes, resulting in 5421 patients. Of note, patients who underwent breast imaging for diagnostic purposes were asymptomatic patients with a personal history of breast cancer or prior biopsy yielding benign high-risk results; in Austria, these patients often do not undergo formal screening but instead undergo diagnostic mammograms for surveillance despite being asymptomatic. Whether for screening or diagnostic purposes, the standard screening imaging protocol with DBT was used. No patients presenting with clinical concerns were included in this study.

Regarding the inclusion and exclusion of patients from the study, of the 5421 patients, 139 patients had imaging findings classified according to the American College of Radiology (ACR) Breast Imaging Reporting and Data System (BI-RADS) as BI-RADS 4 or 5 findings. However, for 25 of 139 patients, BI-RADS 4 or 5 classification was made based on US findings rather than DBT findings; thus, these patients were excluded from the study. Of note, the Austrian Breast Cancer Screening Program mandates US in women with ACR mammographic breast density categories C and D. The study sample was randomly matched at a ratio of 1:3 (patients with higher BI-RADS categories of 4 or 5 matched to those with lower BI-RADS categories of 1 or 2). The variables used for matching were age decade, breast density, and time period of the imaging examination. This resulted 456 matched patients, with 114 patients with BI-RADS 4 or 5 findings and 342 patients with BI-RADS 1 or 2 findings. Patients with BI-RADS 3 findings were excluded from our study for the following reasons specific to the Austrian Breast Cancer Screening Program. In Austria, assigning a BI-RADS 3 category during screening is discouraged, as imaging interpretation is performed while the patient is present, and any detected imaging abnormalities during screening would be immediately discussed with a second reader and addressed on the spot with mammography and/or US to minimize patient anxiety. Further, in cases of asymmetries without US correlation, MRI is often used to assign a final category other than BI-RADS 3. Moreover, the inclusion of patients with BI-RADS 3 findings would have required a 2-year follow-up period to confirm benignity and reclassification as BI-RADS 2, followed by another 2-year follow-up (the screening interval in Austria), delaying the analysis and publication of study results without substantial contribution to them. As expected, the number of patients with true BI-RADS 3 findings from screening mammography during the study period was very low (<3%), leading to their exclusion from the study.

Several further exclusions were subsequently performed. First, three patients for whom either AI system could not compute a score for their DBT examination (for reasons unknown to us) were excluded. Second, of the patients with BI-RADS 4 or 5 findings, two patients were excluded for the sake of comparability, as one was diagnosed with Merkel cell carcinoma and the other was diagnosed with lymph node metastasis from lung cancer. Third, of the patients with BI-RADS 1 or 2 findings, 28 who did not have reference standard 2-year follow-up imaging results were excluded. To evaluate the AI systems in patients with both higher and lower BI-RADS categories, four symptomatic patients were also excluded from the patient sample. The final study sample comprised 419 patients. Figure 1 depicts the flow of patient inclusion in the study.

The reference standard for the AI systems were histologic results from subsequent biopsy in the case of patients with BI-RADS 4 and 5 findings and 2-year follow-up imaging in patients with BI-RADS 1 and 2 findings; in other words, the reference standard was established by radiologist

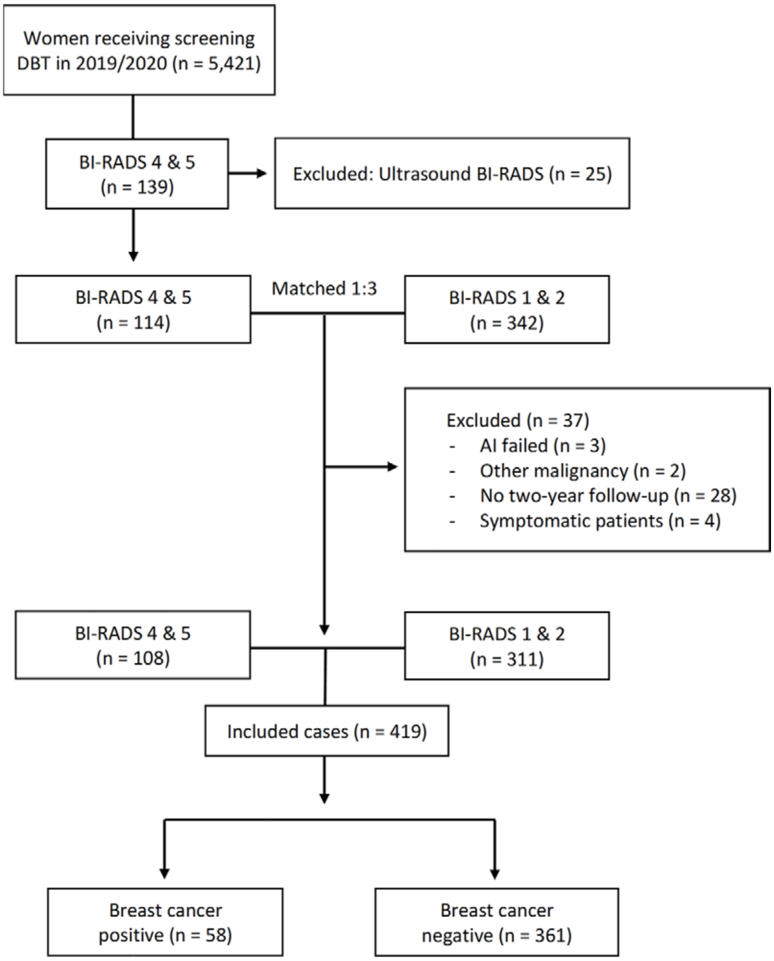
Women receiving screening DBT in 2019/2020 (n = 5,421)
BI-RADS 4 & 5 (n = 139)
Excluded: Ultrasound BI-RADS (n = 25)
BI-RADS 4 & 5 (n = 114)
Matched 1:3
BI-RADS 1 & 2 (n = 342)
Excluded (n = 37)
- AI failed (n = 3)
- Other malignancy (n = 2)
- No two-year follow-up (n = 28)
- Symptomatic patients (n = 4)
BI-RADS 4 & 5 (n = 108)
BI-RADS 1 & 2 (n = 311)
Included cases (n = 419)
Breast cancer positive (n = 58)
Breast cancer negative (n = 361)

