JAMA Network Open

Original Investigation | Imaging

Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms

Thomas Schaffter, PhD; Diana S. M. Buist, PhD, MPH; Christoph I. Lee, MD, MS; Yaroslav Nikulin, MS; Dezső Ribli, MSc; Yuanfang Guan, PhD; William Lotter, PhD; Zequn Jie, PhD; Hao Du, BEng; Sijia Wang, MSc; Jiashi Feng, PhD; Mengling Feng, PhD; Hyo-Eun Kim, PhD; Francisco Albiol, PhD; Alberto Albiol, PhD; Stephen Morrell, B Bus Sc, MiF, M Res; Zbigniew Wojna, MSI; Mehmet Eren Ahsen, PhD; Umar Asif, PhD; Antonio Jimeno Yepes, PhD; Shivanthan Yohanandan, PhD; Simona Rabinovici-Cohen, MSc; Darvin Yi, MSc; Bruce Hoff, PhD; Thomas Yu, BS; Elias Chaibub Neto, PhD; Daniel L. Rubin, MD, MS; Peter Lindholm, MD, PhD; Laurie R. Margolies, MD; Russell Bailey McBride, PhD, MPH; Joseph H. Rothstein, MSc; Weiva Sieh, MD, PhD; Rami Ben-Ari, PhD; Stefan Harrer, PhD; Andrew Trister, MD, PhD; Stephen Friend, MD, PhD; Thea Norman, PhD; Berkman Sahiner, PhD; Fredrik Strand, MD, PhD; Justin Guinney, PhD; Gustavo Stolovitzky, PhD; and the DM DREAM Consortium

Abstract

IMPORTANCE Mammography screening currently relies on subjective human interpretation. Artificial intelligence (AI) advances could be used to increase mammography screening accuracy by reducing missed cancers and false positives.

OBJECTIVE To evaluate whether AI can overcome human mammography interpretation limitations with a rigorous, unbiased evaluation of machine learning algorithms.

DESIGN, SETTING, AND PARTICIPANTS In this diagnostic accuracy study conducted between September 2016 and November 2017, an international, crowdsourced challenge was hosted to foster AI algorithm development focused on interpreting screening mammography. More than 1100 participants comprising 126 teams from 44 countries participated. Analysis began November 18, 2016.

MAIN OUTCOMES AND MEASUREMENTS Algorithms used images alone (challenge 1) or combined images, previous examinations (if available), and clinical and demographic risk factor data (challenge 2) and output a score that translated to cancer yes/no within 12 months. Algorithm accuracy for breast cancer detection was evaluated using area under the curve and algorithm specificity compared with radiologists' specificity with radiologists' sensitivity set at 85.9% (United States) and 83.9% (Sweden). An ensemble method aggregating top-performing AI algorithms and radiologists' recall assessment was developed and evaluated.

RESULTS Overall, 144 231 screening mammograms from 85 580 US women (952 cancer positive ≤12 months from screening) were used for algorithm training and validation. A second independent validation cohort included 166 578 examinations from 68 008 Swedish women (780 cancer positive). The top-performing algorithm achieved an area under the curve of 0.858 (United States) and 0.903 (Sweden) and 66.2% (United States) and 81.2% (Sweden) specificity at the radiologists' sensitivity, lower than community-practice radiologists' specificity of 90.5% (United States) and 98.5% (Sweden). Combining top-performing algorithms and US radiologist assessments resulted in a higher area under the curve of 0.942 and achieved a significantly improved specificity (92.0%) at the same sensitivity.

CONCLUSIONS AND RELEVANCE While no single AI algorithm outperformed radiologists, an ensemble of AI algorithms combined with radiologist assessment in a single-reader screening environment improved overall accuracy. This study underscores the potential of using machine

(continued)

Key Points

Question How do deep learning algorithms perform compared with radiologists in screening mammography interpretation?

Findings In this diagnostic accuracy study using 144 231 screening mammograms from 85 580 women from the United States and 166 578 screening mammograms from 68 008 women from Sweden, no single artificial intelligence algorithm outperformed US community radiologist benchmarks; including clinical data and prior mammograms did not improve artificial intelligence performance. However, combining best-performing artificial intelligence algorithms with single-radiologist assessment demonstrated increased specificity.

Meaning Integrating artificial intelligence to mammography interpretation in single-radiologist settings could yield significant performance improvements, with the potential to reduce health care system expenditures and address resource scarcity experienced in population-based screening programs.

+ Invited Commentary

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

learning methods for enhancing mammography screening interpretation.

JAMA Network Open. 2020;3(3):e200265.
Corrected on March 30, 2020. doi:10.1001/jamanetworkopen.2020.0265

Introduction

Mammography screening is one of the most widely deployed tools for early breast cancer detection and has been shown to decrease mortality in multiple randomized clinical trials.[1] However, screening mammography is imperfect with 1 in 8 cancers missed at time of interpretation in US community practice.[2] Roughly 9% to 10% of the 40 million US women who undergo routine breast screening each year are recalled for additional diagnostic imaging; only 4% to 5% of women recalled are ultimately diagnosed as having breast cancer.[2] These false positives lead to preventable harms, included patient anxiety, benign biopsies, and unnecessary intervention or treatment.[3] High false-positive rates incur significant resources and contribute to the annual $10 billion mammography screening costs in the United States.[4]

Currently, mammograms are interpreted by radiologists and rely on human visual perception to identify relevant traits,[5] leaving its benefit dependent on subjective human interpretation to maximally extract all diagnostic information from the acquired images.[6] In 1998, computer-assisted detection software was developed for mammography with the hopes of improving radiologist performance; however, computer-assisted detection has not improved interpretive accuracy.[7,8] Recent deep learning advances, and the ever increasing large computational power and digital mammography (DM) data availability, renewed the interest in evaluating whether more sophisticated models based on quantitative imaging features can match or even outperform human interpretation alone.[9-17] Such efforts could aid in improving specificity and overall performance in single-radiologist settings. In double-radiologist interpretive settings such as in Europe, highly accurate algorithms could alleviate the person power needed for double-radiologist interpretation and consensus.

Throughout the last decade, crowdsourced competitions or challenges have been popularized as highly effective mechanisms for engaging the international scientific community to solve complex scientific problems.[18,19] The Dialogue on Reverse Engineering Assessment and Methods (DREAM) initiative has run dozens of biomedical challenges, establishing robust and objective computational benchmarks in multiple disease areas and across multiple data modalities.[19] This report describes the DM DREAM challenge, which was designed to develop and validate breast cancer detection algorithms to determine whether machine learning methods applied to mammography data can improve screening accuracy.[20,21]

Methods

The study followed the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline.[22] We conducted an international crowdsourced challenge to assess whether artificial intelligence (AI) algorithms could meet or beat radiologists' interpretive screening mammography performance.[21] The challenge asked participants to develop algorithms inputting screening mammography data and outputting a score representing the likelihood that a woman will be diagnosed with breast cancer within the next 12 months (eAppendix 2 in the Supplement). Digital mammogram images included different views (eTable 2 in the Supplement) from the most recent screening examination. Subchallenge 2 provided access to images for the current and, when available, previous screening examinations, as well as clinical and demographic risk factor information typically available to interpreting radiologists (eAppendix 3 and eTable 1 in the Supplement).

The DM DREAM challenge was hosted between November 2016 and November 2017. There were 4 phases (eAppendix 1 and eFigure 1 in the Supplement): open phase (September 2016-November 2016), leaderboard phase (November 2016-March 2017), and validation phase (March 2017-May 2017), which together constitute the competitive phase and the community phase (July 2017-November 2017). A first data set was provided by Kaiser Permanente Washington (KPW) (eAppendix 5 in the Supplement), which was used during the competitive and community phases. A second data set was provided by the Karolinska Institute (KI) in Sweden (eAppendix 5 in the Supplement), which was only used for trained algorithm validation. To protect data privacy, both data sets were securely protected behind a firewall and were not directly accessible to challenge participants, who had to submit their algorithms for automated training and testing behind the firewall (eAppendix 7 in the Supplement).

