Diagnostic and Interventional Imaging (2019) **100**, 553—566

SOLICITED REVIEW / *Breast imaging*

Artificial intelligence and breast screening: French Radiology Community position paper

I. Thomassin-Naggara[a,b,*], C. Balleyguier[c], L. Ceugnart[d], P. Heid[e], G. Lenczner[f], A. Maire[g], B. Séradour[d], L. Verzaux[e], P. Taourel[c], and Conseil national professionnel de la radiologie et imagerie médicale (G4)

[a] *Collège des Enseignants de Radiologie de France (CERF), 75013 Paris, France*
[b] *Sorbonne Université & Tenon Hospital-AP—HP, 75020 Paris, France*
[c] *Société d'Imagerie de la Femme (SIFEM), 75014 Paris, France*
[d] *Société Francaise de Sénologie Pathologie Mammaire (SFSPM), 67000 Strasbourg, France*
[e] *Société Francaise de Radiologie, 75013 Paris, France*
[f] *Fédération Nationale des Médecins Radiologues, 75007 Paris, France*
[g] *AP—HP, Équipe Wind, 75012 Paris, France*

KEYWORDS
Artificial intelligence;
Breast cancer;
Screening program;
Digital breast tomosynthesis;
Digital mammography

Abstract The objective of this article was to evaluate the evidence currently available about the clinical value of artificial intelligence (AI) in breast imaging. Nine experts from the disciplines involved in breast disease management — including physicists and radiologists — convened a meeting on June 3, 2019 to discuss the evidence for the use of this technology in plenary and focused sessions. Prior to the meeting, the group performed a literature review on predefined topics. This paper presents the consensus reached by this working group on recommendations for the future use of AI in breast screening and related research topics.

Abbreviations: AI, Artificial intelligence; MG, Mammography; MLO, Medio-lateral-oblique; CC, Cranio-caudal; CAD, Computer-aided diagnosis; ANN, Artificial neural network; CNN, Convolutional neural network; Synth2DMG, Synthetic mammography; QC, Quality control; DL, Deep learning; R2, Second reader for breast screening; FDA, US Food and Drug Administration; DBT, Digital breast tomosynthesis; AP-HP, Assistance Publique—Hôpitaux de Paris.

* Corresponding author. Department of Radiology, Hôpital Tenon, 4 rue de la Chine, 75020 Paris, France.
E-mail address: isabelle.thomassin@aphp.fr (I. Thomassin-Naggara).

https://doi.org/10.1016/j.diii.2019.08.005
2211-5684/

Introduction

During the last 10 years, the number of publications on artificial intelligence (AI) in radiology has markedly increased to reach 700—800 per year. Breast screening is one of the major applications of AI in radiological imaging [1,2]. This particular interest in breast cancer can be explained by the following:

- firstly, breast cancer is a public health issue with 58,459 incident cancers in France in 2019 and 12,146 annual deaths with an overall survival of 87% [3];
- secondly, breast screening is based on a clinical and radiographic examination, namely mammography (MG), which includes four 2D incidences (2 medio-lateral-oblique [MLO] and 2 cranio-caudal [CC]) with an examination every 2 years in Europe and annually in the United States of America (USA);
- thirdly, MG is limited by a wide variability in interpretation by the radiologist [4] and a second reading of each mammogram can be necessary often because of the masking effect of dense tissue [5];
- and finally, computer-aided diagnosis (CAD), which was widely used in MG since its rapid uptake in the USA 20 years ago, has been largely abandoned due to a high level of false positives leaving breast radiologists somewhat skeptical about this concept [6,7].

The French Society of Radiology (SFR) and the College des Enseignants de Radiologie (CERF) organized a consensus meeting in Nimes, France on June 3 and 4, 2019 to evaluate the evidence currently available about the clinical value of AI in breast imaging. Nine experts from the disciplines involved in breast disease management, including physicists and radiologists, discussed the evidence for the use of this technology in plenary and focused sessions. Prior to the meeting, the group performed a literature review on predefined topics and identified questions to be discussed at the meeting. The articles were selected based on the following Pubmed search: ''(''breast neoplasms''[MeSH Terms] OR (''breast''[All Fields] AND ''neoplasms''[All Fields]) OR ''breast neoplasms''[All Fields] OR (''breast''[All Fields] AND ''cancer''[All Fields]) OR ''breast cancer''[All Fields]) AND (''artificial intelligence''[MeSH Terms] OR (''artificial''[All Fields] AND ''intelligence''[All Fields]) OR ''artificial intelligence''[All Fields]) AND (''Bildgebung''[Journal] OR ''imaging''[All Fields]) ($n = 775$). Finally, 41 articles were selected as fitting with the five main breast imaging fields of the application of AI to which this article is dedicated: image acquisition, automatic assessment of breast density, the current AI CAD programs working on 2D mammography (2DMG) and breast tomosynthesis, synthetic mammography (synth2DMG), and personalized screening. Moreover, two chapters of this article detail the implementation of AI in France taking into account the specific French screening organization, and present the different French databases. Each chapter was reviewed and discussed by the working group during the plenary session to reach a consensus about recommendations for future research topics for AI. Following the meeting, the literature review was updated to June, 30th 2019. The working group intends to further update these recommendations as and when new relevant evidence becomes available.

Background

In the field of breast imaging, as in all radiology fields, five main groups of application exist in AI [8]. They include lesion classification, image processing, generative tasks, regression and workflow and efficiency. Lesion classification consists of predicting the nature of a group of pixels to distinguish, for example, between a tumor versus normal tissue and a malignant versus benign lesion. Image processing, mainly including tissue and lesion segmentation, consists of identifying which pixels are parts of a structure of interest, which pixels are abnormal within an identified structure, and labeling each pixel in an image with its type (semantic segmentation). Some authors suggest that the notion of ''image-omics'' (*i.e.*, radiological classifiers and precision medicine for prognostic imaging) may be better than tissue genomics [9]. Generative task, which is currently in development, consists of creating new images based on existing images. Regression, which consists of predicting a continuous variable from inputs (*e.g.*, predicting age from a hand radiograph) [10]. Workflow and efficiency may result in reducing radiation doses and acquisition time.

The term ''AI'' covers many different training techniques including artificial neural networks (ANNs), machine learning (ML) and deep learning (DL) [11]. ML is a data driven learning approach with a mathematical model based on the observed ''training'' data in which there are two main models. Supervised AI describes learning based on features from labeled images and consists of ANNs including convolutional neural networks (CNNs), support vector machines, random forest, linear discriminant analysis, and decision trees. The highest accuracy in the support vector machine method is observed in the results of a study which uses an appropriate segmentation method for obtaining the desired area in the image. The shape and intensity of the extracted features has the greatest effect in classification. The combination of gray-level co-occurrence matrix (GLCM) and ratio features along with morphological features result in the highest accuracy. These types of algorithms are classic ML (type ANN) that need relatively low computational requirements in comparison with DL architectures algorithms (type CNN) that require millions of parameters and thus high-performance computing hardware. Unsupervised AI (clustering) describes learning when the data has no diagnosis or normal/abnormal labels and is represented mainly (77%) by k-nearest neighbor (k-NN). One of the most well-known applications in breast pathology is the description of intrinsic molecular subtypes for breast cancer [12].

While supervised learning algorithms are primarily used, with area under the curve (AUC) values from receiver operating characteristic (ROC) curve analysis ranging from 0.74 to 0.98 (median, 0.87) and with that from prognostic imaging ranging from 0.62 to 0.88 (median, 0.80), unsupervised learning is mainly used for image processing purposes [8]. Currently, ANN, support vector machine, and clustering are the most frequently used algorithms, accounting for 66% of all AI imaging publications [8].