Figure 1: Flow diagram depicts patient inclusion. AI = artificial intelligence, BI-RADS = Breast Imaging Reporting and Data System, DBT = digital breast tomosynthesis.

double-reading (the standard of care in Europe for breast cancer screening). After the exclusion of two patients (ie, one with Merkel cell carcinoma and one with lymph node metastasis from lung cancer) for the sake of comparability from the matched patient group with BI-RADS 4 and 5 findings, 58 patients with histologically proven breast cancer remained, including nine patients with in situ carcinoma and 49 with invasive components. Another 39 patients had benign histologic findings, and another 11 patients refused biopsy and underwent 2-year MRI follow-up with no evidence of malignancy; findings in these patients were classified as benign.

Imaging Protocol and Clinical Imaging Evaluation

All DBT examinations were performed with a Hologic 3Dimensions mammography system. All examinations comprised four standard views (ie, bilateral craniocaudal and mediolateral oblique views). Clinical imaging evaluation entailed a double-reading system, whereby initial reads were performed by two board-certified physicians specialized in breast imaging with at least 10 years of experience (A.R. and another physician) who were blinded to each other's decisions. Both readers had access to patient records while performing their assessment, including breast cancer history and images from prior imaging examinations. If one or both readers assigned a BI-RADS category higher than 2, a consensus conference was held whereby a group of readers guided by a leading physician reconciled differences in interpretation.

Imaging Evaluation by DL Systems

Two commercially available AI systems were used to evaluate mammography with DBT. These systems were chosen because they were the only AI systems for DBT that had already been approved for clinical use in Europe (ie, both had already received a European Commission mark) and the United States (ie, both had already received approval from the U.S. Food and Drug Administration) at the time of the study.

The first system, Transpara 1.7.0 (ScreenPoint Medical), is a Food and Drug Administration–approved commercial system for the automated interpretation of both full-field digital mammography and DBT. Its architecture is based on deep convolutional neural networks. A previous version of Transpara, Transpara 1.4.0, was trained and validated on over 9000 images with cancers as well as 180 000 images without abnormalities (14). Mammograms with DBT used for training and validating Transpara 1.4.0 were obtained from several institutions across the United States, Asia, and Europe and were acquired using platforms from four vendors (Siemens, Hologic, General Electric, and Philips) (14). To date, Transpara has been trained on over a million mammograms (15). When reading a mammogram, Transpara assigns an examination score from 1 to 10, reflecting the level of suspicion that a cancer is present. There are three score categories that correspond to the level of cancer risk: scores 1–7, 8–9, and 10 correspond to low, intermediate, and elevated risk, respectively. According to ScreenPoint Medical, examinations in the lowest category have a 99.97% negative predictive value (16). Examinations graded as 8–9 have a similar risk to the average patient in the screening population. Examinations with a score of 10 include over 85% of screening-detected cancers and over 30% of interval cancers. Examinations in this highest category are associated with an 8.5 times higher risk of screening-detected cancer compared with lower categories. For this study, we used a vendor-recommended rule-out criterion of score 7 or lower because it showed a sensitivity of 100% and a rule-in criterion of higher than score 9 because according to ScreenPoint Medical, this category includes the highest probability of detecting cancer. Transpara assigns regional scores for each detection and overall examination scores that comprise all regional scores. In this study, only examination scores were used. Furthermore, Transpara offers different marks for calcifications and masses. The number of detection marks is limited to two per image for each category. Transpara does not access existing information about a patient, such as prior mammograms (14,17).