In the competitive phase, the KPW data set was randomly split into 3 data sets matched on age, body mass index, and race/ethnicity (eAppendix 3 in the Supplement): leaderboard phase training (50%), leaderboard phase evaluation (20%) (eTable 3 in the Supplement), and final evaluation data set (30%) (eTable 4 in the Supplement). The leaderboard phase allowed competitors to train algorithms using the KPW training data or external (public or private) data and submit their trained algorithms for evaluation. To minimize overfitting, a maximum of 9 submissions per team were scored in the leaderboard data set with publicly posted results.[20] During the validation phase, participants were scored using the KPW final evaluation data set (**Figure 1**A) using area under the curve (AUC) and partial AUC as evaluation metrics (eAppendix 6 in the Supplement) assessed at the examination level, on which the final ranking of team performances was determined.

The 8 top-performing competitive phase teams were invited to collaborate during the community phase to further refine their algorithms (Figure 1B). The output of this phase was an ensemble model, consisting of a weighted aggregation of algorithm predictions into a new algorithm called the *challenge ensemble method* (CEM). The CEM model was further integrated with the radiologists' assessment into another ensemble model called the *CEM+R model*. The CEM and CEM+R models were trained using the training and leaderboard data of the competitive phase, with performance assessed using the KPW and KI final evaluation data sets (Figure 1C and eTable 5 in the Supplement).

Data Sources and Characteristics

Kaiser Permamente Washington provided the primary data set for the challenge, with images linked to curated data as part of the Breast Cancer Surveillance Consortium.[21] These data include prospectively collected patient-reported breast cancer risk factor information linked to breast imaging data, radiologist interpretations and recommendations, and benign and malignant breast tumor biopsy results. An additional independent validation data set was provided by the KI and comprised women screened in the Stockholm region of Sweden between April 2008 to December 2012 with comparable data provided by KPW. Both data sets were deidentified full-field DMs. This collaboration received institutional review board approval at Sage Bionetworks with a waiver of consent and Health Insurance Portability and Accountability Act.

Each screening examination was labeled as cancer negative or cancer positive at the breast level, defined as a tissue biopsy yielding an invasive cancer or ductal carcinoma in situ positive result within 1 year of the screening examination. Images were weakly labeled, meaning the presence or absence of cancer was reported per screening examination but not the actual location of cancer on each image. Breast-level performance was used in the competitive phase, whereas examination-level performance was used in the collaborative phase to make results directly comparable to radiologists' performance (eAppendix 4, eAppendix 6, and eFigure 2 in the Supplement).

Training and Evaluating Models

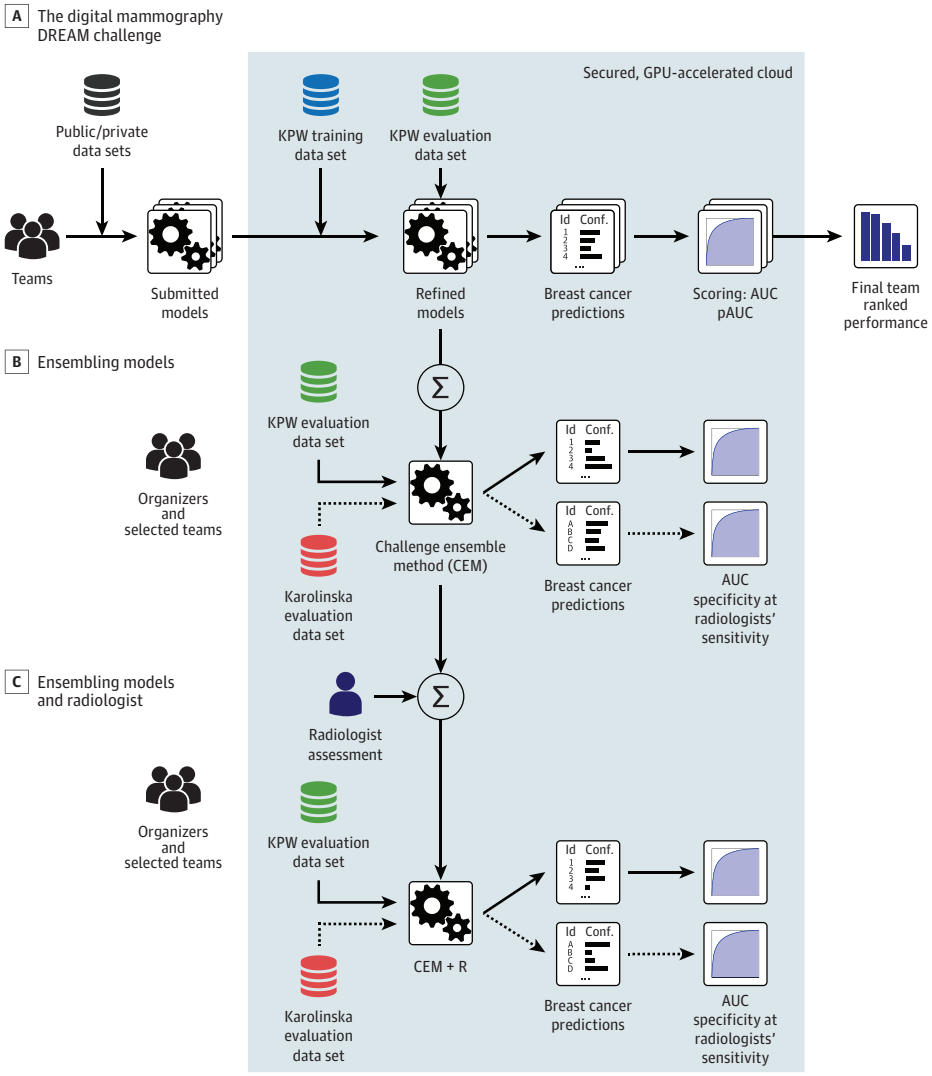
Challenge data providers required the data to be protected from being downloaded, viewed, or otherwise directly accessed by participants. Therefore, we used a model-to-data approach[23] that

required participants to submit containerized algorithms (ie, software that was executed on a participant's behalf in a secure cloud environment) using the IBM and Amazon cloud platforms. Both clouds were donated especially for this challenge (eAppendix 7 and eFigures 3-5 in the Supplement). The model-to-data system was implemented in Synapse (Sage Bionetworks),[24] a web-based platform for running biomedical challenges. Participants were asked to structure their models in the form of a lightweight portable machine image called a Docker image.[25]

Ensemble Method

It has been shown that aggregating the predictions of a set of heterogeneous algorithms can improve performance over the individual algorithms.[26,27] We developed the CEM and CEM+R ensemble classifiers using a meta-learner stacking method[26] (Appendix 8 in the Supplement). Given the data for a screening examination, each individual algorithm outputs a confidence level (a number between 0 and 1) indicative of the likelihood estimated by the algorithm that the woman will develop breast cancer within a year of the screening test. The CEM ensemble algorithm takes as inputs the confidence levels of each of the community phase 8 top-performing methods. The CEM+R takes the same inputs as the CEM ensemble plus the radiologist assessment, represented with a 1 if the woman

Figure 1. Training and Evaluation of Algorithms During the Digital Mammography DREAM Challenge



Training and evaluation Kaiser Permamente Washington (KPW) and Karolinskia Institute data were not directly available to challenge participants; they were stored behind a firewall in a secure cloud (gray box). To access the data, participants submitted models to be run behind the firewall, in the graphics processing unit (GPU)–accelerated cloud (IBM). A, Training and evaluation of models submitted by teams during the Digital Mammography Dialogue on Reverse Engineering Assessment and Methods (DREAM) Challenge. B, A subset of the 8 best models in the evaluation KPW dataset were combined into the Challenge Ensemble Method (CEM), trained using the KPW training set and evaluated in the KPW and Karolinska Institute evaluation datasets. C, We developed a final ensemble method incorporating radiologists' interpretation in a method called CEM+radiologist (CEM+R). AUC indicates area under the curve; pAUC, partial area under the curve.

was recalled or a 0 otherwise. This input information is combined to generate the CEM or the CEM+R ensemble prediction. For the meta-learner classifier, we chose an elastic net regularized logistic regression[28] and the R package caret for construction (R Foundation for Statistical Computing).[29] We trained the meta-learner (ie, tuned the parameters of the logistic regression) on the KPW training data using 10-fold crossvalidation. Final performance assessment was done by applying exactly once the CEM and CEM+R methods to the evaluation KPW and KI data sets.

