

Breast Cancer Risk Estimation With Artificial Neural Networks Revisited

Discrimination and Calibration

Turgay Ayer, MS^{1,2}; Oguzhan Alagoz, PhD¹; Jagpreet Chhatwal, PhD³; Jude W. Shavlik, PhD⁴; Charles E. Kahn, Jr, MD, MS⁵; and Elizabeth S. Burnside, MD, MPH, MS^{1,2,6}

BACKGROUND: Discriminating malignant breast lesions from benign ones and accurately predicting the risk of breast cancer for individual patients are crucial to successful clinical decisions. In the past, several artificial neural network (ANN) models have been developed for breast cancer-risk prediction. All studies have reported discrimination performance, but not one has assessed calibration, which is an equivalently important measure for accurate risk prediction. In this study, the authors have evaluated whether an artificial neural network (ANN) trained on a large prospectively collected dataset of consecutive mammography findings can discriminate between benign and malignant disease and accurately predict the probability of breast cancer for individual patients. **METHODS:** Our dataset consisted of 62,219 consecutively collected mammography findings matched with the Wisconsin State Cancer Reporting System. The authors built a 3-layer feedforward ANN with 1000 hidden-layer nodes. The authors trained and tested their ANN by using 10-fold cross-validation to predict the risk of breast cancer. The authors used area under the receiver-operating characteristic curve (AUC), sensitivity, and specificity to evaluate discriminative performance of the radiologists and their ANN. The authors assessed the accuracy of risk prediction (ie, calibration) of their ANN by using the Hosmer-Lemeshow (H-L) goodness-of-fit test. **RESULTS:** Their ANN demonstrated superior discrimination (AUC, 0.965) compared with the radiologists (AUC, 0.939; $P < .001$). The authors' ANN was also well calibrated as shown by an H-L goodness of fit P -value of .13. **CONCLUSIONS:** The authors' ANN can effectively discriminate malignant abnormalities from benign ones and accurately predict the risk of breast cancer for individual abnormalities. *Cancer* 2010;000:000-000. © 2010 American Cancer Society.

KEYWORDS: breast cancer, neural networks, risk assessment, discrimination, calibration, computer-assisted diagnosis, computer-assisted radiographic image interpretation, computer-assisted decisions.

Successful breast cancer diagnosis requires systematic image analysis, characterization, and integration of many clinical and mammographic variables.¹ An ideal diagnostic system would discriminate between benign and malignant findings perfectly. Unfortunately, perfect discrimination has not been achieved, so radiologists must make decisions based on their best judgment of breast cancer risk amid substantial uncertainty. When there are numerous interacting predictive variables, ad hoc decision strategies based on experience and memory may lead to errors² and variability in practice.^{3,4} That is why there is intense interest in developing tools that can calculate an accurate probability of breast cancer to aid in making decisions.⁵⁻⁷

Discrimination and calibration are the 2 main components of accuracy in a risk-assessment model.^{8,9} Discrimination is the ability to distinguish benign abnormalities from malignant ones. Although assessing discrimination with area under the receiver-operating characteristic (ROC) curve (AUC) is a popular method in the medical community, it may not be optimal in assessing risk prediction models that stratify individuals into risk categories.¹⁰ In this setting, calibration is also an important tool for accurate risk assessment of individual patients. Calibration measures how well the probabilities generated by the risk prediction model agree with the observed probabilities in the actual population of interest.¹¹

Corresponding author: Elizabeth S. Burnside, MD, MPH, MS, Department of Radiology, University of Wisconsin Medical School, E3 of 311, 600 Highland Avenue, Madison, WI 53792-3252; Fax: (608) 265-1836; eburnside@uwhealth.org

¹Industrial and Systems Engineering Department, University of Wisconsin, Madison, Wisconsin; ²Department of Radiology, University of Wisconsin, Madison, Wisconsin; ³Health Economic Statistics, Merck Research Laboratories, North Wales, Pennsylvania; ⁴Department of Computer Science, University of Wisconsin, Madison, Wisconsin; ⁵Department of Radiology, Medical College of Wisconsin, Milwaukee, Wisconsin; ⁶Department of Biostatistics and Medical Informatics, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin

DOI: 10.1002/cncr.25081, **Received:** July 22, 2009; **Revised:** September 18, 2009; **Accepted:** October 14, 2009, **Published online** in Wiley InterScience (www.interscience.wiley.com)

There is a trade off between discrimination and calibration, and a model typically cannot be perfect in both.¹⁰ In general, risk-prediction models need good discrimination, when their aim is to separate malignant findings from benign ones, and good calibration, when their aim is to stratify individuals into higher or lower risk categories, to aid in decisions and communication.¹¹

Computer models have the potential to help radiologists increase the accuracy of mammography examinations in both detection¹²⁻¹⁵ and diagnosis.¹⁶⁻²⁰ Existing computer models in the domain of breast-cancer diagnosis can be classified under 3 broad categories: prognostic, computer-aided detection (CAD), and computer-aided diagnostic (CADx) models. *Prognostic models*, such as the Gail model,²¹⁻²⁴ use retrospective risk factors such as a woman's age, her personal and family histories of breast cancer, and clinical information to predict breast cancer risk during a time interval in the future for treatment or risk-reduction decisions.²⁴ These models provide guidance for clinical trial eligibility, tailored disease surveillance, and chemoprevention strategies.²⁵ Because risk stratification is of primary interest in prognostic models, the performance of these models is assessed principally by calibration measures.¹¹ *Detection or CAD models*^{12-15,26-28} are developed to assist radiologists in identifying possible abnormalities in radiologic images, leaving the interpretation of the abnormality to the radiologist.²⁹ Because discrimination is most important, and calibration is less critical in detection, the performance of CAD models is typically evaluated in terms of ROC curves.¹¹ *Diagnostic or CADx models*³⁰⁻³⁹ characterize findings from mammograms (eg, size, contrast, shape) identified either by a radiologist or a CAD model²⁹ to help radiologists classify lesions as benign or malignant by providing objective information, such as the risk of breast cancer.⁴⁰ CADx models are similar to prognostic models in 1 way; they estimate the risk of breast malignancy to help physicians and patients improve decisions.²⁹ On the other hand, CADx models differ from prognostic models in the sense that their risk estimation is based on mammography findings and at a single time point (ie, at the time of mammography) to aid in further imaging or intervention decisions. Both discrimination and calibration are important features of a CADx model. High discrimination is needed because helping radiologists to distinguish malignant findings from benign ones is the primary purpose of CADx models.¹¹ In addition, good calibration is needed to stratify risk and communicate

the risk with patients as in the example of prognostic models.¹¹

However, existing CADx studies that use ANNs to assess the risk of breast cancer have ignored calibration and focused only on discrimination ability.^{31,36,38,39} Most of these studies have good discrimination but may be very poorly calibrated.⁴¹ For example, 4 such models report that no cancers would be missed if the threshold to defer biopsy was set to 10%-20%.^{31,35,37,42} By suggesting a threshold in this range to defer biopsy, these models not only substantially exceed the accepted biopsy threshold in clinical practice of 2%,⁴³ but they also indicate a systematic overestimation of malignancy risk. This discrepancy is likely attributable to suboptimal calibration.

In addition, existing studies have several potential limitations that make them impractical for clinical implementation. First, the size of training datasets used for building ANNs in these previous studies has been relatively small (104-1288 lesions)^{31,35,36,38,39} to obtain reliable models. Second, the majority of these studies developed models by using only findings that underwent biopsy,^{30,31,35-37,39} or were referred to a surgeon,³⁸ and excluded other findings in their analysis, which may lead to biased models.

Our research team has developed 2 CADx models that use the same dataset to discriminate malignant mammography findings from benign ones.^{33,34} This study differs from our previous research in 2 different ways. First, this study uses a different modeling technique (an artificial neural network [ANN]) than our previous research, which used logistic regression and a Bayesian network. Second, this study considers calibration, whereas our previous research, like many other CADx models, did not evaluate calibration but only evaluated discrimination.