2DMG today and in the near future: what can we expect from AI?

Radiation dose optimization and quality control

DL algorithms have been developed to improve images by speeding up acquisition time and outperforming traditional noise reduction techniques in image reconstruction. Generative adversarial networks (GANs) in particular are going to have a marked impact in radiology [9].

Manufacturers have developed different tools to reduce radiation doses. A first approach was introduced by Siemens Healthineers with an optional software (Progressive Reconstruction Intelligently Minimizing Exposure [PRIME®], which can be used for breast thicknesses of up to 70 mm and correct for scatter without the use of an anti-scatter grid. Instead of using a grid, the software identifies structures in the breast that cause scatter and subtracts the calculated scatter. Fewer mAs are thus required as the X-rays are not absorbed by a grid. The degree of dose saving depends on breast thickness and structure.

To implement tomosynthesis in breast cancer screening, manufacturers should optimize image quality in synth2DMG. Today, only ''real'' 2DMG with an image acquisition can be used in Europe. This implies that the radiation dose can be double for a complete examination with 2D and 3D images. Synth2DMG is currently integrated in some systems and is being developed for others.

In the USA, Hologic® received Food and Drug Administration (FDA) approval for C-View™ in 2013. Thus, synth2DMG may replace the conventional 2D image. Approval was based on a Hologic® study showing that 3D + C-View is non-inferior to conventional 2DMG [13].

The FDA also approved ''high definition'' breast tomosynthesis technology from Siemens Healthineers in 2017. This technology incorporates software called EMPIRE® (enhanced multiple parameter iterative reconstruction), a combination of iterative and ML algorithms, and has been approved as a 3D-only examination. However, it also includes ''Insight 2D and 3D'', synthetic software that generates tomosynthesis volumes in 3D from which 2D images can be obtained without additional dose for the 2D exposure.

Finally, it should be noted that while quality control (QC) has been well defined and applied for full field digital mammography (FFDM), QC guidelines have not been finalized for tomosynthesis as yet. To date, tomosynthesis has not yet been approved in national screening programmes in Europe and its use is limited to some experimental breast cancer screening studies, such as the TOMMY Trial in the United Kingdom (UK). Nevertheless, the EUREF group published tomosynthesis QC guidelines on their web site in 2015. In the UK, the National Co-ordinating Centre for the Physics of Mammography is running QC tests based on the EUREF guidelines on behalf of the National Health Service Breast Screening Programme and has developed tools to analysed 3D QC images. However, the time required to analyze QC data remains long today.

The situation in France is more complicated. More than 400 units have been used daily without any QC (daily/weekly/monthly radiographer's tests or semestrial physicists' tests) for years, not within the framework of the breast cancer screening program but for all monitoring mammograms and additional diagnostic examinations. The first evaluation tests done in France, based on the EUREF guidelines, show a huge disparity with the manufacturers in terms of image quality and dose, even when applying manufacturer adjustments for the same brand of systems. Without a normalized evaluation and a strict QC, it is not possible to know the physical performances of the installed systems and therefore the potential clinical results. In the meantime, QC is performed in France by private companies using technicians who sometimes lack adequate training. The lack of expert medical physicists in France and the number of private radiology centers practicing MG (more than 2500) poses a problem for the implementation of effective QC in tomosynthesis and in synthetic view evaluation.

Quality of image acquisition

Positioning parameters

An important issue in breast cancer screening is the radiation dose, which correlates with breast thickness and therefore the correct positioning of the patient [14]. AI could be used to help the technician achieve optimal positioning by defining the right compression force and showing the right exposure parameters. Volpara Solutions has developed software that can give feedback to radiographers and radiologists on the quality of all these parameters. After each exposure, anonymized data are sent to an external cloud database that analyzes the final quality of each diagnostic image and advises on all acquisition parameters. General Electric Healthcare is currently developing similar tools that provide analysis of each examination based on qualitative criteria. These various software options which check the quality of the acquisition parameters can therefore be used as a continuous training tool.

Automatic breast density assessment

Dense breasts are associated with a higher risk of breast cancer [15—17]. In addition, breast density assessment is crucial because dense tissue may mask cancer detection leading to a higher proportion of cancers occurring between MG screening sessions (known as interval cancers) [18]. It is well known that breast density assessment varies widely not only from one radiologist to another but also for the same radiologist at two different times. A large prospective multicenter observational study ($n = 216,783$ including 34,271 patients seen several times), showed low inter- and intra-observer agreement in 17% of women who were assigned to different categories on successive mammogram readings [19]. Therefore, there is a need for more reproducible software.

The first software tools to allow automatic assessment of breast density were created 10 years ago with a view to improving performance and reproducibility. Most of these were based on segmentation techniques and either did not reach, or barely reached, the accuracy of the College of Radiology Breast Imaging Reporting and Data System (BI-RADS) subjective assessment in predicting breast cancers. These software tools have recently been redesigned to integrate the DL model (QUANTRA 2.2®) so that further evaluations are needed.

In a recent publication, a deep CNN was compared with human analysis of breast density according to the BI-RADS lexicon 2013 [20—22] on a set of 20,578 mammographic images resulting from data augmentation from 12,932 MLO and CC images [23]. This study demonstrated very good accordance in distinguishing between MG for fatty breast (breast density rated A and B) from dense breast (breast density rated C and D) with an agreement of 99% for MLO views and 96% for CC views.

The results of Ciritsis et al. study [23] are in line with those of a subsequent publication by Lehman et al. [24] who also developed a deep CNN, ResNet-18, with PyTorch (2018, version 0.31; pytorch.org) using 58,894 randomly selected digital mammograms from 39,272 women screened between January 2009 and May 2011. The resulting AI algorithm was tested in 10,763 consecutive screening digital mammograms from January to May 2018 by the authors who showed that the DL model matched the radiologist's interpretation in 78% of mammograms for four-way BI-RADS categorization, and in 94% for binary categorization of dense or non-dense breasts [24].

Other authors correlated an automatic assessment of breast density with commercial software based on an AI technique with screening population characteristics in Norway ($n = 107,949$) on 307,015 MG [25]. They concluded that screening examinations of women with dense breasts assessed by automated software resulted in a higher recall rate, lower sensitivity, larger tumor diameter, and more lymph node—positive disease compared with women with nondense breasts [25].

Finally, in addition to breast density, Kontos et al. recently published a study defining radiomic phenotypes as descriptors of the complexity of breast parenchyma [26]. On a cohort of 2241 women with MG + digital breast tomosynthesis (DBT) (both views), they performed an unsupervised hierarchical clustering in a training population ($n = 1339$) and a validation set ($n = 690$) and classified the parenchyma into four levels of complexity. In their article, they demonstrate that low-intermediate-complex parenchyma have an OR higher than 2 to develop a breast cancer, independently to breast density [26].

Research issues

- To develop cloud software that automatically analyze daily images and detect errors or instabilities to help medical physicists decrease analysis time during quality control (QC).
- To create automatic analysis of QC criteria to simplify daily QC and decrease time for technologists.
- To propose automatic patient assessment QC for technologists (positioning, compression, blurring, artifact) (report for technologist's self-assessment).

To optimize automatic assessment of parenchymal evaluation including density, complexity, heterogeneity to equal BI-RADS classification accuracy for predicting breast cancer risk for radiologists.

Which AI algorithms have been trained and validated on 2DMG?

CAD performance

The first CAD in the field of breast imaging was approved in 1998 and used from 2002 after obtaining reimbursement status in the USA. In 2006, CAD was used for 92% of all screening mammograms [27]. CAD can be used not only for detection but also for diagnosis to help classify and interpret images. In our field, CAD is probably mainly used for detection but the new generation of CAD systems can be used for both detection and characterization.