Statistical Analysis

Analysis began November 18, 2016. We used AUC as our primary metric for evaluating and ranking algorithm performance during the competitive phase. To assess the algorithms' sensitivity and specificity, we computed the radiologists' sensitivity for the data set under study: 85.9% for KPW and 77.1% (single reader) and 83.9% (consensus reading) for KI, which served as the algorithms' specificity prediction threshold. Spearman correlation was used to test for rank distribution similarity of AUCs and specificities between the data sets. We used binomial proportion CIs to compare the significance between specificity of radiologists and CEM+R. The DeLong test of significance was used to compare the area under the receiver operating characteristic curve of 2 correlated receiver operating characteristic curves. One-tailed *P* values had a statistical significance threshold of .05. All analyses were completed in R statistical software, version 3.5.1 (R Foundation for Statistical Computing).

Results

After curation (eAppendix 1 in the Supplement), the KPW data set included 144 231 examinations from 85 580 women, of whom 952 (1.1%) were cancer positive (697 cancers [73.2%] were invasive) (**Table**). The portion of this data set used to evaluate the methods amounts to 30%, which corresponds to the data of 25 657 women, of whom 283 (1.1%) were cancer positive (202 cancers [71.3%] were invasive), while the remaining examinations were used for training. The KI data set provided 166 578 examinations from 68 008 women, of whom 780 (1.1%) were cancer positive (681 cancers [87.3%] were invasive). Women from the KI data set were younger than those in KPW (mean [SD] age, 53.3 [9.4] years vs 58.4 [9.7] years, respectively). Time between screening examinations was bimodal in the 2 data sets and tended to be longer in the KI set (mean [SD] time for first mode: 18.9 [0.9] months; second mode: 24.9 [1.1] months) vs the KPW set (mean [SD] time for first mode: 12.9 [1.7] months; second mode: 24.2 [2.1] months).

Table. Composition of the Data Sets From Kaiser Permanente Washington and Karolinska Institute

Characteristic	No. (%)		
	Kaiser Permanente Washington		Karolinska Institute
	Training	Evaluation	Evaluation
Screening examinations, No.[a]	100 974	43 257	166 578
Women, No.[b]	59 923	25 657	68 008
Women diagnosed with breast cancer within 12 mo of mammogram	669 (1.1)	283 (1.1)	780 (1.1)
Women without a breast cancer diagnosis within 12 mo of mammogram	59 254 (98.9)	25 374 (98.9)	67 228 (98.9)
Invasive breast cancers	495 (74.0)	202 (71.4)	681 (87.3)
Ducal carcinoma in situ	174 (26.0)	81 (28.6)	99 (12.7)
Age, mean (SD), y	58.4 (9.7)	58.4 (9.7)	53.3 (9.4)
BMI, mean (SD)	28.2 (6.9)	28.1 (6.8)	NA
Women with ≥1 prior mammogram	27 165 (45.3)	11 651 (45.4)	50 358 (74.2)
Time since last mammogram, mean (SD), mo			
Mode 1	12.8 (1.7)	12.9 (1.7)	18.9 (0.9)
Mode 2	24.2 (2.1)	24.2 (2.1)	24.9 (1.1)

Abbreviations: BMI, body mass index (calculated as weight in kilograms divided by height in meters squared); NA, not applicable.

[a] Subchallenge 2 provided access to all screening images for the most recent screening examination and, when available, previous screening examinations.

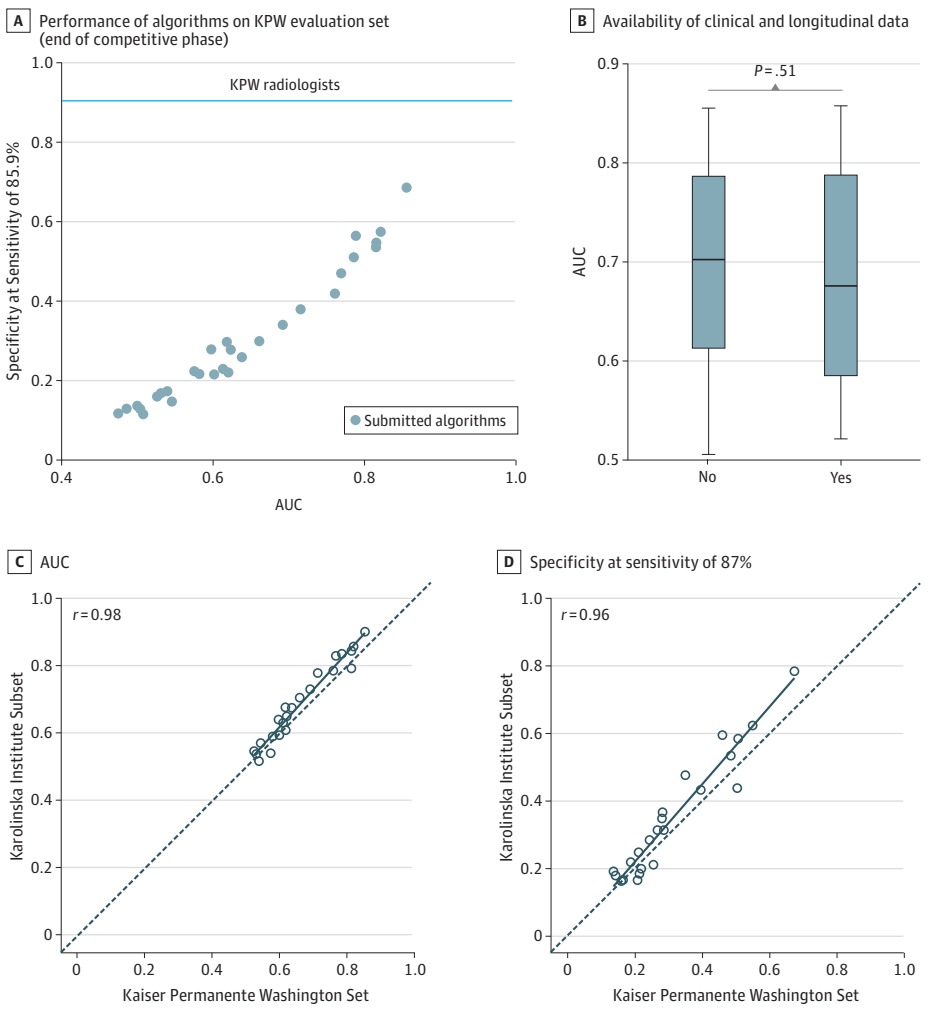
[b] Subchallenge 1 provided access only to the digital mammogram images from the most recent screening examination.

Submitted Algorithms

The DM DREAM challenge included more than 1100 individuals participating, comprising 126 teams from 44 countries (eAppendix 9 in the Supplement). Thirty-one teams submitted their methods for final validation in the competitive phase on the KPW evaluation set (**Figure 2**A). In subchallenge 1, median AUC performance for all teams was 0.611 (interquartile range, 0.54-0.77), with the best-performing method achieving a 0.855 AUC and a specificity of 68.5% at sensitivity of KPW radiologists of 85.9%. The AUC had little improvement when algorithms were able to use clinical, demographic, and longitudinal information (AUC = 0.858 with specificity = 66.3% at sensitivity 85.9%). We observed no improvement across teams in performance measured by AUC (P = .51; t = 0.024) when comparing their results in subchallenge 2 to subchallenge 1 (Figure 2B).

To assess the generalizability of these methods, we evaluated the top 20 methods on the KI data set. The best-performing method on the KPW data achieved top performance on KI data (AUC = 0.903; specificity = 81.2% at the 83.9% KI radiologists' sensitivity). Ranking individual methods on the data sets were found to be significantly correlated by AUC (r = 0.98; 95% CI, 0.95-1.00; P < .001; Figure 2C) and by specificity at Breast Cancer Surveillance Consortium's average sensitivity[2] of 86.9% (r = 0.96; 95% CI, 0.92-1.00; P < .001; Figure 2D).