The purpose of our study is to evaluate whether an ANN trained on a large prospectively collected dataset of consecutive mammography findings can discriminate between benign and malignant disease and accurately predict the probability of breast cancer for individual patients.

MATERIALS AND METHODS

The institutional review board exempted this Health Insurance Portability and Accountability Act (HIPAA)-compliant, retrospective study from requiring informed consent. The data used in this study have been presented in our previous studies^{33,34} and is repeated here for the convenience of the reader.

Table 1. Distribution of Study Population

| Study Population | Malignant (%) | Benign (%) | Total (%) |
|--------------------------|----------------------|-------------------|------------------|
| No. of mammograms | 477 (1) | 48,267 (99) | 48,744 (100) |
| Age groups, y | | | |
| <45 | 66 (13.84) | 9529 (19.74) | 9595 (19.68) |
| 45-49 | 49 (10.27) | 7524 (15.59) | 7573 (15.54) |
| 50-54 | 56 (11.74) | 7335 (15.2) | 7391 (15.16) |
| 55-59 | 71 (14.88) | 6016 (12.46) | 6087 (12.49) |
| 60-64 | 59 (12.37) | 4779 (9.9) | 4838 (9.93) |
| ≥65 | 176 (36.9) | 13,084 (27.11) | 13,260 (27.20) |
| Breast density | | | |
| Predominantly fatty | 61 (12.79) | 7226 (14.97) | 7287 (14.95) |
| Scattered fibroglandular | 201 (42.14) | 19,624 (40.66) | 19,825 (40.67) |
| Heterogeneously dense | 174 (36.48) | 17,032 (35.29) | 17,206 (35.30) |
| Extremely dense tissue | 41 (8.6) | 4385 (9.08) | 4426 (9.08) |
| BI-RADS category | | | |
| 1 | 0 (0) | 21,094 (43.7) | 21,094 (43.28) |
| 2 | 13 (2.73) | 10,048 (20.82) | 10,061 (20.64) |
| 3 | 32 (6.71) | 8520 (17.65) | 8552 (17.54) |
| 0 | 130 (27.25) | 8148 (16.88) | 8278 (16.98) |
| 4 | 137 (28.72) | 364 (0.75) | 501 (1.03) |
| 5 | 165 (34.59) | 93 (0.19) | 258 (0.53) |

BI-RADS indicates Breast Imaging Reporting and Data System.

Data Collection

All of the screening and diagnostic mammograms performed at the Froedtert and Medical College of Wisconsin Breast Care Center between April 5, 1999 and February 9, 2004 were included in our dataset for retrospective evaluation. We consolidated our database in the National Mammography Database (NMD) format, a data format based on the standardized Breast Imaging Reporting and Data System (BI-RADS) lexicon developed by the American College of Radiology (ACR) for standardized monitoring and tracking of patients.^{44,45} The study comprised 48,744 mammograms belonging to 18,269 patients (Table 1).

Each mammogram was prospectively interpreted by 1 of 8 radiologists. Four of these radiologists were general radiologists, 2 of them were fellowship trained in breast imaging, and the other 2 had extensive experience in breast imaging. These radiologists had between 1-35 years of experience interpreting mammography. Each radiologist reviewed 6994 mammograms on average (median, 2924; range, 49-22,219) in our dataset.

Each mammographic finding, if any, was recorded as a unique entry in our database. In case of a negative mammogram, a single entry showing only demographic data (age, personal history, prior surgery, and hormone replacement therapy) and BI-RADS assessment category was entered. If an image had more than 1 reported finding

with only 1 of them being cancer, we considered the other findings as false positives. Throughout the current article, the term “finding” will be used to denote the single record for normal mammograms or each record denoting an abnormality on a mammogram. Both radiologists (for mammography findings) and technologists (for demographic data) used PenRad (Minnetonka, Minn) mammography reporting/tracking data system, which records clinical data in a structured format. (ie, Point-and-click entry of information populates the clinical report and the database simultaneously.) We included in our ANN model all of the demographic risk factors and BI-RADS descriptors that were routinely collected in the practice and predictive of breast cancer (Table 2). We obtained the reading radiologist’s information by merging the PenRad data with the radiology information system at the Medical College of Wisconsin. We could not assign 504 findings to a radiologist during our matching protocol. We elected to keep these unassigned findings in our dataset to maintain its consecutive nature.