The first generation of CAD were able to mark 86% of missed calcifications, 72% of masses, and, in some studies, 42% of very subtle mammographic cancer findings deemed occult for the radiologist [28]. However, in clinical practice, the majority of these marks were considered as false-positives and dismissed by the radiologist. Nevertheless, even if this traditional CAD did not identify all the cancers detected by a radiologist, they did improve sensitivity and even reached accuracy of a second reader (R2) for detection albeit with less specificity. In a 2007 study, Fenton et al. found that the specificity of CAD was significantly lower and resulted in a 20% increase in the biopsy rate, lowering overall accuracy (AUC of 0.807 vs. 0.919) [6]. In summary, traditional CAD have very high sensitivity for the detection of calcifications (99%), lower sensitivity for masses (75—99%) and poor sensitivity for architectural distortion (38%). The use of CAD in general practice was evaluated in a large retrospective study from the Breast Cancer Surveillance Consortium (BCSC) in 600,000 mammograms read with and without CAD by 271 radiologists across 66 facilities [7]. With a cancer detection rate of 4.1%, sensitivity and specificity were identical with or without the use of CAD. Performance with CAD was even poorer for 107 radiologists, especially for sensitivity (83% with CAD vs. 89% without CAD).

The technical limitations of traditional CADs were small datasets, poor quality image (digitized image, no quality standard), insufficiency of computer processing (unable to include multiple views and prior examinations), an absence of dynamic improvement (limited to periodic software upgrades) and a selection of reference cases and images depending on human expertise.

New CAD with emerging DL algorithm

Today, software programs known as ''AI CAD'' are able to work with very large databases and improve performance by learning from new cases. Nevertheless, developing new CAD software is complex even with huge computer potentiality. It requires working with a large database which can be supervised (each image is labelled by a human) or unsupervised (computers alone discern from the image characteristics from non-labeled databases). The supervised approach is costly and not exempt from mistakes and approximations, whereas the unsupervised approach requires high quality raw data in full resolution mammograms in each view and, if possible, prior images which generate very high data volume. The second major problem of the unsupervised approach is the complexity of the algorithm which can comprise between 30—150 layers. Users need to understand how

it works and assess learned parameters to avoid overfitting (learning about idiosyncratic variation without understanding the clinical impact of these variations). For CAD to be implemented in the screening process, four aspects need to be improved. The first is performance. For a given high sensitivity, the major drawback is the false-positive flags which alter the performance of the screening test and is currently a major topic of research. In a recent study, AI CAD (cmAssit from CureMetrix®) was used to reduce false positive results in screening compared with traditional CAD (ImageChecker-Hologic®) [29]. There was 69% reduction in false positive mark lesions with AI CAD for the same performance for both masses and calcifications. The second is interpretation time which AI CAD increases by approximately 20%. Although reading time is one of the most important features of a centralized screening program, it is not of major concern for the French program in the first reading setting. It can, nevertheless, be a key issue for the second reading session. In the same study [29], the reading time was decreased by 64% with AI CAD compared to traditional CAD. This potential time saving implies a 10% increase in screening capacity. However, the retrospective design of this study and the very small sample size with few cases of cancer (250 FFDM) constitute two major limitations. The third way in which AI CAD can be implemented is in the work flow. Using AI CAD in current practice implies complete integration in the post-acquisition process, especially working on existing workstations, and with no delay in presenting images on the screen. Thus, AI CAD needs to be compatible with all manufacturers and, similarly, manufacturers need to open their system to CAD providers. Finally, AI CAD must be shown to be cost-effective for implementation. This is a major concern especially in France where CAD is not currently reimbursed unlike in some other countries. One way in which AI CAD could be interesting in terms of cost cutting (pending evaluation) is in reducing the need for two readings which would considerably reduce the volume of mammograms (each radiologist reading currently costs 4 euros representing 20 million euros/year for two processes).

Review of the level of development of the various algorithms available on the market

Many AI algorithms, mainly based on deep CNN techniques, have been developed over the last 5 years. However, the level of validation for clinical implementation differs and should be well understood before use in a clinical setting (Table 1).

Transpara™, which has been approved for use both in Europe and the USA, was developed by ScreenPoint Medical and is currently distributed by Incepto™ in France. The algorithm was developed on the basis on 9000 truly positive and 180,000 truly negative mammograms provided by different constructors and trained on 2DMG and DBT images. The radiologists provided a risk of malignancy based on the BI-RADS assessment scale that ranges from 0 to 6 while this AI system offers three different decision tools: an interactive decision tool that provides a local cancer likelihood score (1—100) activated by clicking on a specific breast region; a traditional lesion marker for computer-detected abnormalities; and a proprietary examination-based cancer likelihood score with a score ranging from 1 to 10, with a score calibrated such that the number of mammograms in each category is approximately equal.

In 2019, a multicenter and multi-reader study demonstrated the non-inferiority of this AI system compared to the average of 101 radiologist readings (AUC = 0.840 vs. 0.814, respectively) [30]. In this study, each dataset consisted of 2DMG acquired with different systems from the four different vendors (GE Healthcare, Siemens Healthineers, Hologic® and Philips Healthcare) and the reference standard which was either histopathological analysis or follow-up in a total of 2652 examinations (prevalence of malignancy 653/2652; 24.6%). The performance of the AI system was superior to more than 61.4% of radiologists. The sensitivity and specificity of the system was also better than the majority of the radiologists, but always worse than the best radiologist. The authors suggest that the AI system could be used as an independent stand-alone first or second reader in countries lacking experienced breast radiologist [31] or as an interactive decision support tool [32]. In this setting, the same authors compared the breast cancer detection rate of radiologists reading 2DMG unaided versus supported by this AI system on an enriched cohort of 240 women (100 showing cancers, 40 with false-positive findings and 100 2DMG with normal findings) [33]. In this study, the AUC of the AI system was similar to the average of 14 board-certified radiologists but lower than the AUC of a radiologist supported by the system. Reading time per examination was similar (146 s vs. 149 s). However, the high prevalence of breast cancers in these validation sets probably resulted in an overestimation of the accuracy of the AI system and further studies are needed especially for use as a stand-alone technique. Moreover, a stand-alone approach imposes the question of who would take ultimate responsibility for any undetected breast cancer (which remains the most litigious situation for medical malpractice lawsuits) by an imperfectly performing AI algorithm [34].

Other AI algorithms have been trained but not yet externally validated in clinical conditions. Therapixel was created in 2013 by two researchers from the French National Institute for Computer Science and Applied Mathematics and took the joint 1st place of the Digital DREAM Mammography Challenge, an international competition in DL applied to mammography. This competition — organized jointly by the National Cancer Institute, the Group Health Cooperative, the Icahn School of Medicine at Mount Sinai, the FDA, Apple and IBM — gathered about 1200 participants to compare the best breast cancer prediction algorithms based on screening mammograms [23]. This challenge was based on a set of 320,000 2DMG with 1200 breast cancers (prevalence: 3.7/1000 patients). After four consecutive rounds, Therapixel reached the co-first place of the challenge with an accuracy of 75%. However, no comparison with radiologist interpretation was available.

Mammography Intelligent Assessment (MIA)® is an AI algorithm developed by Kheiron as part of a National Health Service grant in the UK. This algorithm was trained on more than one million MG images from the UK's breast screening program and validated in a retrospective multicenter study with 3854 MG from four UK screening sites (prevalence 6.9%). MIA® displayed a sensitivity of 85%—97% and a specificity of 50%—94% in the four sites. No comparison was available with radiological interpretation in this study which

Table 1 Degree of validation of the main available artificial intelligence algorithms for 2D breast screening based on deep learning convolutional neural networks.