Figure 2. Performance of the Algorithms Submitted at the End of the Competitive Phase



Individual algorithm performance submitted at the end of the competitive phase on Kaiser Permamente Washington (KPW) and Karolinska Institute data. A, Area under the curve (AUC) and specificity computed at KPW radiologists' sensitivity of 85.9% of 31 methods submitted to the Digital Mammography Digital Mammography Dialogue on Reverse Engineering Assessment and Methods Challenge and evaluated on KPW evaluation set. B, The performance of methods is not significantly higher when clinical, demographic, and longitudinal data are provided. C-D, The AUC and specificity computed at Breast Cancer Surveillance Consortium's sensitivity of 86.9% of methods evaluated on the KPW evaluation set generalize to the Karolinska Institute data.

Ensemble Models and Radiologists' Predictions

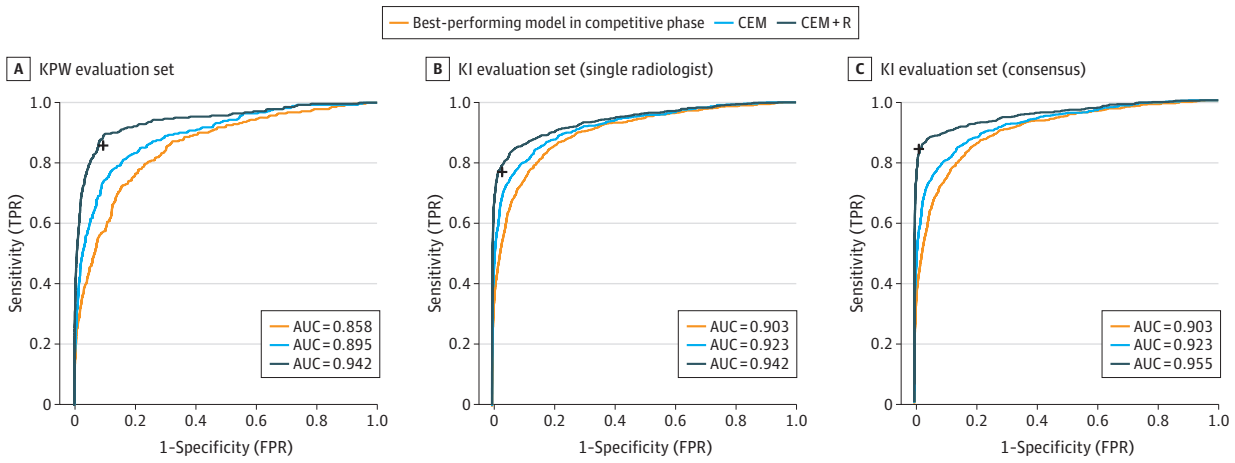
We evaluated whether an ensemble approach could improve overall performance accuracy. Focusing first on the KPW evaluation data, the CEM significantly increased performance (AUC = 0.895; $P < .001$; $z = 6.7$) when compared with the best-performing team (AUC = 0.858) (**Figure 3**A). To compare the dichotomous screening interpretation (recall/no recall) of the radiologist with continuous AI predictions, we examined the specificity using the fixed sensitivity of the radiologist in each of the cohorts. At the KPW radiologist sensitivity of 85.9%, the specificity of the top model, CEM, and radiologist was 66.3%, 76.1%, and 90.5%, respectively (**Figure 4**A). Because AI was consistently inferior to the radiologists' performance, we evaluated whether CEM+R could improve performance. Evaluating the CEM+R on KPW data yielded an AUC of 0.942, with 92% specificity (95% CI, 91.7%-92.3%) (Figure 3A and Figure 4A), higher than the radiologists' specificity of 90.5% (95% CI, 90.1%-90.9%; $P < .001$).

Accuracy assessments of the CEM and CEM+R models were repeated in patient subpopulations by invasive vs ductal carcinoma in situ, age group, and time since examination. We observed that the CEM+R model consistently resulted in a significantly higher specificity compared with the radiologists' assessments alone except for women with ductal carcinoma in situ in KPW (Figure 4A), women with at least 1 previous mammogram done 9 months or earlier in KPW (resulting in a tie between radiologists and CEM+R) (Figure 4A) and women in the oldest age groups in both KPW and KI evaluation data sets (Figure 4A and Figure 4B).

Because the KI data set includes data from a countrywide screening program that completes biennial screening with each mammogram undergoing double reading by 2 radiologists, we used the first KI reader interpretation to directly compare with the US data set. Like the KPW analysis, the CEM method achieved a higher AUC (0.923) compared with the top-performing model AUC (0.903) (Figure 3B) on the KI data set. At the first readers' sensitivity of 77.1%, the specificity of the top model, CEM, and radiologist was 88%, 92.5%, and 96.7%, respectively (Figure 4B). The CEM+R specificity was 98.5% (95% CI, 98.4%-98.6%) (KI AUC: 0.942; Figure 3B and Figure 4B), higher than the radiologist alone specificity of 96.7% (95% CI, 96.6%-96.8%; $P < .001$).

We evaluated whether our ensemble method could be generalized to the double-reading context. Using consensus readings (double reading) from the KI data set, we found sensitivity and specificity of 83.9% and 98.5%, respectively, which outperformed the first readers' sensitivity and specificity of 77.1% and 96.7%, respectively. The CEM+R algorithm, using the consensus readers'

Figure 3. Receiver Operating Characteristic Curves of the Best Individual CEM and CEM+R Methods

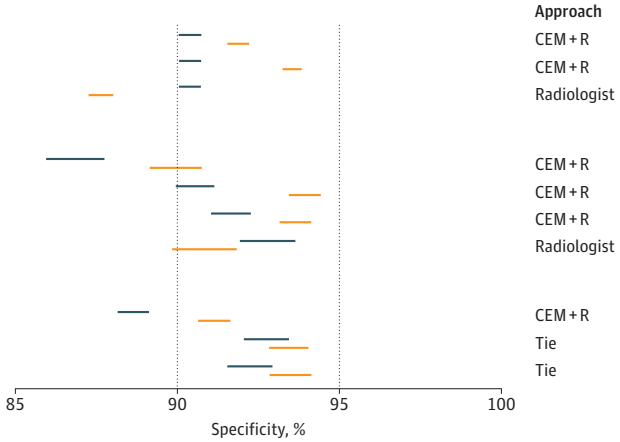


Receiver operating characteristic curves of the best individual method (orange), challenge ensemble method (CEM) (light blue), and challenge ensemble method + radiologist (CEM+R) method (dark blue) in Kaiser Permamente Washington (KPW) (A) and Karolinska Institute (KI) (B-C) data sets for single radiologist and consensus. The black cross reports the sensitivity and specificity achieved by the radiologist(s) in the corresponding cohort. AUC indicates area under the curve; FPR, false-positive rate; TPR, true-positive rate.