We analyzed discrimination and calibration accuracy at the finding level because this is the level at which recall and biopsy decisions are made in clinical practice. We believe this is the level at which computer-assisted models will help radiologists improve performance. However, because conventional analysis of mammographic data is at the mammogram level (where findings from a

Table 2. Variables from the NMD Used in Our ANN

| Variables | Instances |
|----------------------------|---|
| Age groups, y | <45, 45-50, 51-54, 55-60, 61-64, ≥65 |
| Hormone therapy | None, <5 y, >5 y |
| Personal history of BCA | No, yes |
| Family history of BCA | None, minor (nonfirst-degree family members), major (1 or more first-degree family members) |
| Breast density | Predominantly fatty, scattered fibroglandular, heterogeneously dense, extremely dense |
| Mass shape | Circumscribed, ill-defined, microlobulated, spiculated, not present |
| Mass stability | Decreasing, stable, increasing, not present |
| Mass margins | Oval, round, lobular, irregular, not present |
| Mass density | Fat, low, equal, high, not present |
| Mass size | None, small (<3 cm), large (≥3 cm) |
| Lymph node | Present, not present |
| Asymmetric density | Present, not present |
| Skin thickening | Present, not present |
| Tubular density | Present, not present |
| Skin retraction | Present, not present |
| Nipple retraction | Present, not present |
| Skin thickening | Present, not present |
| Trabecular thickening | Present, not present |
| Skin lesion | Present, not present |
| Axillary adenopathy | Present, not present |
| Architectural distortion | Present, not present |
| Prior history of surgery | No, yes |
| Postoperative change | No, yes |
| Microcalcifications | |
| Popcorn | Present, not present |
| Milk | Present, not present |
| Rodlike | Present, not present |
| Eggshell | Present, not present |
| Dystrophic | Present, not present |
| Lucent | Present, not present |
| Dermal | Present, not present |
| Round | Scattered, regional, clustered, segmental, linear ductal |
| Punctate | Scattered, regional, clustered, segmental, linear ductal |
| Amorphous | Scattered, regional, clustered, segmental, linear ductal |
| Pleomorphic | Scattered, regional, clustered, segmental, linear ductal |
| Fine Linear | Scattered, regional, clustered, segmental, linear ductal |
| BI-RADS category | 0, 1, 2, 3, 4, 5 |

NMD indicates National Mammography Database; BCA, breast cancer; BI-RADS, Breast Imaging Reporting and Data System.

single study are combined), we also calculated the cancer detection rate, the early stage cancer detection rate, and the abnormal interpretation rate at the mammogram level for comparison. We specify whether analyses in this study are based on mammograms or findings.

Data obtained from the Wisconsin Cancer Reporting System (WCRS), a statewide cancer registry, was used as our reference standard. The WCRS has been collecting information from hospitals, clinics, and physicians since 1978. The WCRS records demographic information, tumor characteristics (eg, date of diagnosis, primary site, stage of disease), and treatment information for all newly diagnosed breast cancers in the state. Under data exchange agreements, out-of-state cancer registries also provide reports on Wisconsin residents diagnosed in their states. Findings that had matching registry reports of ductal carcinoma in situ or any invasive carcinoma within 12 months of a mammogram date were considered positive. Findings shown to be benign by biopsy or without a registry match within the same time period were considered negative.