Software	Internal validation	External validation Publications	Positive points	Negative points
Transpara™ (ScreenPoint Medical/Incepto™)	Over 9000 MG with cancer (one-third of which are presented as lesions with calcifications) and 180,000 MG without abnormalities. The MG originated from devices from 4 vendors (Hologic®, GE Healthcare, Siemens Healthineers, Philips healthcare, Fuji) and institutions from Europe, USA, and Asia Validation: independent dataset representative of screening population with enriched prevalence of cancer	Rodriguez-Ruiz et al. [30] Rodriguez-Ruiz et al. [33] Rodriguez-Ruiz et al. [45] Lång K et al. [Presented at ECR 2019, Vienna, SB-0696]	Provides 3 different outputs to aid radiologists in detection, classification and triage/workflow optimization This system can be applied to processed DM images and DBT volumes Multi-vendor Support for combined examinations (with both 2D and 3D digital breast DBT)	Currently, the AI system does not use information from prior MG Two different modules to detect calcifications and soft-tissue lesions
Therapixel	Vendors: Hologic®, GE Healthcare, Siemens Healthineers, Philips healthcare, Fuji) Origin: USA, France and UK Quantity: several hundred thousand MG Enriched cohort with annotations of all cancers Validation: independent dataset with a 10% prevalence of cancer	Data challenge: 320,000 cases representative of screening population with enriched prevalence of cancer A new multicenter study is ongoing	Enriched cohort Multicenter cases	
iCAD Inc.	North America and Europe > 2000 cases to train and internally test the algorithm, including > 400 cancer cases Vendors: GE Healthcare, Hologic® and Siemens Healthineers Roughly 50% of the cases were used for training, and 50% for testing Prevalence of cancer was 50%	A separate independent regulatory set was used for the reader study and reporting the standalone performance Reading was performed by independent company (Intrinsic Imaging)		

AI: Artificial intelligence; MG: mammogram; DBT: digital breast tomosynthesis.

was presented in an industrial workshop at the annual meeting of the Radiology Society of North America in 2018. This algorithm is CE approved and waiting for FDA approval.

Arterys, a pioneer in cloud-based medical imaging software, is also developing an AI algorithm in partnership French developers to train and validate their new algorithm. They are currently in the preliminary phase of training.

AI Research issues
- To validate artificial intelligence (AI) models in a French population.
- To test the ability of AI systems to integrate comparisons with previous mammograms.
- To test the performance of AI algorithms on combined two-view MG analysis.
- To compare two readings versus one reading + AI software.
- To test AI algorithms on 2DMG according to the different vendors.

Breast tomosynthesis: the future of MG

Have any AI algorithms been trained and validated on breast tomosynthesis?

According to the six published meta-analyses about the value of DBT for screening [35—40], DBT combined with 2DMG is more sensitive and more specific than 2DMG alone with a higher detection rate of invasive cancers. Moreover, retrospective studies have demonstrated a lower recall rate. Thus, in the near future, DBT will become the standard technique for breast screening implying that AI algorithms should be trained and validated as soon as possible on DBT data.

The interest of AI in DBT is threefold:
- to detect more lesions while maintaining an acceptable rate of false positives;
- to improve the characterization of breast lesions, whether they were detected by MG or by DBT;
- to decrease the time of reading, which may be relevant for a screening method which has been shown to double the reading time.

Nevertheless, several issues in the use of AI in DBT require attention:
- should the model use projection view images, DBT reconstruction slices, a combination of both, 3D reconstructed volume or the derivatives of the reconstructed images such as synthetic mammograms?
- should the AI model use large data collected from MG [41] or should it be built exclusively with DBT data?
- how can the region-based CNNs (rCNNs) be used — computationally expensive and highly time-consuming in a process with a huge set of data such as DBT — without overly slowing down the reading, and what is the value of faster rCNNs? [42].

Two companies have industrially developed AI in DBT: iCAD with ProFound AI™ and ScreenPoint Medical, which has developed Transpara™. Transpara™ for DBT analyzes and interprets full 3D data from the DBT volumes. Similar to Transpara™ for MG it delivers interactive decision support including the detection of soft-tissue lesions and calcifications, interpretation of suspicious regions and automated linking of MLO and CC views. Furthermore, Transpara™ for DBT uses synthetic images for intelligent navigation in both MLO and CC views. However, no results about the use of Transpara™ for DBT in clinical conditions have been presented to date.

The first results of a new version of a commercial software called ProFound AI® (iCAD) which was trained and validated on DBT images, were recently presented at the 2019 European Congress of Radiology (ECR). In this retrospective, fully-crossed, multi-reader (24 radiologists) study based on 260 cases with 127 dense breasts and 133 nondense breasts, radiologists performed better when supported by the AI algorithm in both dense and non-dense breasts. Sensitivity and specificity significantly increased by 7% and 9.9% in dense breasts and 9% and 4% in nondense breasts. Moreover, reading times decreased for the 24 radiologists with AI by 57.4% in dense breasts and 47.6% in nondense breast. The main limitation of this study is the fact that only one vendor was represented, and these results must be consolidated by a multivendor study. This algorithm was trained on a database of 12,000 DBT including 4000 proven cancers. It was able to detect and diagnose calcified and non-calcified suspicious lesions. Depending on the version of the workstation, the radiologist may select one of three sensitivity levels of detection: low (88% overall sensitivity, and 72% specificity); medium (91% overall sensitivity and 59% specificity); and high (95% overall sensitivity and 31% specificity). The program provides two types of information: a case score (*i.e.*, the probability that the entire case is malignant), and a lesion score (*i.e.*, the probability that a marked lesion is malignant). The probability score ranges from 0 to 100% with a higher score indicating a higher level of confidence in the malignancy of the detection or case. The scores were calibrated on an enriched cohort with a 25% prevalence of cancer. Thus, these scores should be interpreted as the probability of detecting a cancer correctly in a population of 25% cancers and 75% non-cancers. This is a main limitation for an application for breast screening as the positive predictive value is more than certainly highly overestimated.

Are AI algorithms useful to optimize synthetic MG?

Two-view DBT is better than one-view DBT for detecting breast cancer and limiting the number of false-positive findings [43]. Synth2DMG is thus of utmost importance to limit the radiation dose. Since synth2DMG will become the standard in DBT, in order to avoid 2D acquisition and subsequent radiation exposure, it is mandatory to know which tool is used by the different vendors to perform synth2DMG. A recent study (the Oslo Trial) which found that the sensitivity of 3D + synth2DMG was not superior to the sensitivity of 2D, was the first to conclude that this ''negative'' result underlines the potential differences in reliability of synth2DMG from one vendor to another [13]. Synth2DMG is increasingly integrating AI, as shown by the development of synth2DMG by Hologic®.

The first generation of synth2DMG was an algorithm using ML techniques to generate a synthesized 2D image from 100-micron tomosynthesis reconstructions. This is much more than a basic maximum intensity projection function of a volume. In order to avoid superimposed tissue mimicking suspicious areas and to improve the visibility of true structures, the C-View algorithm analyzes each tomosynthetic slice as well as adjacent slices (above and below) to differentiate normal structures from suspicious ones and to find microcalcifications barely visible in conventional 2D due to breast thickness attenuation.