The second system, ProFound AI 3.0 (iCAD), is also based on deep convolutional neural networks. The system was trained on 30 000 cases, including 8000 biopsy-proven cancers, and subsequently validated on 3500 cases, including 1200 biopsy-proven cancers, altogether involving approximately 6 million images (18). The ProFound AI algorithm analyzes each individual image or section and identifies potentially malignant lesions. Each examination is given a lesion-specific score as well as an overall case score. In this study, the case scores provided by ProFound AI were collected and compared with Transpara examination scores. Lesion-specific scores range from 0 to 100 and impact the algorithm's confidence in the malignancy of a case, with higher scores indicating higher confidence. Case scores combine information from all detections within a case (19) and represent how confident the algorithm is that a case is malignant. These scores are divided into three categories representing different levels of cancer prevalence: 0–29, 30–69, and 70–100. According to iCAD, scores from 0–29, 30–69, and 70–100 correspond to 0.37, 1.54, and 4.1 cancers per 1000 patients, respectively (unpublished presentation by iCAD). ProFound AI also allows users to adjust the threshold for lesion detection based on the desired sensitivity and specificity level. Depending on the sensitivity level, the system displays more marks, but the overall case score remains the same across all sensitivity levels. For this study, the sensitivity level was set to medium, as recommended by iCAD and which is most commonly used in the clinical setting; this sensitivity level corresponds to 92% sensitivity and 59% specificity according to iCAD. Unlike Transpara, the number of marks per image is not limited. Like Transpara, ProFound AI does not access existing information about a patient.

Statistical Analysis

The following variables were extracted for every patient: age, BI-RADS category, Transpara examination score, ProFound AI case score, ACR breast density category, and biopsy result if available. Statistical analysis was conducted using Microsoft Excel (version 16.82) and MedCalc (version 20.218). AI system performance, whether for Transpara or ProFound AI, was assessed using receiver operating characteristic (ROC) analysis; the area under the ROC curve (AUC) was obtained for the overall patient sample as well as for different patient subgroups based on mammographic breast density. The DeLong test was used to calculate the AUC as well as compare AUCs between Transpara, ProFound AI, and human double-reading. Before ROC analysis, the outputs of each AI system were dichotomized to enable comparison with the recorded binary decisions of human double-reading. Cases were divided into two categories: 0 denoting benign cases and 1 indicating malignant cases. Category 0 encompassed all BI-RADS 1, BI-RADS 2, as well as histopathologically based B1, B2, and B3 cases. Category 1 consisted of histopathologically based B5 results. These classifications were then compared with the risk assessments of the AI systems. Operating points at high sensitivity (rule-out criterion at a negative likelihood ratio ≤0.1) and high specificity (rule-in criterion at a positive likelihood ratio ≥4) were also determined. AI system performance was also benchmarked against human double-reading BI-RADS results according to the AUC. A prior sample size estimate indicated that a balanced sample of 212 studies was required to detect a 10% AUC difference (0.9 vs 0.8). $P < .05$ was considered indicative of a statistically significant difference.

Results

Sample Characteristics

The final study sample comprised 419 patients (median age, 60 years [IQR, 52–70 years]; all female). Table 1 summarizes the patient and lesion characteristics of the study sample. There were 27 patients with BI-RADS 1 findings, 284 with BI-RADS 2 findings, 68 with BI-RADS 4 findings, and 40 with BI-RADS 5 findings. A total of 58 of 419 patients (14%) (median age, 60 years [IQR, 52–70 years]) had biopsy-proven cancer. Biopsy-proven cancers were most frequent across ACR B and C categories. Slightly over 5% of all cancers were found in extremely dense breasts (ACR D), even though 12% of all investigated breasts were classified as ACR D.

AI System Results

Table 2 illustrates the distribution of malignant findings compared with the total number of findings within each examination score category for the two AI systems. Transpara assigned an examination score of 10 (ie, the category representing elevated risk) to 56 of 58 cancers (97%) and an

Table 1: Patient and Lesion Characteristics

Characteristic	Value
Patient characteristics (*n* = 419)	
Age (y)*	60 (52–70)
Biopsy-proven cancer lesion	
Yes	58 (14)
No	361 (86)
BI-RADS category	
BI-RADS 1	27 (6)
BI-RADS 2	284 (68)
BI-RADS 4	68 (16)
BI-RADS 5	40 (10)
ACR breast density category	
ACR A	84 (20)
ACR B	154 (37)
ACR C	130 (31)
ACR D	51 (12)
Lesion characteristics (*n* = 58)	
ACR A	12 (21)
ACR B	24 (41)
ACR C	19 (33)
ACR D	3 (5)

Note.—Unless otherwise noted, data are numbers of patients or lesions, with percentages in parentheses. ACR = American College of Radiology, BI-RADS = Breast Imaging and Reporting Data System.

* Age is reported as the median, with IQR in parentheses.

examination score of 8–9 (ie, the category representing intermediate risk) to two of 58 cancers (3%). No cancer was assigned a score below 8. ProFound AI assigned the highest score category (70–100), indicating the highest confidence of malignancy, to 54 of 58 cancers (93%) and the middle score category (30–69) to four of 58 cancers (3%). No cancer was assigned a score of 29 or below. Scores from ProFound AI and Transpara are also shown in a scatterplot (Fig 2).