Model

We built a 3-layer, feed-forward, neural network by using Matlab 7.4 (Matlab, The Mathworks, Natick, Mass) with a backpropagation learning algorithm⁴⁶ to estimate the likelihood of malignancy. The layers included an input layer of 36 discrete variables (mammographic descriptors, demographic factors, and BI-RADS final assessment categories as entered by the radiologists; Table 2), a hidden layer with 1000 hidden nodes, and an output layer with a single node generating the probability of malignancy for each finding. We designed our ANN to have a large number of hidden nodes, because ANNs with a large number of hidden nodes generalize better than networks with small number of hidden nodes when trained with backpropagation and “early stopping”.⁴⁷⁻⁴⁹ (See Discussion, this article).

To train and test our ANN, we used a standard machine-learning method called 10-fold cross-validation, which ensures that a test sample is never used for training. In our 10-fold cross-validation, the data was divided into 10 subsets that were approximately equal in size. In the first iteration, 9 of these subsets were combined and used for training. The remaining 10th set was used for testing the performance of our ANN on unseen cases. We repeated this process for 10 iterations until all subsets were used once for testing. In addition to 10-fold cross-validation, to assess the robustness of our ANN, we performed the following supplementary analyses: 1) we trained our ANN on the first half of the dataset and tested on the second half, 2) we trained our ANN on the second half of the dataset and tested on the first half.

We used “early stopping (ES)” procedure to prevent our ANN from overfitting and to keep it generalizable to future cases.^{50,51} Generalizability is the ability of a model to demonstrate similar predictive performance on data not used for training but consisting of unseen cases from the same population. A model lacks generalizability when overfitting occurs, a phenomenon whereby the model “memorizes” the cases in the training data but fails to generalize to new data. When overfitting occurs, ANNs obtain spuriously good performance by learning anomalous patterns unique to the training set but generate high error resulting in low accuracy when presented with unseen data.⁵² We performed ES by using a validation (tuning) set, in addition to a training and a testing set, to calculate the network error during training and to stop training early if necessary to prevent overfitting.⁵⁰⁻⁵²

Model Evaluation

We evaluated the discriminative ability of our ANN against radiologists at an aggregate level and at an individual-radiologist level. We plotted the receiver-operator characteristic (ROC) curve for our ANN by using the probabilities generated for all findings by means of our 10-fold cross-validation technique. We constructed the ROC curves for all radiologists individually and in aggregate by using BI-RADS assessment categories assigned by the radiologists to each finding. We ordered BI-RADS assessment categories by the increasing likelihood of malignancy ($1 < 2 < 3 < 4 < 5$) for this purpose. We measured area under the curve (AUC), sensitivity, and specificity to assess the discriminative ability of our ANN and the radiologists (in aggregate and individually). We used a 2-tailed DeLong method⁵³ to measure and compare AUCs because it accounts for correlation between the ROC curves obtained from the same data.

We calculated sensitivity and specificity of our ANN and the radiologists at recommended levels of performance: sensitivity at a specificity of 90% and specificity at a sensitivity of 85%, as they represent the minimal performance thresholds for screening-mammography.⁵⁴ When calculating the sensitivity and specificity of the radiologists, we considered BI-RADS 0, 4, and 5 positive, whereas BI-RADS 1, 2, and 3 were designated negative.⁴⁵ We used 1-tailed McNemar test to compare sensitivity and specificity between the radiologists and our ANN.⁵⁵ A McNemar test accounts for correlation between the sensitivity and specificity ratios and is not defined when the ratios are equal, nor when 1 of the ratios is 0 or 1. We used the Wilson method to generate confidence intervals

for sensitivity and specificity.⁵⁶ We considered $P < .05$ to be the level of statistical significance.