The second version of synth2DMG was built on an advanced ML technique operating on Hologic® 70 μm tomosynthesis reconstructions to generate a synthesized 2D image. Once again to avoid superimposed tissue mimicking suspicious area and to improve visibility of true structures, the algorithm (named I2D) analyzes each tomosynthetic slice as well as adjacent slices to differentiate normal structures from suspicious ones and to find microcalcifications. These findings will be better depicted in I2D because of the advanced AI identification, which increases the conspicuity of the identified lesions. The higher resolution of the tomosynthesis reconstructions, along with advanced AI algorithms, theoretically enables better identification of suspicious lesions while reducing the enhancement of false positives, compared to synth2DMG. In addition, AI was used to take into account details of the images such as breast density and parenchymal arrangement when creating the synthesized image so that the synthesized image resembles a conventional 2D image as much as possible while maintaining the increased conspicuity of suspicious lesions.

AI research issues

- To compare the accuracy of AI algorithms in one- and two- view digital breast tomosynthesis (DBT) images by assessing readings with and without AI of one- and two-view DBT images with synthetic reconstruction.
- To evaluate the added value of AI in terms of the characteristics of the cancer detected (size, grade).
- To evaluate the modifications due to AI in terms of false positives and false positives requiring a biopsy (a potential issue of DBT).
- To evaluate the time of reading with and without DBT.
- To optimize synth2DMG quality by an AI algorithm.
- To test AI algorithms on synth2DMG and according to different vendors.

Organization of breast screening in France in 2019: what is the expected impact of artificial intelligence?

In France, 2.5 million women undergo a breast screening MG each year [3]. Seventy percent of the MG units are digital (DR) and 30% are computed radiography units. Each regional center performs between 5000 to 70,000 mammograms a year. The number of first reader radiologists (> 500 mammograms read per year) per department is between 5 and 200 and the number of second readers (> 2500 mammograms read per year) is between 5 and 50. A second reading is performed with the previous mammograms and the knowledge of the clinical examination [44].

In France, 175,000 breast screening mammograms (7%) are considered as abnormal at the first reading (rated BI-RADS 0, 3,4, 5) while 2,325,000 mammograms are considered as normal at the first reading but referred for a second reading. In addition, 7000 mammograms (4%) considered as abnormal before supplementary incidences and or ultrasonography (performed at the same time as the first reading) and are finally reclassified as normal but are also referred for a second reading. Thus, 2,332,000 mammograms with normal findings are referred to the regional screening center for a second reading after which 1% are reclassified as ''suspicious'' (n = 23,320). After complementary incidences and or ultrasonography performed for these mammograms considered as positive by the R2, 22% are confirmed to be abnormal which corresponds to 0.3—0.4/1000 cancers. Finally, the French breast screening program detects 7 cancers per thousand screened women, including 5% of breast cancers detected by an R2 in 2019. This percentage has decreased over the last years form about 10% [44].

The French screening system is very different to that of other countries and few studies can be implemented in France. First, the screening attendance rate is highly heterogeneous varying from less than 30% in Paris to more than 65% in some departments in the west or center of France. This can be partially explained by the rate of individual screening (*i.e.*, women presenting for office-based MG outside of the national screening program) in some regions especially in the south of France and in the Parisian region, but there are also huge differences between towns and even between districts in a given town. The attendance rate varies also with age with a higher participation for women between 55—65 years than for women over 70 years. Other differences in participation can be seen between different socio-economic groups. A first application of AI from the national French medical database crossed with other databases like taxes, social care or the unemployment benefits registry may be helpful to identify and communicate specifically with these populations to improve screening attendance.

A second application of AI is related to the second reading in France. The screening program is organized by the regions and performed by experienced radiologists but radiologists in France can perform complementary or ultrasonographic examination (*i.e.*, immediate diagnostic work-up) when they detect a clinical or MG abnormality. With this French specificity, the recall rate after a second reading is only 1.1% of all participants who are sent back to the first radiologist for a deferred diagnostic work-up. Of these women, 22% will have a positive diagnosis resulting in detection of 6% of all cancers or 0.4 cancer for 1000 women. The performance of the R2 varies from region to region for reasons which remain to be clarified. The second reading is always done on film which is costly at around 20 million euros per year (film transfer to the centralized R2 unit, film display on light box by technician, radiologist fees, data monitoring, screening results and sending the film back to the woman and

her medical practitioner). Some small studies about data dematerialization are currently underway but have not been published to date.

The value of a second reading is still under debate especially following the recent governmental decision in France to prohibit screen-film MG and the current system of transfer from computed radiography (CR) to digital radiography (DR) as CR system detects less cancer than DR system. Moreover, the potential implementation in the screening program of DBT which improves the first reading cancer detection rate and decreases the recall rate is also an argument against a second reading. The results of the recently published Oslo trial shows that digital MG plus DBT performs better than FFDM and a double reading [13].

As mentioned above, before CAD is used as an R2, great improvements will have to be made in terms of false positive marks, interpretation time, the development of an easy work flow, and finally in terms of cost-effectiveness.

Thus, organization of breast screening in France may benefit from the development of AI in the following fields:

- improvement in image quality by helping the technician to optimize positioning to reduce blurring, and by validating this step before sending the mammogram to the radiologist for interpretation;
- better selection of women who may benefit from breast ultrasonographic examination. Breast ultrasonography is currently performed in 20% of all women after a normal mammography for dense breast, and is the only examination performed in 76% of women during immediate diagnostic work-up, but detects only 2.4% of cancers. We need software which can determine not only global density but also, as radiologists do, focal high breast density which is sometimes a good indication for breast ultrasonography. Another research topic is the development of AI CAD for breast ultrasonography. Two potential devices are currently being developed: one is an automated breast ultrasound (ABUS) system called QVCAD (Qview Medical) and the other a cloud-based CAD system developed by Koios Medical able to work on picture archiving and communication systems (PACS) images (Koios Medical). However, there are no published studies about these systems to date;
- better first-round triage and possibly avoiding a second reading for some MGs, thereby lowering the cost and reducing the time it takes to get the results to the patient and medical referee. One study tested the ability of an AI algorithm (Transpara, ScreenPoint) to exclude examinations with the lowest likelihood ratio of malignancy [45]. This study was conducted on 2562 examinations and demonstrated that excluding mammograms with a score 1 or 2 resulted in a 17% decrease in the number of examinations and 5% of false positive cases (reducing the recall rate) while only missing 1% of cancers. The radiologist's performance was unchanged in a new cohort with a higher prevalence of breast cancer;
- improvement in the radiologist's performance, which is important in our decentralized program and would reduce the need of second readings.

AI Research issues

- Organization issue
  - To identify reasons for non-participation in the breast screening program in the French national medical database.
  - To analyze differences in the cancer detection rates at a local level (town, districts) to better select women at higher risk.
  - To establish prediction parameters to improve cost effectiveness of organized breast screening programs.
- Reading issues
  - To predict comorbidities (vascular calcification correlated with heart attack).
  - To improve screening performance of radiologists with different levels of experience.
  - To better select patients for ultrasonographic evaluation.

What do we need to develop AI for breast screening in France?

Dexter Hadley estimates that screening algorithms should be trained on millions of mammograms and suggested that AI researchers should embrace bitcoin technology [46]. To date, many validation cohorts are enriched with a high proportion of cancer-positive mammograms than detected in routine clinical screening. This type of cohort may induce bias with radiologists being overcautious resulting in higher recall rates and lower accuracy. We can split the development of an AI algorithm for detecting breast cancer using DL CNN into three main phases:

- first, the model is built on an enriched data cohort with a high volume of breast cancers, ideally including all types of cancers seen on MG (spiculated mass, round mass, cluster of microcalcifications, architectural distortion). Most of time, there is another data set in the same cohort not used for the training which will be used for internal validation;
- second, the model needs to be externally validated in another cohort with a lower prevalence of cancer and possibly with more subtle cancers to improve its accuracy;
- the last step is to test the model on an independent data set with a prevalence representative of the screening population before clinically validating the model in a randomized trial comparing the accuracy of the model with radiologists' performance.