Figure 4. Comparison of the Specificity of Radiologist(s) and CEM+R on Kaiser Permanente Washington (KPW) and Karolinska (KI) Data

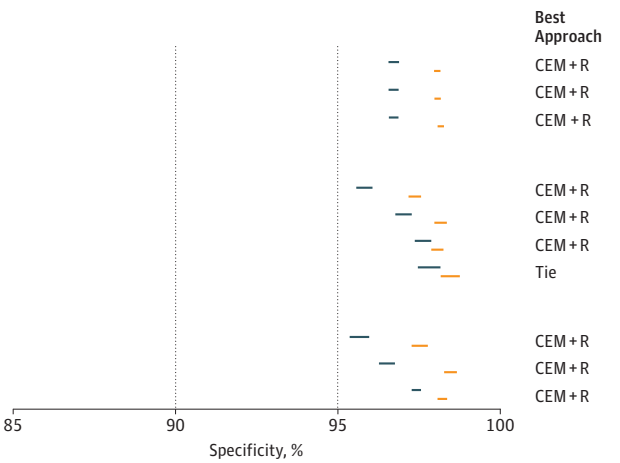
A KPW evaluation set

Characteristic	All Individuals	Cancer-Positive Individuals	Cancer-Negative Individuals	Sensitivity, %	Best Approach
All individuals	25 657	283	25 374	85.9	CEM + R
Invasive and negatives	25 576	202	25 374	81.7	CEM + R
DCIS and negatives	25 455	81	25 374	96.3	Radiologist
Age, y					
40-49	5268	29	5239	79.3	CEM + R
50-59	8655	80	8575	86.3	CEM + R
60-69	8020	114	7906	84.2	CEM + R
≥70	3489	58	3431	93.1	Radiologist
Time since most recent examination					
No previous examination	14 006	164	13 842	86.0	CEM + R
9-21 mo	5990	70	5920	81.4	Tie
≥21 mo	5661	49	5612	91.8	Tie



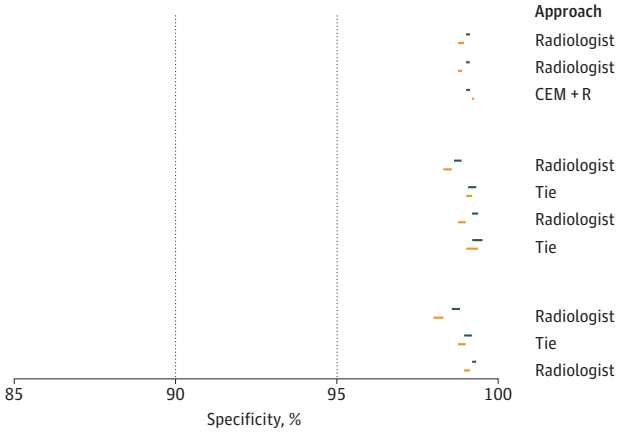
B KI evaluation set (single radiologist)

Characteristic	All Individuals	Cancer-Positive Individuals	Cancer-Negative Individuals	Sensitivity, %	Best Approach
All individuals	67 831	703	67 128	77.1	CEM + R
Invasive and negatives	67 739	611	67 128	76.1	CEM + R
DCIS and negatives	67 220	92	67 128	83.7	CEM + R
Age, y					
40-49	26 324	173	26 151	64.7	CEM + R
50-59	18 351	170	18 181	77.1	CEM + R
60-69	17 509	300	17 209	82.0	CEM + R
≥70	5643	60	5583	88.3	Tie
Time since most recent examination					
No previous examination	17 472	286	17 186	79.0	CEM + R
9-21 mo	15 946	96	15 850	55.2	CEM + R
≥21 mo	34 412	321	34 091	81.9	CEM + R



C KI evaluation set (consensus radiologist)

Characteristic	All Individuals	Cancer-Positive Individuals	Cancer-Negative Individuals	Sensitivity, %	Best Approach
All individuals	67 831	703	67 128	83.9	Radiologist
Invasive and negatives	67 739	611	67 128	83.3	Radiologist
DCIS and negatives	67 220	92	67 128	88.0	CEM + R
Age, y					
40-49	26 324	173	26 151	73.4	Radiologist
50-59	18 351	170	18 181	83.5	Tie
60-69	17 509	300	17 209	88.7	Radiologist
≥70	5643	60	5583	91.7	Tie
Time since most recent examination					
No previous examination	17 472	286	17 186	87.1	Radiologist
9-21 mo	15 946	96	15 850	61.5	Tie
≥21 mo	34 412	321	34 091	87.9	Radiologist



Comparison of the specificity of radiologist(s) and challenge ensemble method + radiologist (CEM+R) for different clinical/demographic conditions on KPW and KI data. For each condition, we report the CI of the specificity of radiologist (blue) and CEM+R (orange) computed at the sensitivity of radiologists. A best performing approach can be identified when the 2 CIs do not overlap. DCIS indicates ductal carcinoma in situ.

calls, did not significantly improve the consensus interpretations alone (98.1% vs 98.5% specificity, respectively). This observation persisted in subpopulation analyses, where consensus radiologist interpretations outperformed the CEM+R ensemble across nearly all subpopulations (Figure 4C).

Top-Performing Algorithmic Methods

The most accurate competitive phase solution was a custom neural network designed for the challenge, initially trained on strongly labeled external data, and subsequently refined using the challenge training data set with 3 teams tied for second (Therapixel model in eAppendix 10 and eFigures 6 and 7 in the Supplement). A second strategy was an adaptation of the Faster R-CNN[30] object detection framework for mammography,[31] which was only trained on external data sets with location annotation for lesions (Dezso Ribli's model in eFigure 8 in the Supplement). Another model used a combination of a higher resolution method to detect calcifications with lower resolution method for masses. A fourth method used a custom neural network with multiple different resolution views of the mammograms[32] (DeepHealth's model in eFigure 9 in the Supplement).

Discussion

The results from our study underscore the promise of using deep learning methods for enhancing the overall accuracy of mammography screening. While no single AI algorithm outperformed US community radiologist benchmarks,[2] an ensemble of AI algorithms combined with single-radiologist assessment was associated with an improved overall mammography performance. Surprisingly, there was no additional improvement in performance when models had access to clinical variables or prior examinations. It is possible that participants did not fully exploit this information, especially the use of prior imaging examinations from the same women. This suggests that future algorithm development would do well to focus on the use of prior images from the same women to detect breast cancer. Furthermore, including additional clinical features not provided in this challenge may result in improved performance.[33] With more than 1100 participants worldwide from 44 countries, more than 1.2 million images representing 310 827 examinations robustly linked to cancer outcomes from 2 population-based screening programs, and a third-party approach for evaluation of AI algorithms on 2 independent data sets, the DM DREAM challenge represents the most objective and rigorous study of deep learning performance for automated mammography interpretation thus far, to our knowledge.

Our trained CEM+R ensemble method used the top AI algorithms resulting from the challenge and the single-radiologist recall assessment available from the KPW training data set. When the CEM+R method was evaluated in 2 independent data sets that included single-radiologist assessments (KPW and KI evaluation sets), the ensemble method had a higher diagnostic accuracy compared with the single radiologist alone. This conclusion is consistent with a recent study demonstrating the AUC of a hybrid model that averaged the probability of malignancies estimated by a neural network and an expert radiologist outperformed the AUC of either.[17] The improvement of the CEM+R method over the radiologist assessment was observed across all women except for the following groups: women 70 years and older in both in the KI and KPW cohorts, women with ductal carcinoma in situ, and women with at least 1 previous screening mammogram done 9 months or more earlier in the KPW cohort. In contrast, when the same ensemble method was evaluated using the consensus interpretation instead of the first radiologist assessment in the Swedish cohort, the ensemble performance did not improve in specificity. This somewhat paradoxical result is likely owing to the fact that the CEM+R ensemble was trained on the single-radiologist interpretation and thereby the importance given by the algorithm to the radiologist's final interpretation may have been less than it would have been if the algorithm had been trained with the consensus interpretations. The performance enhancement of the CEM+R ensemble over the single-reader assessment underscores the potential value of AI as a second digital reader in a single-radiologist environment such as the United States. In the double-reading and consensus environment seen in Sweden and

many other European countries, the addition of AI may not have as great an effect on improving overall diagnostic accuracy, even though it is likely that training an ensemble of AI algorithms and radiologists consensus assessments would improve over the consensus assessments alone. Taken together, our results suggest that adding AI to mammography interpretation in single-radiologist settings could yield significant performance improvements, with the potential to reduce health care system expenditures and address the recurring radiologist person-power issues experienced in population-based screening programs.