For the four cancers that were scored below 70 by ProFound AI, histologic reports were searched to gather more information. Two were ductal carcinomas in situ with a size less than 1 cm, and two invasive breast cancers were circumscribed round masses measuring 5 and 6 mm. With respect to false-positive findings, Transpara assigned a score of 10 to 78 of 361 benign cases (22%), and ProFound AI assigned a score of 70–100 to 64 of 361 benign cases (18%) that all had either some BI-RADS features suspicious for malignancy, such as architectural distortions in postsurgical patients and microcalcification, or were benign masses such as fibroadenomas.

ROC Analysis Comparing Transpara, ProFound AI, and Human Double-Reading

Figure 3 illustrates the respective ROC curves of the two AI systems and human double-reading. Table 3 shows the AUCs of the two AI systems as well as human double-reading for cancer detection, both overall and within subgroups based on mammographic density, and Table 4 shows pairwise comparisons of the ROC curves between ProFound AI and Transpara and between each AI system and human double-reading.

Overall, Transpara had an AUC of 0.86 (95% CI: 0.85, 0.91), ProFound AI had an AUC of 0.93 (95% CI: 0.90, 0.95), and human double-reading had an AUC of 0.98 (95% CI: 0.96, 0.99). ProFound AI yielded better performance compared with Transpara (difference between AUCs, 0.04; *P* = .004).

Within the subgroup of 238 patients with nondense breasts (ACR A or B), ProFound AI performed significantly better than Transpara (*P* < .001), and human double-reading performed better than either AI system, with an AUC of 0.99 (95% CI: 0.97, >0.99) versus 0.93 (95% CI: 0.88, 0.96; *P* = .003) and 0.86 (95% CI: 0.80, 0.90; *P* < .001), respectively. In the subgroup of 181 patients with dense breasts (ACR C or D), there was no evidence of a difference in performance between the two systems (*P* = .97), nor with the human readers. Both Transpara and ProFound AI had an AUC of 0.92 (95% CI: 0.88, 0.96) versus 0.96 (95% CI: 0.92, 0.98) for human double-reading. Table 5 shows results based on various operating points chosen for either high sensitivity or high specificity. For Transpara, a rule out-criterion of score 7 or lower yielded 100% (95% CI: 93.8, 100.0) sensitivity and 60.9% (95% CI: 55.7, 66.0) specificity at a negative likelihood ratio of 0. The rule-in criterion of higher than score 9 yielded 96.6% sensitivity (95% CI: 88.1, 99.6) and 78.1% specificity (95% CI: 73.8, 82.5) at a positive likelihood ratio of 4.47 (95% CI: 3.65, 5.47). For ProFound AI, a rule-out criterion of below score 51 yielded 100% (95% CI: 93.8, 100) sensitivity and 67.0% (95% CI: 62.2, 72.1) specificity at a negative likelihood ratio of 0.00. The rule-in criterion of higher than score 69 for ProFound AI yielded 93.1% sensitivity (95% CI: 83.3, 98.1) and 82.0% specificity (95% CI: 77.9, 86.1) at a positive likelihood ratio of 5.25 (95% CI: 4.16, 6.63).

Discussion

The aim of the present exploratory study was to compare the performance of two commercially available and clinically approved (in Europe and the United States) AI systems for mammography with DBT with one another and with human performance in the real-life clinical setting of asymptomatic patients undergoing mammography with DBT as part of the Austrian Breast Cancer Screening Program. In this clinical setting, based on standard-of-care human double-reading, biopsy or imaging follow-up was subsequently performed, the results of which were used as the reference standard to evaluate the performance of the AI systems. Results for the human double-reading were well within national screening benchmarks (20) and still outperformed both AI systems. Based on this reference standard, ProFound AI outperformed Transpara (overall AUC, 0.93 and 0.86, respectively; *P* = .004), particularly in patients with nondense breasts (AUC, 0.93 and 0.86, respectively). Importantly, neither AI system missed any cancer when assigning a low-risk category.

Table 2: Distribution of Malignant Findings Compared With the Total Number of Findings Within Each Score Category of the Two AI Systems for Mammography with Digital Breast Tomosynthesis

System and Score Category	Malignant Findings (*n* = 58)	Total Findings
Transpara examination score categories		
0–7	0 (0)	173
8–9	2 (3)	112
10	56 (97)	134
ProFound AI case score categories		
0–29	0 (0)	153
30–69	4 (7)	148
70–100	54 (93)	118

Note.—Data are numbers of findings, with percentages in parentheses. AI = artificial intelligence.