We need all these steps because incremental improvement in the AUC is not directly translatable to improved patient outcomes in the clinical setting. It is uncertain what proportion of examinations that a commercial AI system would flag as having more than 2% malignancy, requiring additional diagnostic work-up under our current clinical practice thresholds. Moreover, we learned from the experience of CAD in MG that adopting promising new technologies too quickly can be a costly mistake; later found to lead to more false positives without improved cancer detection [47].

Therefore, an algorithm needs more than one million 2D MG/DBT to truly demonstrate its efficacy and a dedicated platform must be developed to build larger validation data sets more representative of the screening population. In France, the government created a Health Data Warehouse in 2018 to pool data, including mammograms, from public/private clinical practice in the same environment. However, breast screening clinical and follow-up data are recorded in regional centers for cancer screening programs. Thus, a connection needs to be created.

What can we expect from regional breast screening centers?

Regional breast screening centers collect all data related to the second reading from the departmental centers, including epidemiological information such as age, menopausal status, personal history of breast biopsy, familial history of breast cancer, the presence/absence of an abnormal MG image, the side, the type of MG abnormality (cluster of microcalcifications, architectural distortion, mass or asymmetric density) and breast density (according to the BI-RADS classification) as well as follow-up data for all abnormal MG (type of biopsy, type of surgery, histopathological findings).

Moreover, the role of breast screening centers today is to record all interval cancers (detected after a normal first or second reading) which is crucial in evaluating the value of a breast screening program.

The regional breast screening center also collects first reading data regarding personal history context, clinical examination, the type of vendors, type of MG abnormalities, the number of ultrasounds performed after normal MG, the type of supplementary incidences performed (including DBT), and the type of pathological findings detected.

Thus a lot of health and clinical data are available in the regional breast screening centers which are connected with the administrative information for each patient (name, birth date, social security number.). These huge amounts of very informative data should shortly be correlated with the Health Data Warehouse where mammograms are stored.

What can we expect from a hospital Health Data Warehouse? The AP—HP model

L'Assistance Publique des Hôpitaux de Paris (AP—HP) is a unique international structure that comprises 39 public hospitals federated in a single legal entity. Each day, the AP—HP information system collects numerous medical data in a wide variety of software and databases used to monitor patients' care pathways. Exploiting this data for AI research, especially in the domain of DL, requires new big-data approaches and the creation of a specific database, known as the Health Data Warehouse (HDW; *Entrepot de Données de Santé* in French).

With more than 7 million patients and more than 20 million radiological examinations in digital imaging and communication in medicine (DICOM) format, the AP—HP, via the HDW, is a unique database in Europe for biomedical research in imaging. Around 150,000 MGs are currently available in the HDW (period 2010—2018) for more than 75,000 patients including more than 13,000 with a 2-year follow-up by MG. Moreover, more than 27,000 DBTs are available (period 2015—2018) for more than 26,000 patients. These MGs have been performed on machines from more than three manufacturers corresponding to stringent criteria for image quality in France. Thus, several industrial partnerships have been contracted to train and validate different types of AI algorithms which is the subject of a large ongoing study called EZ mammo.

In this setting, the creation of the HDW database, and specifically for medical imaging data (DICOM data), was associated with the definition of legal and ethical rules that take into account data specificity, especially concerning access of data for internal/external scientists or manufacturers. APHP has created processes to inform patients of the possible use of their data for research purposes and to give them the right to opt out, define the limit of use of data and the rules of access, and define data exploitation rules. A specific scientific and ethics committee has been set up in AP—HP (*i.e.*, the Comité Scientifique et Ethique) composed of healthcare and technical professionals and patients. Their role is to ensure that the rules are adhered to and to evaluate each access request for data collected in the HDW of the AP—HP. As all data are stored in a single structure, direct access to the database is not possible because of the risk of stress in the case of multiple and big requests which could negatively impact the operation of care services. Thus, specific research databases are created by periodically and automatically collecting data. During the collection of research data in HDW, a specific data de-identification process pipeline can be applied. In the EZ mammo study, DL techniques are used to automatically search for words in the imaging and pathology reports in order to constitute cohorts with different prevalences of malignancy.

For DICOM data in the HDW of the AP—HP, this process would require duplicating all DICOM files in PACS. As duplication is not recommended, the solution tested by the HDW of the AP—HP consists of collecting minimal metadata for DICOM like identity, description or modality information. This metadata can be used to select DICOM of interest to build a research cohort for a specific project. Once done, the DICOM data is collected by a PACS to PACS (C-MOVE) process. During this process the DICOM file metadata or pixel will be also de-identified.

At the end of the process the scientist must be able to have access to this research database that is composed of structured data and a DICOM repository. This implies that data will be exported out of the AP—HP server. However, for security and exploitation aspects AP—HP has chosen to not export data and has created a data management and research platform that allows data-scientists to analyze data within a mega-data cluster without exporting them off the secure servers of the AP—HP.

During a research project, the database will undergo considerable transformation and will need to be structured to answer a specific question. Annotations may be added to the images, for example, and these need to be reintegrated in the database that will subsequently be enriched by the various research projects. Generally, the radiologist can annotate information in a private DICOM tag. This means

all DICOM files must be re-sent to PACS and if no annotation management is designed, the private tag annotation may be erased if another annotation is created. In the HDW of the AP—HP, a specific annotation collection web service compatible with a PACS solution, called SPHERE, will be added. At the end of data labeling and research processing, the data could be used to test AI algorithms.

Briefly, the challenges for an adapted mega-data PACS solution is to allow high speed transfer of DICOM files and reduce storage costs, to define a specific data model and structure for medical imaging data research, to find and build solutions to assist and secure the annotation process, and to report and facilitate the creation of re-usable datasets.

What can we expect from individual breast screening? An example of a radiologist's initiative

In France, SENOLOG is a database for individualized breast screening performed in private structures outside of the organized program. This database contains all information from MG reports (clinical history, clinical examination, radiological features and BI-RADS classification) and especially the follow-up of MGs classified as BI-RADS 1 or 2. The first challenge for private radiologists will be to standardize their work methods as there is a huge heterogeneity in the MG units, archiving methods, and report standardization. Recently, an initiative for the radiology community, called DRIM France AI, which consists of a platform dedicated to radiological examination has been designed to train and test AI algorithms in a centralized system. However, several questions about DRIM remain unanswered to date: will the data be anonymized by the radiologist, by DRIM, or by the software provider? How are the radiologists going to achieve data flow? Who will be responsible for recovering and storing the data via a common channel: DRIM or the platform's master installer? If it is DRIM that retrieves the data, besides the major problem of the General Data Protection Regulation and data security, it will require storage servers with huge capacities.

A system should be built that allows people not only to share their medical data with researchers easily and securely, but also to retain control of their data [46]. This method, which is based on the blockchain technology that underlies the cryptocurrency Bitcoin, will soon be put to the test and could constitute a solution to build a massive database that is needed to train DL algorithms. In any case, there should be a complementarity and a partnership between radiologists and AI.

AI Research issues
- To create universal annotation tools.
- To develop an AI tool for virtually managing an increasingly enriched cohort.
- To structure a platform adapted to radiological volume.