This challenge included 2 large population-based mammography data sets from 2 countries that prospectively collected consecutive screening examinations linked to clinical and longitudinal data with robust capture of breast cancer outcomes (≤12 months' follow-up) to inform ground truth. These independent data sets differ by screening interval, cancer composition, radiologist interpretive practices, and some technical parameters (mean compression force), all of which may contribute to the algorithm performance differences between these 2 cohorts. The top-performing algorithm achieved specificities of 66.2% and 81.2% in the KPW and KI data sets, respectively, at the radiologists' sensitivity. We believe that the reason for this difference between the 2 data sets is 2-fold. First, the 2 specificities correspond to 2 different sensitivity operating points in the 2 data sets: a sensitivity of 85.9% in the KPW data set and a sensitivity of 83.9% in the KI data set. Everything else being equal, a lower sensitivity in the KI data set will result in a higher specificity. Second, we believe the longer screening intervals in the KI data set may make the KI data set easier for cancer detection. This is supported by the difference in the AUC between the KPW and KI data sets for the top-performing algorithm (0.858 and 0.903, respectively), despite the fact that the training data set provided in the challenge consisted of an independent data set collected at KPW. Despite the important differences between these cohorts and screening programs, the performance concordance of the algorithms underscores the generalizability of our findings.

To our knowledge, this was the first study in AI and mammography benchmarking requiring teams to submit their algorithms to the challenge organizers, which permitted the evaluation of their algorithms in an unbiased and fully reproducible manner. We believe this to be an important new paradigm for data sharing and cloud-based AI algorithm development,[23] allowing highly sensitive and restricted data such as screening mammograms to be used for public research and AI algorithm assessment. Moreover, as a stipulation of the DREAM organization and challenge funder, our fully documented algorithms are freely available to the larger research community for use and assessment in future studies of automated and semiautomated mammography interpretation.[34]

Limitations

This study has some limitations. We recognize it is currently theoretical to combine radiologist interpretation and AI algorithms. We did not study the interaction of a human interpreter with AI algorithm results and how AI would influence radiologists' final assessments is an area requiring greater research efforts.[5,35] Challenge participants were unable to download and manipulate the larger training and validation image data sets, and mammography images were not strongly labeled, meaning cancer regions were not localized. We observed top-performing challenge teams using external data sets containing spatially annotated tumor information for model development had significantly higher performance in the KPW evaluation data than teams without access to strongly labeled external data (eFigure 10 in the Supplement). During the community phase with additional training data, the faster R-CNN based approach[31] surpassed the top-performing teams' method, which contributed to the improved performance of the ensemble model. This likely reflects that although our data sets are large, they are limited by the relatively small number of positive cases. Consequently, large comparable data sets with spatial annotation will be needed for training original algorithms or vastly larger cohorts will be required to train the next generation of AI models.

Conclusions

In summary, the DM DREAM challenge represents the largest objective deep learning benchmarking effort in screening mammography interpretation to date. An AI algorithm combined with the single-radiologist assessment was associated with a higher overall mammography interpretive accuracy in independent screening programs compared with a single-radiologist interpretation alone. Our study suggests that a collaboration between radiologists and an ensemble algorithm may reduce the recall rate from 0.095 to 0.08, an absolute 1.5% reduction. Considering that approximately 40 million women are screened for breast cancer in the United States each year, this would result in more than half a million women annually who would not have to undergo unnecessary diagnostic work-up. Confirmation of these estimates will require additional validation and testing in clinical settings.

ARTICLE INFORMATION

Accepted for Publication: December 26, 2019.

Published: March 2, 2020. doi:10.1001/jamanetworkopen.2020.0265

Correction: This article was corrected on March 30, 2020, to fix an error in the Author Affiliations.

Corresponding Author: Gustavo Stolovitzky, PhD, IBM Translational Systems Biology and Nanobiotechnology Program, IBM Thomas J. Watson Research Center, 1101 Kitchawan Rd, Yorktown Heights, NY 10598 (gustavo@us.ibm.com).

Author Affiliations: Computational Oncology, Sage Bionetworks, Seattle, Washington (Schaffter, Hoff, Yu, Chaibub Neto, Friend, Guinney); Kaiser Permanente Washington Health Research Institute, Seattle, Washington (Buist); University of Washington School of Medicine, Seattle (Lee); Therapixel, Paris, France (Nikulin); Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary (Ribli); Department of Computational Medicine and Bioinformatics, Michigan Medicine, University of Michigan, Ann Arbor (Guan); DeepHealth Inc, Cambridge, Massachusetts (Lotter); Tencent AI Lab, Shenzhen, China (Jie); National University of Singapore, Singapore (Du); Integrated Health Information Systems Pte Ltd, Singapore (Wang); Department of Electrical and Computer Engineering, National University of Singapore, Singapore (J. Feng); National University Health System, Singapore (M. Feng); Lunit Inc, Seoul, Korea (Kim); Instituto de Física Corpuscular (IFIC), CSIC-Universitat de València, Valencia, Spain (F. Albiol); Universitat Politecnica de Valencia, Valencia, Valenciana, Spain (A. Albiol); Centre for Medical Image Computing, University College London, Bloomsbury, London, United Kingdom (Morrell); Tensorflight Inc, Mountain View, California (Wojna); University of Illinois at Urbana-Champaign, Urbana (Ahsen); IBM Research Australia, Melbourne, Australia (Asif, Jimeno Yepes, Yohanandan, Harrer); IBM Research Haifa, Haifa University Campus, Mount Carmel, Haifa, Israel (Rabinovici-Cohen, Ben-Ari); Stanford University, Stanford, California (Yi); Department of Biomedical Data Science, Radiology, and Medicine (Biomedical Informatics), Stanford University, Stanford, California (Rubin); Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden (Lindholm); Department of Diagnostic, Molecular and Interventional Radiology, Icahn School of Medicine at Mount Sinai, New York, New York (Margolies); Department of Pathology, Molecular and Cell-Based Medicine, Icahn School of Medicine at Mount Sinai, New York, New York (McBride); Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York (Rothstein); Department of Population Health Science and Policy, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York (Sieh); Fred Hutchinson Cancer Research Center, Seattle, Washington (Trister); Bill and Melinda Gates Foundation, Seattle, Washington (Norman); Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, Maryland (Sahiner); Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden (Strand); Breast Radiology, Karolinska University Hospital, Stockholm, Sweden (Strand); IBM Research, Translational Systems Biology and Nanobiotechnology, Thomas J. Watson Research Center, Yorktown Heights, New York (Stolovitzky).

The DM DREAM Consortium Authors: Lester Mackey, PhD; Joyce Cahoon, MS; Li Shen, PhD; Jae Ho Sohn, MD, MS; Hari Trivedi, MD; Yiqiu Shen, MS; Ljubomir Buturovic, PhD; Jose Costa Pereira, PhD; Jaime S. Cardoso, PhD; Eduardo Castro, MSc; Karl Trygve Kalleberg, MD, PhD; Obioma Pelka, MSc; Imane Nedjar, MSc; Krzysztof J. Geras, PhD; Felix Nensa, MD; Ethan Goan, BE; Sven Koitka, MSc; Luis Caballero, PhD; David D. Cox, PhD; Pavitra Krishnaswamy, PhD; Gaurav Pandey, PhD; Christoph M. Friedrich, PhD; Dimitri Perrin, PhD; Clinton Fookes, PhD;

Bibo Shi, PhD; Gerard Cardoso Negrie, MSc; Michael Kawczynski, MS; Kyunghyun Cho, PhD; Can Son Khoo, BSc; Joseph Y. Lo, PhD; A. Gregory Sorensen, MD; Hwejin Jung, PhD.

Affiliations of The DM DREAM Consortium Authors: DeepHealth Inc, Cambridge, Massachusetts (Sorensen); Instituto de Física Corpuscular (IFIC), CSIC-Universitat de València, Valencia, Spain (Caballero); Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York (Pandey); Microsoft New England Research and Development Center, Cambridge, Massachusetts (Mackey); North Carolina State University, Raleigh (Cahoon); Icahn School of Medicine at Mount Sinai, New York, New York (L. Shen); Department of Radiology and Biomedical Imaging, University of California, San Francisco, San Francisco (Sohn); Emory University, Atlanta, Georgia (Trivedi); New York University, New York (Y. Shen, Cho); Clinical Persona, East Palo Alto, California (Buturovic); Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal (Pereira, Cardoso, Castro); KolibriFX, Oslo, Norway (Kalleberg); Department of Computer Science, University of Applied Sciences and Arts, Dortmund, Germany (Pelka, Koitka, Friedrich); Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Essen, Germany (Pelka, Nensa); Biomedical Engineering Laboratory Tlemcen University, Tlemcen, Algeria (Nedjar); Department of Radiology, NYU School of Medicine, New York, New York (Geras, Koitka); Queensland University of Technology, Brisbane, Australia (Goan, Perrin, Fookes); MIT-IBM Watson AI Lab, IBM Research, Cambridge, Massachusetts (Cox); Institute for Infocomm Research, A*STAR, Singapore (Krishnaswamy); Icahn Institute for Data Science and Genomic Technology, New York, New York (Pandey); Carl E. Ravin Advanced Imaging Laboratories, Department of Radiology, Duke University School of Medicine, Durham, North Carolina (Shi); Satalia, London, United Kingdom (Cardoso Negrie); Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco (Kawczynski); University College London, London, United Kingdom (Khoo); Department of Radiology, Duke University School of Medicine, Durham, North Carolina (Lo); Korea University, Seoul, Korea (Jung).