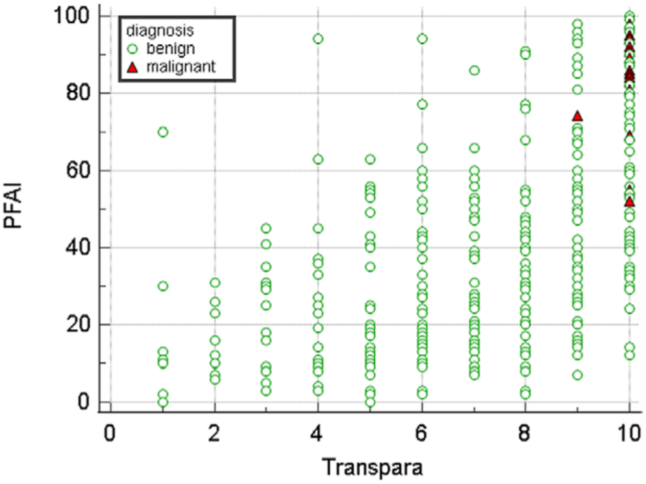

Figure 2: Scatterplot denotes malignancy scores from ProFound AI (PFAI) and Transpara systems. For the Transpara system, three score categories correspond to the level of cancer risk: scores 1–7, 8–9, and 10 correspond to low, intermediate, and elevated risk, respectively (x-axis). ProFound AI scores are divided into three categories representing different levels of cancer prevalence, from lowest to highest: 0–29, 30–69, and 70–100 (y-axis).

Transpara assigned a score of 10 to 78 of 361 benign cases (22%), and ProFound AI assigned a score of 70–100 to 64 of 361 benign cases (18%).

The false-positive findings observed in this study may be due to our mixed study sample, which included many patients who had previously undergone surgery, including lumpectomy, excisional biopsy, or mastopexy or reduction mammoplasty. Of note, neither AI system had the ability to access prior imaging, which would have been beneficial to reduce the number of both false-negative and false-positive findings. According to Olivotto et al (21), the misinterpretation of cancers with somewhat benign features could lead to progression of tumor size and delayed diagnosis, whereas it is not clear if "suspicion bias" has an influence on prognosis.

Our findings are in agreement with the prior literature. Of the studies investigating the performance of Transpara for DBT in particular, Romero-Martín et al (10) conducted a retrospective study involving 15 999 mammograms (both DBT and 2D digital mammograms) to investigate the standalone performance of Transpara 1.7.0, the same version of Transpara that was used in our study. For DBT, Transpara achieved noninferior sensitivity compared with radiologists but at the cost of an increased recall rate of up to 12%. Its performance was better for 2D digital mammograms, where it reduced the recall rate while maintaining noninferior sensitivity compared with radiologists. A study by Raya-Povedano et al (12) evaluated the ability of Transpara to reduce the radiologist workload while maintaining sensitivity in 15 987 DBT examinations and 2D digital mammograms. Transpara was used as a preselection tool to determine whether examinations were likely normal (all examinations graded with a score ≤7) and therefore did not require reading by radiologists. The results showed that screening strategies based on AI systems can reduce the radiologist workload by up to 70% without reducing sensitivity by more than 5%. Elsewhere, van Winkel et al (22) investigated the performance of radiologists reading an enriched cancer data set of 240 bilateral DBT examinations, with and without the support of Transpara. Transpara increased the radiologists' AUC while also reducing their reading time. It was also shown that DBT as a stand-alone tool was noninferior compared with radiologists. Last, Lång et al (23) conducted a study to evaluate the ability of Transpara to identify normal mammograms, showing that Transpara correctly identified normal mammograms and reduced false-positive findings. While we were unable to assess reductions in workload due to our study design, where BI-RADS 4 and 5 cases were matched to BI-RADS 1 and 2 cases, our study confirmed that the diagnostic performance of Transpara in a real-life clinical setting matches the assertion by the vendor, ScreenPoint, that 99.97% of examinations are negative in the lowest category of 1–7 and that category 10 includes over 85% of screening-detected cancers (16). Based on its diagnostic performance alone, Transpara is indeed a valuable application and relevant for population-based screening.

Regarding ProFound AI, our results confirm the assertion by its vendor, iCAD, that the highest prevalence of cancers occurs in the score category 70–100 (unpublished presentation by iCAD). Published original studies investigating the performance of ProFound AI for DBT in particular, however, are lacking. One retrospective study by Graewingholt and Duffy (24) in 35 000 women showed that single-reading of 2D digital mammography with ProFound AI was noninferior compared with the current clinical screening system of double-reading of 2D digital mammography. Another study by Conant et al (19) showed that reading times were significantly reduced, while sensitivity, specificity, recall rate, and AUC improved in their nonclinical reader study when ProFound AI was used alongside reader interpretation of DBT.

Using a different AI system from either Transpara or ProFound AI, Benedikt et al (25) demonstrated in a reader study that concurrent reading with CAD enhancement by PowerLook Tomo Detection (iCAD) resulted in 23.5% faster reading time and noninferior performance compared with radiologist interpretation. Similar results were achieved by Balleyguier et al (26). Regarding breast density, the distribution of ACR breast density categories among patients in our study sample was similar to that in the wider screening population. In the subgroup analysis in our study, ProFound AI significantly outperformed Transpara in nondense breasts (AUC, 0.93 vs 0.86, respectively; $P < .001$), whereas we found no evidence of a difference in performance between the two systems in dense breasts (AUC, 0.92 vs 0.92, respectively; $P = .97$). Interestingly, only a little over 5% of cancers in our study were detected in extremely dense breasts (ACR D); this is an underrepresentation considering that 12% of the entire patient sample had extremely dense breasts, which is considered an independent risk factor for developing breast cancer (27). This may be because malignant lesions in extremely dense breasts are more challenging to detect at mammography, regardless of whether it is 2D digital mammography or DBT, due to overlying fibroglandular tissue (28). Transpara assigned a score of 10 to only four cancers in the ACR D category (of 134 total cases, whether benign or malignant, that received a score of 10). Meanwhile, ProFound AI assigned a score of 70–100 to only eight patients in the ACR D category (of a total of 118 cases, whether benign or malignant, that received a score of 70–100). These low numbers confirm the poor visibility of cancers in dense tissue. Additionally, as mentioned, symptomatic patients are examined with mandatory US in Austria. Indeed, our study excluded 25 patients whose BI-RADS 4 or 5 classification was based on US findings only; in these patients, another five cancers were found, but those were excluded from further analysis.