Better assessing the risk of breast cancer: AI and personalized programs

French national recommendations for the management of women at high risk of breast cancer were published in 2014 by the Haute Autorité de Santé (HAS) [48]. However, there are no specific breast-screening recommendations for women at intermediate risk of breast cancer, and women at normal risk are largely debated. The future will probably include a personalized approach depending on the individual risk of developing a breast cancer. For women with the lowest risk of breast cancer, the interval between two rounds of mammographic screening might increase whereas for women with a higher risk, screening might be performed annually. To deal with this issue, two international trials have begun: the WISDOM TRIAL in the USA [49] and the MyPEBS in Europe, funded by a European H2020 grant [50]. The main endpoint of the European study is the ability to decrease the number of stage 2 or higher cancers detected by a personalized breast screening in comparison with the conventional approach. The trial will include 85,000 women between 40 and 70 years without any risk factors of breast cancer in five countries (France, Belgium, Italy, Israel, and the UK). In France, 20,000 women will be included. The women will be randomized in two arms: a standard arm (screening according to each country's recommendation), and a risk-stratified arm. The risk will be evaluated with the BCSC score or with the IBIS tool (Tyrer-Cuzick model) for women with more than one first-line first-degree relative with breast cancer. Women in this arm will be invited to perform a score of 300 polymorphisms and a risk prediction score. The final risk score will determine the strategy. This risk prediction model, based on AI called MammoRisk®, was developed by Predlife and includes first-degree breast cancer, personal history of breast biopsy, and breast density. In this setting, the software algorithm was developed using a concept of Manhattan distance to compare a patient's mammographic image to reference mammograms with an assigned breast mean density category. Reference databases were built from a total of 2289 pairs (CC and MLO views) of 2D FFDM. A validation set of an additional 800 image pairs was evaluated for breast mean density both by the software and seven blinded radiologists specialized in breast imaging. The software showed a substantial agreement with the radiologists' consensus (unweighted kappa = 0.68; 95% CI: 0.64—0.72) when considering the four breast density categories, and an almost perfect agreement (unweighted kappa = 0.84; 95% CI: 0.80—0.88) when considering clinically significant nondense (A—B) and dense (C—D) categories [51].

A third risk-based screening trial (the PROCAS study) was conducted in the UK between 2011 and 2013 and included women who presented for their 3-yearly MG [52]. Overall, 53,596 women were recruited (representing 37% of the 68% of the women attending their national 3-yearly screening). Of these, 10,000 women participated in a saliva DNA collection study. Risk data comprised age at menarche/menopause, hormone replacement therapy, a family history of breast cancer, weight/height, and breast biopsies. Ten-year risks were identified. Breast density assessment was done by visual analog assessment of the MG, and with two software tools (Volpara and Quantra). Eighteen single

nucleotide polymorphisms were identified and 632 prospective breast cancers occurred in 53,184 women. The best predictor for breast cancer was the visual analog assessment of breast density (OR: 3.59). Using a mammogram density-adjusted TC model in PROCAS improved risk stratification (AUC = 0.6) and identified significantly higher rates (4.7 per 10,000 vs. 1.3 per 10,000; $P < 0.001$) of high-stage cancers in women with an above-average risk of breast cancer. The model performed particularly well in predicting higher stage 2+ invasive cancer. This combined approach using the Tyrer-Cuzick model, mammographic density assessment and polygenic risk score provides an accurate risk stratification, particularly for poor prognosis cancers [53].

Yala et al. recently published a breast cancer risk model based on DL MG on 88,994 consecutive screening MGs in 39,571 women. In this study, the authors combined traditional risk factors and mammograms in a hybrid DL model and demonstrated that this model was better (AUCs of 0.70) than the Tyrer-Cuzick model (0.62; $P < 0.001$) or a risk-factor-based logistic regression model (0.67; $P = 0.01$) [54].

AI research issues

- To develop an AI model that improves the definition of breast risk assessment to select patients for different screening modalities (HR+ patients, breast density, B3 lesions) — better define intermediate-risk patients.
- To build an AI model that integrates all epidemiological and clinical data (physical activity) to improve personalized screening programs.

Conclusion

The main question is how will the radiologist cohabit with AI in the future? Maybe we have to imagine this as an airplane with a human or virtual pilot. The radiologist will always be the last bastion to AI errors and must be aware of the strengths and weaknesses of AI so that the performance of the two together will be stronger than either taken individually. This also raises the question of training for residents: like simulators that allow pilots to train for critical situations, will the radiologist of tomorrow be even better trained on the most complicated cases? The huge database will give residents the opportunity to access all the mammogram to train themselves. Finally, senior radiologists will be able to annotate mammograms rated normal by AI resulting in a stronger, better and faster assessment than the machine alone.

Ethical statements

Informed consent and patient details

The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s).

Funding

This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

CRediT authorship contribution statement

All authors performed Writing—Review & Editing.

In addition, I.T.N performed also Supervision.

Disclosure of interest

Conflicts of interest ITN (GE, Hologic, Siemens, Bard, Guerbet, Samsung, Canon), CB (Icad, GE). The authors declare that they have no competing interest.

References

[1] SFR-IA Group, CERF. French Radiology Community. Artificial intelligence and medical imaging 2018: French Radiology Community white paper. Diagn Interv Imaging 2018;99:727—32.
[2] Beregi JP, Zins M, Masson JP, Cart P, Bartoli JM, Silberman B, et al. Radiology and artificial intelligence: an opportunity for our specialty. Diagn Interv Imaging 2018;99:677—8.
[3] Santé Publique France. Dossier thématique Cancer du sein; 2019 https://www.santepubliquefrance.fr/maladies-et-traumatismes/cancers/cancer-du-sein.
[4] Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331:1493—9.
[5] Trister AD, Buist DSM, Lee CI. Will machine learning tip the balance in breast cancer screening? JAMA Oncol 2017;3:1463—4.
[6] Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007;356:1399—409.
[7] Lehman CD, Wellman RD, Buist DSM. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 2015;175:1828—37.
[8] Codari M, Schiaffino S, Sardanelli F, Trimboli RM. Artificial intelligence for breast MRI in 2008-2018: a systematic mapping review. AJR Am J Roentgenol 2019;212:280—92.
[9] Erickson BJ, Korfiatis P, Kline TL, Akkus Z, Philbrick K, Weston AD. Deep learning in radiology: does one size fit all? J Am Coll Radiol 2018;15:521—6.
[10] Tong C, Liang B, Li J, Zheng Z. A deep automated skeletal bone age assessment model with heterogeneous features learning. J Med Sys 2018;42:249.
[11] Jiang F, Jiang Y, Zhi H. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2:230—43.
[12] Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature 2000;406:747—52.
[13] Skaane P, Bandos AI, Niklason LT, Sebuødegård S, Østerås BH, Gullien R, et al. Digital mammography versus digital mammog-