Author Contributions: Drs Schaffter and Stolovitzky had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs Schaffter, Buist, and Lee contributed equally as co–first authors. Drs Guinney and Stolovitzky contributed equally as co–senior authors.

Concept and design: Schaffter, Buist, Lee, Nikulin, Ribli, Jie, J. Feng, M. Feng, Kim, A. Albiol, Yepes, Yi, Yu, Margolies, McBride, Sieh, Ben-Ari, Harrer, Trister, Friend, Norman, Sahiner, Guinney, Stolovitzky, Y. Shen, Pereira, Castro, Pelka, Goan, Pandey, Cardoso Negrie.

Acquisition, analysis, or interpretation of data: Schaffter, Buist, Lee, Nikulin, Ribli, Guan, Lotter, Jie, Du, Wang, J. Feng, M. Feng, F. Albiol, Morrell, Wojna, Ahsen, Asif, Yepes, Yohanandan, Rabinovici-Cohen, Hoff, Chaibub Neto, Rubin, Lindholm, Margolies, McBride, Rothstein, Sieh, Ben-Ari, Harrer, Norman, Sahiner, Strand, Guinney, Stolovitzky, Mackey, Cahoon, L. Shen, Sohn, Trivedi, Buturovic, Pereira, Cardoso, Kalleberg, Nedjar, Geras, Nensa, Koitka, Caballero, Cox, Krishnaswamy, Friedrich, Perrin, Fookes, Shi, Kawczynski, Cho, Khoo, Lo, Sorensen, Jung.

Drafting of the manuscript: Schaffter, Buist, Lee, Nikulin, Ribli, Lotter, J. Feng, M. Feng, Kim, Morrell, Ahsen, Yohanandan, Hoff, Yu, Rubin, Sieh, Ben-Ari, Friend, Strand, Guinney, Stolovitzky, Mackey, Trivedi, Nedjar, Cardoso Negrie, Kawczynski, Khoo, Jung.

Critical revision of the manuscript for important intellectual content: Schaffter, Buist, Lee, Ribli, Guan, Jie, Du, Wang, M. Feng, F. Albiol, A. Albiol, Morrell, Wojna, Asif, Yepes, Rabinovici-Cohen, Yi, Hoff, Chaibub Neto, Lindholm, Margolies, McBride, Rothstein, Sieh, Ben-Ari, Harrer, Trister, Norman, Sahiner, Strand, Guinney, Stolovitzky, Cahoon, L. Shen, Sohn, Y. Shen, Buturovic, Pereira, Cardoso, Castro, Kalleberg, Pelka, Geras, Nensa, Goan, Koitka, Caballero, Cox, Krishnaswamy, Pandey, Friedrich, Perrin, Fookes, Shi, Cho, Lo, Sorensen.

Statistical analysis: Schaffter, Nikulin, Ribli, Lotter, Du, Wang, M. Feng, A. Albiol, Morrell, Wojna, Ahsen, Yohanandan, Chaibub Neto, McBride, Rothstein, Sieh, Ben-Ari, Harrer, Sahiner, Guinney, Stolovitzky, Mackey, Cahoon, L. Shen, Y. Shen, Pereira, Nedjar, Goan, Caballero, Perrin, Cardoso Negrie, Kawczynski, Cho, Khoo.

Obtained funding: Buist, Lee, M. Feng, Trister, Friend, Norman, Guinney, Stolovitzky, Nensa.

Administrative, technical, or material support: Schaffter, Buist, Ribli, Guan, Jie, J. Feng, Morrell, Wojna, Asif, Yepes, Rabinovici-Cohen, Hoff, Yu, Rubin, Lindholm, Margolies, McBride, Rothstein, Sieh, Ben-Ari, Friend, Norman, Strand, Guinney, Stolovitzky, L. Shen, Nensa, Koitka, Cox, Pandey, Sorensen, Jung.

Supervision: Schaffter, Buist, Lee, M. Feng, F. Albiol, Yepes, Margolies, Sieh, Ben-Ari, Harrer, Norman, Guinney, Stolovitzky, Cardoso, Cox, Friedrich, Fookes.

Conflict of Interest Disclosures: Dr Buist reported grants to Kaiser Permanente Washington from the Arnold Foundation, National Institutes of Health, Patient Centered Outcomes Research Institute, and the Agency for Healthcare Research and Quality during the conduct of the study. Dr Lee reports a research grant from GE Healthcare to their institution; textbook royalties from McGraw-Hill, Oxford University Press, and Wolters Kluwer Health; resesarch consulting fees from GRAIL Inc for work on a data safety monitoring board; and personal fees for serving on the editorial board of the *Journal of the American College of Radiology*. Dr Nikulin reports that the solution they submitted for this challenge (which won first place) became the base of the product currently being

developed by Therapixel (where they currently work). Drs Rabinovici-Cohen, Ben-Ari, and Stolovitzky report that IBM, which has employees who work in the area of screening mammography using artificial intelligence, is their employer. Drs Margolies and McBride report grants from Laura and John Arnold Foundation Grant subaward to Icahn School of Medicine at Mount Sinai for the Digital Mammography DREAM challenge during the conduct of the study. Drs Rothstein and Sieh report grants from Laura and John Arnold Foundation during the conduct of the study. Dr Ben-Ari had patents to P201801121US01 and P20170845US01 pending and patents to US10037601 and US9918686B2 issued. Dr Sohn reported grants from National Institute of Biomedical Imaging and Bioengineering during the conduct of the study. Dr Kalleberg reported personal fees from Age Labs AS outside the submitted work. Dr Cox reported an equity stake in DeepHealth Inc. Dr Kawczynski reported personal fees from Genentech and Roche outside the submitted work. Dr Cho reported serving on the advisoty board of Lunit outside the submitted work. Dr Sorensen reported employment with DeepHealth Inc during the conduct of the study; personal fees from Siemens Healthineers, Konica Minolta, Hitachi, and National Institutes of Health; and grant funding from National Institutes of Health, National Science Foundation, and the US Air Force. No other disclosures were reported.

Funding/Support: Funding for the Digital Mammography DREAM challenge was provided by the Laura and John Arnold Foundation. Drs Buist and Lee are supported by the National Cancer Institute (grants HHSN26120110003 and P01CA154292). Dr Lee is also supported by the National Cancer Institute (grant R37CA240403) and the American Cancer Society (grant 126947-MRSG-14-160-01-CPHPS). Dr Guinney is supported by the National Cancer Institute (grant 5U24CA209923).