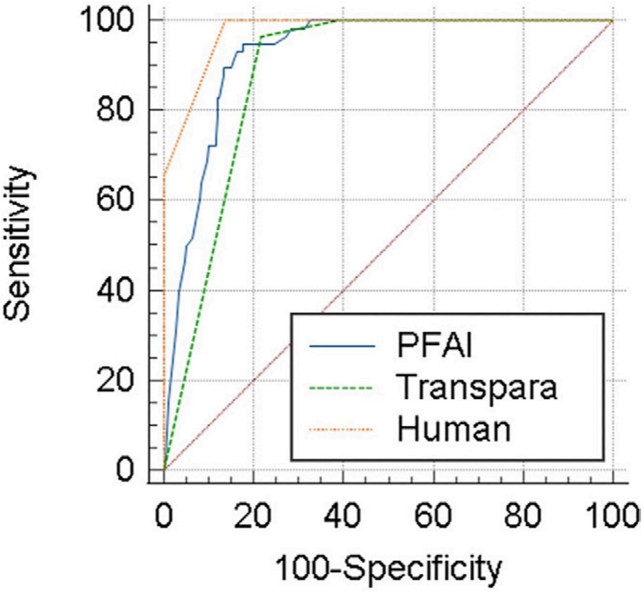

Figure 3: Comparison of the receiver operating characteristic curves for ProFound AI (PFAI), Transpara, and human double-reading. The solid red line represents the reference line.

Despite the previous publications and promising results in this field, there are still unanswered questions that limit use of AI in the daily clinical practice of breast radiologists. For algorithms to meet high performance expectations, a large amount of breast imaging data, ideally diverse, is required. To ensure the generalizability of the algorithm, it is important that socioeconomically disadvantaged groups are represented in both the training and validation sets, as these patients are often underdiagnosed. However, obtaining high-quality reference standard data is laborious and costly (29). To date, most studies investigating AI in breast imaging are either retrospective trials or small reader studies. Yet, it is essential that large multi-institutional prospective trials assess whether AI tools will perform as expected in everyday clinical practice (30). Practitioners and academic publishers have also demanded greater transparency and a better understanding of the working principle of AI tools. The so-called black box problem lies in the fact that crucial decisions are made while the decision-making process remains completely opaque (29). This lack of traceability also means that human errors or biases incorporated into the programming of the system have an unnoticed negative impact on the final result (31). The question of legal responsibility for the decisions made, such as when an AI system misses a cancer, remains controversial (29). In any case, it is crucial that regulatory frameworks for AI quality control are established (31).

Table 3: AUCs for Both AI Systems and Human Double-Reading

Study Sample (*n* = 419)	Transpara	ProFound AI	Human Double-Reading
Overall	0.86 (0.85, 0.91)	0.93 (0.90, 0.95)	0.98 (0.96, 0.99)
By mammographic density			
Nondense breasts (ACR A or B)	0.86 (0.80, 0.90)	0.93 (0.88, 0.96)	0.99 (0.97, 0.99)
Dense breasts (ACR C or D)	0.92 (0.88, 0.96)	0.92 (0.88, 0.96)	0.96 (0.92, 0.98)

Note.—Data are areas under the receiver operating characteristic curve (AUCs), with 95% CIs in parentheses. ACR = American College of Radiology, AI = artificial intelligence.

Although this was not investigated in the current study, we hypothesize that the two particular AI systems used in this study may differ in clinical applicability. Transpara is designed to function independently as a reader in a double-reading screening scenario (mandated in Europe), whereby in low-risk cases, it can fully replace one human reader, and in high-risk cases, it can be used in addition to the standard double-reading approach as an assistant. Unlike Transpara, ProFound AI primarily functions as an assistant to the human reader rather than functioning independently as a reader; in this way, in a single-reading screening scenario, ProFound AI supports the human reader to achieve comparable performance to that of human double-reading (24).

There were limitations to our study. First, the study was a retrospective study. Second, all mammograms were acquired from a single site, and DBT was performed using a platform by a single vendor, posing a risk for selection bias. Third, the matching performed in this study was suboptimal and could introduce selection bias. It should also be noted that BI-RADS 3 cases were not included in the analysis, which also poses a selection bias. The performance of the AI systems may differ in studies with BI-RADS 3 cases in the study sample. Fourth, the study sample was small and possibly underpowered (power was based on an AUC difference of 0.1), limiting the conclusions that can be drawn; thus, larger-scale studies will be necessary to corroborate our findings.