raphy plus tomosynthesis in breast cancer screening: the Oslo Tomosynthesis Screening Trial. Radiology 2019;291:23—30.
[14] Brnić Z, Hebrang A. Breast compression and radiation dose in two different mammographic oblique projections: 45 and 60 degrees. Eur J Radiol 2001;40:10—5.
[15] McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiol Biomarkers Prev 2006;15:1159—69.
[16] Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, et al. Mammographic density and the risk and detection of breast cancer. N Engl J Med 2007;356:227—36.
[17] Brandt KR, Scott CG, Ma L, Mahmoudzadeh AP, Jensen MR, Whaley DH, et al. Comparison of clinical and automated breast density measurements: implications for risk prediction and supplemental screening. Radiology 2016;279:710—9.
[18] Kerlikowske K, Zhu W, Tosteson AN, Sprague BL, Tice JA, Lehman CD, et al. Identifying women with dense breasts at high risk for interval cancer: a cohort study. Ann Intern Med 2015;162:673—81.
[19] Sprague BL, Conant EF, Onega T, et al. Variation in mammographic breast density assessments among radiologists in clinical practice: a multicenter observational study. Ann Intern Med 2016;165:457—64.
[20] Breast Imaging Reporting & Data System. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads. Accessed July 24, 2019.
[21] Kinkel K. The never-ending success story of BI-RADS. Diagn Interv Imaging 2017;98:177—8.
[22] Spak DA, Plaxco JS, Santiago L, Dryden MJ, Dogan BE. BI-RADS® fifth edition: a summary of changes. Diagn Interv Imaging 2017;98:179—90.
[23] Ciritsis A, Rossi C, Vittoria De Martini I. Determination of mammographic breast density using a deep convolutional neural network. Br J Radiol 2018:20180691.
[24] Lehman CD, Yala A, Schuster T, Dontchos B, Bahl M, Swanson K, et al. Mammographic breast density assessment using deep learning: clinical implementation. Radiology 2019;290:52—8.
[25] Moshina N, Sebuødegård S, Lee CI, Akslen LA, Tsuruda KM, Elmore JG, et al. Automated volumetric analysis of mammographic density in a screening setting: worse outcomes for women with dense breasts. Radiology 2018;288:343—52.
[26] Kontos D, Winham SJ, Oustimov A, Pantalone L, Hsieh MK, Gastounioti A, et al. Radiomic phenotypes of mammographic parenchymal complexity: toward augmenting breast density in breast cancer risk assessment. Radiology 2019;290:41—9.
[27] Keen JD, Keen JM, Keen JE. Utilization of computer-aided detection for digital screening mammography in the United States, 2008 to 2016. J Am Coll Radiol 2018;15:44—8.
[28] Birdwell RL, Ikeda DM, O'Shaughnessy KF, Sickles EA. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. Radiology 2001;219:192—202.
[29] Mayo RC, Kent D, Sen LC, Kapoor M, Leung JWT, Watanabe AT. Reduction of false-positive markings on mammograms: a retrospective comparison study using an artificial intelligence-based CAD. J Digit Imaging 2019;32:618—24.
[30] Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 2019, http://dx.doi.org/10.1093/jnci/djy222.
[31] Gilbert FJ, Astley SM, Gillan MG, Agbaje OF, Wallis MG, James J, et al. Single reading with computer-aided detection for screening mammography. N Engl J Med 2008;359:1675—84.
[32] Hupse R, Samulski M, Lobbes MB, Mann RM, Mus R, den Heeten GJ, et al. Computer-aided detection of masses at mammography: interactive decision support versus prompts. Radiology 2013;266:123—9.
[33] Rodríguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology 2019;290:305—14.
[34] Arleo EK, Saleh M, Rosenblatt R. Lessons learned from reviewing breast imaging malpractice cases. J Am Coll Radiol 2014;11:1186—8.
[35] Lei J, Yang P, Zhang L, Wang Y, Yang K. Diagnostic accuracy of digital breast tomosynthesis versus digital mammography for benign and malignant lesions in breasts: a meta-analysis. Eur Radiol 2014;24:595—602.
[36] Houssami N, Lång K, Bernardi D, Tagliafico A, Zackrisson S, Skaane P. Digital breast tomosynthesis (3D-mammography) screening: a pictorial review of screen-detected cancers and false recalls attributed to tomosynthesis in prospective screening trials. Breast 2016;26:119—34.
[37] Yun SJ, Ryu CW, Rhee SJ, Ryu JK, Oh JY. Benefit of adding digital breast tomosynthesis to digital mammography for breast cancer screening focused on cancer characteristics: a meta-analysis. Breast Cancer Res Treat 2017;164:557—69.
[38] Phi XA, Tagliafico A, Houssami N, Greuter MJW, de Bock GH. Digital breast tomosynthesis for breast cancer screening and diagnosis in women with dense breasts - a systematic review and meta-analysis. BMC Cancer 2018;18:380.
[39] Marinovich ML, Hunter KE, Macaskill P, Houssami N. Breast cancer screening using tomosynthesis or mammography: a meta-analysis of cancer detection and recall. J Natl Cancer Inst 2018;110:942—9.
[40] Marinovich ML, Macaskill P, Bernardi D, Houssami N. Systematic review of agreement between tomosynthesis and pathologic tumor size for newly diagnosed breast cancer and comparison with other imaging tests. Expert Rev Med Devices 2018;15:489—96.
[41] Samala RK, Chan H-P, Hadjiiski L, Helvie MA, Wei J, Cha K. Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography. Med Phys 2016;43:6654.
[42] Fan M, Li Y, Zheng S, Peng W, Tang W, Li L. Computer-aided detection of mass in digital breast tomosynthesis using a faster region-based convolutional neural network. Methods 2019;166:103—11.
[43] Zackrisson S, Lång K, Rosso A, Johnson K, Dustler M, Förnvik D, et al. One-view breast tomosynthesis versus two-view mammography in the Malmö Breast Tomosynthesis Screening Trial (MBTST): a prospective, population-based, diagnostic accuracy study. Lancet Oncol 2018;19:1493—503.
[44] Quelles performances pour le programme de dépistage organisé du cancer du sein en France ? /liste-des-actualites/quelles-performances-pour-le-programme-de-depistage-organise-du-cancer-du-sein-en-france. (https://www.santepubliquefrance.fr/les-actualites/2019/quelles-performances-pour-le-programme-de-depistage-organise-du-cancer-du-sein-en-france). Accessed July 24, 2019.
[45] Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Teuwen J, Broeders M, Gennaro G, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. Eur Radiol 2019;29:4825—32.
[46] Maxmen A. AI researchers embrace Bitcoin technology to share medical data. Nature 2018;555:293—4.
[47] Fenton JJ, Lee CI, Xing G, Baldwin LM, Elmore JG. Computer-aided detection in mammography: downstream effect on

diagnostic testing, ductal carcinoma in situ treatment, and costs. JAMA Intern Med 2014;174:2032—4.
[48] Dépistage du cancer du sein en France : identification des femmes à haut risque et modalités de dépistage. Haute Autorité de Santé. (https://www.has-sante.fr/jcms/c_1741170/fr/depistage-du-cancer-du-sein-en-france-identification-des-femmes-a-haut-risque-et-modalites-de-depistage). Accessed July 24, 2019.
[49] Esserman LJ. The WISDOM Study: breaking the deadlock in the breast cancer screening debate. NPJ Breast Cancer 2017;3:34.
[50] Delaloge S, Bachelot T, Bidard FC, Espie M, Brain E, Bonnefoi H, et al. Breast cancer screening: on our way to the future. Bull Cancer 2016;103:753—63.
[51] Balleyguier C, Arfi-Rouche J, Boyer B, Gauthier E, Helin V, Loshkajian A, et al. A new automated method to evaluate 2D mammographic breast density according to BI-RADS® Atlas Fifth Edition recommendations. Eur Radiol 2019;29:3830—8.
[52] van Veen EM, Brentnall AR, Byers H, Harkness EF, Astley SM, Sampson S, et al. Use of single-nucleotide polymorphisms and mammographic density plus classic risk factors for breast cancer risk prediction. JAMA Oncol 2018;4:476—82.
[53] Evans DGR, Harkness EF, Brentnall AR, van Veen EM, Astley SM, Byers H, et al. Breast cancer pathology and stage are better predicted by risk stratification models that include mammographic density and common genetic variants. Breast Cancer Res Treat 2019;176:141—8.
[54] Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology 2019;292:60—6.