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

The DM DREAM Consortium: Lester Mackey, PhD (Microsoft Research, Cambridge, MA); Hossein Azizpour, PhD (Division of Robotics, Perception, and Learning, KTH Royal Institute of Technology, Stockholm, Sweden); Joyce Cahoon, MS (North Carolina State University, Raleigh, NC); Kevin Smith, PhD (School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden; Science for Life Laboratory, Solna, Sweden); Bibo Shi, PhD (Carl E. Ravin Advanced Imaging Laboratories, Department of Radiology, Duke University School of Medicine, Durham, NC); Li Shen, PhD (Icahn School of Medicine at Mount Sinai, New York, NY); Jae Ho Sohn, MD, MS (University of California San Francisco, Radiology and Biomedical Imaging, San Francisco, CA); Hari Trivedi, MD (Emory University, Atlanta, GA); Yiqiu Shen (New York University, New York, NY); Ljubomir Buturovic, PhD (Clinical Persona Inc, East Palo Alto, CA); Jose Costa Pereira, PhD (INESC TEC, Porto, Portugal); Jaime S. Cardoso, PhD (INESC TEC and University of Porto, Porto, Portugal); Michael Kawczynski, MS (Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA); Eduardo Castro, MSc (INESC TEC, Campus da Faculdade de Engenharia da Universidade do Porto, Porto, Portugal); Karl Trygve Kalleberg, MD, PhD (KolibriFX, Oslo, Norway); Obioma Pelka, MSc (Department of Computer Science, University of Applied Sciences and Arts, Dortmund, Germany; Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Essen, Germany); Imane Nedjar, MSc (Biomedical Engineering Laboratory Tlemcen University, Tlemcen, Algeria); Kyunghyun Cho, PhD (New York University, New York); Krzysztof J. Geras, PhD (Department of Radiology, NYU School of Medicine, New York, NY); Felix Nensa, MD (Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Essen, Germany); B.E. Ethan Goan, PhD (Queensland University of Technology, Brisbane, Australia); Sven Koitka, MSc (Department of Computer Science, University of Applied Sciences and Arts Dortmund, Dortmund, Germany; Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Essen, Germany); Can Son Khoo, BSc (University College London, London, United Kingdom); Luis Caballero, PhD (Instituto de Física Corpuscular [IFIC], CSIC–Universitat de València, Valencia, Spain); Joseph Y. Lo, PhD (Department of Radiology, Duke University School of Medicine, Durham, North Carolina); David D. Cox, PhD (MIT-IBM Watson AI Lab, IBM Research, Cambridge, MA); Pavitra Krishnaswamy, PhD (Institute for Infocomm Research, A*STAR, Singapore); A. Gregory Sorensen, MD (DeepHealth, Inc, Cambridge MA); Hwejin Jung, PhD (Korea University, Seoul, Republic of Korea); Bibo Shi, PhD (Carl E. Ravin Advanced Imaging Laboratories, Department of Radiology, Duke University School of Medicine, Durham, NC); Gerard Cardoso Negrie, MSc (Satalia, London, United Kingdom); Michael Kawczynski, MS (Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA); Kyunghyun Cho, PhD (New York University, New York, NY); Can Son Khoo, BSc (University College London, London, United Kingdom); Joseph Y. Lo, PhD (Department of Radiology, Duke University School of Medicine, Durham, NC) (eAppendix 11 in the Supplement).

Disclaimer: IBM and Amazon donated computer and storage resources. The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the US Department of Health and Human Services.

REFERENCES

1. Nelson HD, Tyne K, Naik A, Bougatsos C, Chan BK, Humphrey L; U.S. Preventive Services Task Force. Screening for breast cancer: an update for the U.S. Preventive Services Task Force. *Ann Intern Med*. 2009;151(10):727-737, W237-42. doi:10.7326/0003-4819-151-10-200911170-00009

2. Lehman CD, Arao RF, Sprague BL, et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. *Radiology*. 2017;283(1):49-58. doi:10.1148/radiol.2016161174

3. Nelson HD, Pappas M, Cantor A, Griffin J, Daeges M, Humphrey L. Harms of breast cancer screening: systematic review to update the 2009 US Preventive Services Task Force recommendation. *Ann Intern Med*. 2016;164(4):256-267. doi:10.7326/M15-0970

4. O'Donoghue C, Eklund M, Ozanne EM, Esserman LJ. Aggregate cost of mammography screening in the United States: comparison of current practice and advocated guidelines. *Ann Intern Med*. 2014;160(3):145. doi:10.7326/M13-1217

5. Houssami N, Lee CI, Buist DSM, Tao D. Artificial intelligence for breast cancer screening: opportunity or hype? *Breast*. 2017;36:31-33. doi:10.1016/j.breast.2017.09.003

6. Trister AD, Buist DSM, Lee CI. Will machine learning tip the balance in breast cancer screening? *JAMA Oncol*. 2017;3(11):1463-1464. doi:10.1001/jamaoncol.2017.0473

7. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med*. 2007;356(14):1399-1409. doi:10.1056/NEJMoa066099

8. Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson AN, Miglioretti DL; Breast Cancer Surveillance Consortium. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med*. 2015;175(11):1828-1837. doi:10.1001/jamainternmed.2015.5231

9. Obermeyer Z, Emanuel EJ. Predicting the future: big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375(13):1216-1219. doi:10.1056/NEJMp1606181

10. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056

11. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2(3):158-164. doi:10.1038/s41551-018-0195-0

12. Bender E. Challenges: crowdsourced solutions. *Nature*. 2016;533(7602):S62-S64. doi:10.1038/533S62a

13. 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(4):618-624. doi:10.1007/s10278-018-0168-6

14. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology*. 2019;290(2):305-314. doi:10.1148/radiol.2018181371

15. 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. doi:10.1093/jnci/djy222

16. 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. doi:10.1148/ryai.2019180096

17. Wu N, Phang J, Park J, et al. Deep neural networks improve radiologists' performance in breast cancer screening [published online October 7, 2019]. *IEEE Trans Med Imaging*. doi:10.1109/TMI.2019.2945514

18. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2017;60(6):84-90. doi:10.1145/3065386

19. Saez-Rodriguez J, Costello JC, Friend SH, et al. Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat Rev Genet*. 2016;17(8):470-486. doi:10.1038/nrg.2016.69

20. SAGE Bionetworks. Digital mammography DREAM challenge. https://sagebionetworks.org/research-projects/digital-mammography-dream-challenge/. Accessed January 22, 2020.

21. Breast Cancer Surveillance Consortium. https://www.bcsc-research.org/. Accessed August 30, 2019.

22. Bossuyt PM, Reitsma JB, Bruns DE, et al; STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology*. 2015;277(3):826-832. doi:10.1148/radiol.2015151516

23. Guinney J, Saez-Rodriguez J. Alternative models for sharing confidential biomedical data. *Nat Biotechnol*. 2018;36(5):391-392. doi:10.1038/nbt.4128

24. SAGE Bionetworks. Synapse. https://www.synapse.org/. Accessed January 22, 2020.

25. Docker. https://www.docker.com/. Accessed January 22, 2020.

26. Wolpert DH. Stacked generalization. *Neural Netw*. 1992;5(2):241-259. doi:10.1016/S0893-6080(05)80023-1

27. Whalen S, Pandey OP, Pandey G. Predicting protein function and other biomedical characteristics with heterogeneous ensembles. *Methods*. 2016;93:92-102. doi:10.1016/j.ymeth.2015.08.016

28. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1-22. doi:10.18637/jss.v033.i01

29. Kuhn M. *A Short Introduction to the Caret Package*. Published July 16, 2015. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.696.4901&rep=rep1&type=pdf. Accessed January 22, 2020.

30. NIPS Proceedings. Faster R-CNN: towards real-time object detection with region proposal networks. https://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks. Accessed January 22, 2020.

31. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. *Sci Rep*. 2018;8(1):4165. doi:10.1038/s41598-018-22437-z

32. Lotter W, Sorensen G, Cox D. A Multi-scale CNN and curriculum learning strategy for mammogram classification. In: Cardoso M (ed); *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. New York, NY: Springer International Publishing; 2017:169-177. doi:10.1007/978-3-319-67558-9_20

33. Akselrod-Ballin A, Chorev M, Shoshan Y, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology*. 2019;292(2):331-342. doi:10.1148/radiol.2019182622

34. GitHub. Code for classifying mammograms using an ensemble of models from the DREAM D.M. Challenge. https://github.com/Sage-Bionetworks/DigitalMammographyEnsemble. Accessed January 22, 2020.

35. Lee CI, Elmore JG. Artificial intelligence for breast cancer imaging: the new frontier? *J Natl Cancer Inst*. 2019;111(9):875-876. doi:10.1093/jnci/djy223

SUPPLEMENT.