In conclusion, the results of the present study evaluating the performance of two AI systems for mammography with DBT, both of which are approved for clinical use in the United States and Europe, show that both systems can detect malignancies without missing any cancer when assigning a low-risk category. Our results showed that standard-of-care human double-reading still outperformed both AI systems. Nevertheless, both AI systems exhibit promising outcomes in terms of cancer detection and effective classification of negative examinations. Notably, none of the examinations identified as low risk by the AI systems revealed the presence of breast cancer. In conjunction with the existing literature, the two systems may differ in clinical applicability. Given these promising results, further prospective studies with larger sample sizes are needed to define the future role and mode of implementation of AI in breast cancer screening.

Table 4: Pairwise Comparison of Receiver Operating Characteristic Curves

Comparison and Parameter	Value
ProFound AI vs Transpara	
Difference between areas	0.04
Standard error*	0.01
95% CI of the difference between areas	0.01, 0.07
z statistic	2.88
P value	.004
ProFound AI vs human double-reading	
Difference between areas	0.05
Standard error*	0.01
95% CI of the difference between areas	0.02, 0.08
z statistic	3.69
P value	<.001
Transpara vs human double-reading	
Difference between areas	0.09
Standard error*	0.01
95% CI of the difference between areas	0.07, 0.11
z statistic	7.23
P value	<.001

* Standard error was calculated using the DeLong method.

Acknowledgments: The authors thank Pascal Baltzer, MD, for his help with statistical software and general guidance, and Joanne Chin, MFA, ELS, for her help in editing this manuscript.

Author contributions: Guarantors of integrity of entire study, **D.R., R.L.G., A.R., K.P.;** study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **D.R., R.L.G., F.S., J.H., A.R., K.P.;** clinical studies, **R.L.G., F.S., K.P.;** statistical analysis, **D.R., F.S., J.H.;** and manuscript editing, **R.L.G., J.T., F.S., A.R., K.P.**

Data sharing: Data generated or analyzed during the study are available from the corresponding author by request.

Disclosures of conflicts of interest: **D.R.** No relevant relationships. **R.L.G.** No relevant relationships. **J.T.** Consulting fees from ScreenPoint Medical. **F.S.** No relevant relationships. **J.H.** No relevant relationships. **A.R.** No relevant relationships. **K.P.** Grants from the Research and Innovation Framework Programme, H2020 FETOPEN, Jubiläumsfonds of the Austrian National Bank, The Vienna Science and Technology Fund, Breast Cancer Research Foundation, and National Institutes of Health; recipient of the MSKCC 2020 Molecularly Targeted Intra-Operative Imaging Award; institution has financial interests in Grail; consulting fees from or rela-

Table 5: Operating Points at High Sensitivity and High Specificity for Both AI Systems

AI System and Score Cutoff	Sensitivity (%)	Specificity (%)	Positive LR	Negative LR
Transpara				
≤7	100 (93.8, 100) [58/58]	60.9 (55.7, 66.0) [220/361]	2.56 (2.25, 2.91)	0.00
>9	96.6 (88.1, 99.6) [56/58]	78.1 (73.8, 82.5) [282/361]	4.47 (3.65, 5.47)	0.04 (0.01, 0.17)
ProFound AI				
≤51	100 (93.8, 100) [58/58]	67.0 (62.2, 72.1) [242/361]	3.06 (2.64, 3.55)	0.00
>69	93.1 (83.3, 98.1) [54/58]	82.0 (77.9, 86.1) [296/361]	5.25 (4.16, 6.63)	0.08 (0.03, 0.22)

Note.—Data in parentheses are 95% CIs, with numbers of patients in brackets. The cutoff values are recommended by the vendor to be used in the clinical setting. AI = artificial intelligence, LR = likelihood ratio.

tionship with Genentech, Merantix Healthcare, AURA Health Technologies, Guerbet, and NeoDynamics; payment for lectures from the European Society of Breast Imaging, Bayer, Siemens Healthineers, IDKD, Olea Medical, and Roche; support for travel or attending meeting from the European Society of Breast Imaging.

References

1. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer 2019;144(8):1941–1953.
2. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023;73(1):17–48.
3. Hickman SE, Baxter GC, Gilbert FJ. Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations. Br J Cancer 2021;125(1):15–22.
4. Baker JA, Rosen EL, Lo JY, Gimenez EI, Walsh R, Soo MS. Computer-aided detection (CAD) in screening mammography: sensitivity of commercial CAD systems for detecting architectural distortion. AJR Am J Roentgenol 2003;181(4):1083–1088.
5. Baker JA, Lo JY, Delong DM, Floyd CE. Computer-aided detection in screening mammography: variability in cues. Radiology 2004;233(2):411–417.
6. Dromain C, Boyer B, Ferré R, Canale S, Delaloge S, Balleyguier C. Computed-aided diagnosis (CAD) in the detection of breast cancer. Eur J Radiol 2013;82(3):417–423.
7. Arun Kumar S, Sasikala S. Review on deep learning-based CAD systems for breast cancer diagnosis. Technol Cancer Res Treat 2023;22:15330338231177977.
8. Dahlblom V, Dustler M, Tingberg A, Zackrisson S. Breast cancer screening with digital breast tomosynthesis: comparison of different reading strategies implementing artificial intelligence. Eur Radiol 2023;33(5):3754–3765.
9. Kerschke L, Weigel S, Rodriguez-Ruiz A, Karssemeijer N, Heindel W. Using deep learning to assist readers during the arbitration process: a lesion-based retrospective evaluation of breast cancer screening performance. Eur Radiol 2022;32(2):842–852.
10. Romero-Martín S, Elías-Cabot E, Raya-Povedano JL, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. Stand-alone use of artificial intelligence for digital mammography and digital breast tomosynthesis screening: a retrospective evaluation. Radiology 2022;302(3):535–542.
11. Pinto MC, Rodriguez-Ruiz A, Pedersen K, et al. Impact of artificial intelligence decision support using deep learning on breast cancer screening interpretation with single-view wide-angle digital breast tomosynthesis. Radiology 2021;300(3):529–536.
12. Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. AI-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: a retrospective evaluation. Radiology 2021;300(1):57–65.
13. Shoshan Y, Bakalo R, Gilboa-Solomon F, et al. Artificial intelligence for reducing workload in breast cancer screening with digital breast tomosynthesis. Radiology 2022;303(1):69–77.
14. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 2019;111(9):916–922.
15. Transpara Breast Care. ScreenPoint Medical. https://transparabreastcare.com/transpara-breast-care/. Published 2020. Accessed August 14, 2023.
16. Artificial Intelligence from ScreenPoint Medical and Volpara Health Demonstrates Ability to Predict Long-Term Breast Cancer Risk. ScreenPoint Medical. https://screenpoint-medical.com/transpara-from-screenpoint-medical-demonstrates-ability-to-predict-long-term-breast-cancer-risk/. Published 2023. Accessed August 14, 2023.
17. Larsen M, Aglen CF, Lee CI, et al. Artificial intelligence evaluation of 122 969 mammography examinations from a population-based screening program. Radiology 2022;303(3):502–511.
18. iCAD Announces FDA Clearance for ProFound AI Version 3.0 for 3D Mammography. iCAD. https://www.icadmed.com/newsroom.html#!/posts/iCAD-Announces-FDA-Clearance-for-ProFound-AI-Version-3.0-for-3D-Mammography/248. Published 2021. Accessed August 14, 2023.
19. Conant EF, Toledano AY, Periaswamy S, et al. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. Radiol Artif Intell 2019;1(4):e180096.
20. Gollmer A, Link T, Weißenhofer S. Third evaluation report on the Austrian breast cancer early detection program. Evaluation report for the years 2014 to 2019 [in German]. https://jasmin.goeg.at/id/eprint/1875/. Published October 15, 2021. Updated August 14, 2022. Accessed January 19, 2024.
21. Olivotto IA, Gomi A, Bancej C, et al. Influence of delay to diagnosis on prognostic indicators of screen-detected breast carcinoma. Cancer 2002;94(8):2143–2150.
22. van Winkel SL, Rodríguez-Ruiz A, Appelman L, et al. Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study. Eur Radiol 2021;31(11):8682–8691.
23. Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. Eur Radiol 2021;31(3):1687–1692.
24. Graewingholt A, Duffy S. Retrospective comparison between single reading plus an artificial intelligence algorithm and two-view digital tomosynthesis with double reading in breast screening. J Med Screen 2021;28(3):365–368.
25. Benedikt RA, Boatsman JE, Swann CA, Kirkpatrick AD, Toledano AY. Concurrent computer-aided detection improves reading time of digital breast tomosynthesis and maintains interpretation performance in a multireader multicase study. AJR Am J Roentgenol 2018;210(3):685–694.
26. Balleyguier C, Arfi-Rouche J, Levy L, et al. Improving digital breast tomosynthesis reading time: a pilot multi-reader, multi-case study using concurrent computer-aided detection (CAD). Eur J Radiol 2017;97:83–89.
27. Boyd NF, Guo H, Martin LJ, et al. Mammographic density and the risk and detection of breast cancer. N Engl J Med 2007;356(3):227–236.
28. Boyd NF, Martin LJ, Yaffe MJ, Minkin S. Mammographic density and breast cancer risk: current understanding and future prospects. Breast Cancer Res 2011;13(6):223.
29. Ozcan BB, Patel BK, Banerjee I, Dogan BE. Artificial intelligence in breast imaging: challenges of integration into clinical practice. J Breast Imaging 2023;5(3):248–257.
30. Bhowmik A, Eskreis-Winkler S. Deep learning in breast imaging. BJR Open 2022;4(1):20210060.
31. Choy G, Khalilzadeh O, Michalski M, et al. Current applications and future impact of machine learning in radiology. Radiology 2018;288(2):318–328.