We assessed the calibration of our ANN by calculating the Hosmer-Lemeshow (H-L) goodness-of-fit statistic⁵⁷ and plotting a calibration curve. The H-L statistic compares the observed and predicted risk within risk categories. A lower H-L statistic and a higher P value ($P > .05$) indicate better calibration. For the H-L statistic, the predicted risks of findings were rank-ordered and divided into 10 groups, based on their predicted probability. Within each predicted risk group, the number of predicted malignancies was accumulated against the number of observed malignancies. The H-L statistic was calculated from this 2×10 contingency table. The H-L statistic was then compared with the chi-square distribution, with degrees of freedom equal to 8. We also plotted a calibration curve to visually compare calibration of our ANN to the perfect calibration in predicting breast malignancy risk. In a calibration curve, a line at a 45° angle (line of identity) indicates perfect calibration. Data points to the right of the perfect calibration line represent overestimation of the risk, and those to the left of the line represent underestimation.⁵⁸ Although a calibration curve does not provide a quantitative measure of reliability for probability predictions, it provides a graphical representation of the degree to which predicted probability of malignancy by our ANN corresponds to actual prevalence.^{58,59} The calibration curve shows the ability of the model to enable prediction of probabilities across all ranges of risk.

RESULTS

After matching to the cancer registry, our final matched dataset contained a total of 62,219 findings [510 (0.8%), malignant and 61,709 (99.2%) benign], in 18,269 patients (17,924 women and 345 men). The mean age of the female patients was 56.5 years (range, 17.7-99.1; SD, 12.7). Women were, on average, 2 years younger compared with men, whose mean age was 58.5 years (range, 18.6-88.5; SD, 15.7).

Our analysis at the mammogram level showed that 14% of the mammographic abnormalities occurred predominantly in fatty tissue, 41% in scattered fibroglandular tissue, 36% in heterogeneously dense tissue, and 9% in extremely dense tissue (Table 1). At the findings level, the cancers included 246 masses, 121 microcalcifications, 27 asymmetries, 18 architectural distortions, 86 combinations of findings, and 12 other.

Cancer registry match revealed a detection rate of 8.9 cancers per 1000 mammograms for the radiologists at the mammogram level (432 cancers for 48,744 mammograms—33 patients had more than 1 cancer resulting in 510 total cancers). The abnormal interpretation rate (considering BI-RADS 0, 4, and 5 abnormal) was 18.5% (9037 of 48,744 mammograms). Of all the 432 cancers, 390 had staging information from the cancer registry, and 42 did not. Of the detected cancers with staging information, only 26.7% (104 of 390) had lymph node metastasis, and 71% (277 of 390) were early stage (ie, stage 0 or 1).

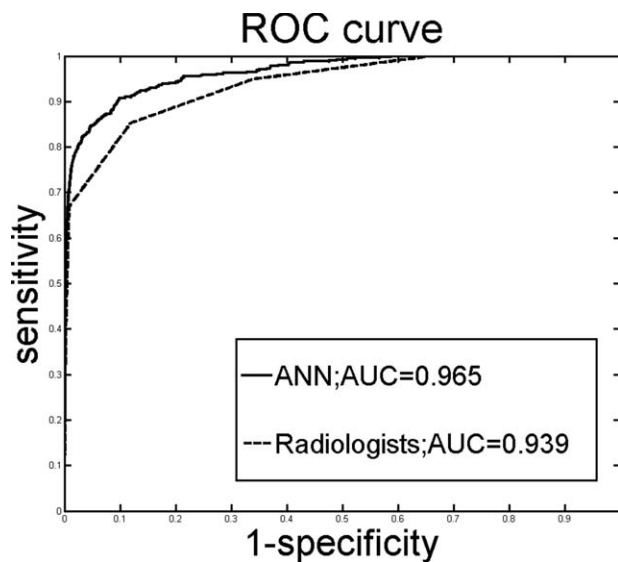


Figure 1. ROC curves were constructed from the output probabilities of our artificial neural network (ANN) and the radiologists' BI-RADS assessment categories. AUC indicates area under the ROC curve.

Following training and testing using 10-fold cross-validation, the AUC of our ANN, 0.965, was significantly higher than that of the radiologists in aggregate, 0.939 ($P < .001$), at the finding level, which implied that our ANN performed better than the radiologists alone in discriminating between benign and malignant findings. The ROC curve of our ANN (aggregate level) dominated the combined ROC curve of all radiologists at all cutoff thresholds (Fig. 1). This trend was preserved when the ANN was trained on the first half of the dataset and tested on the second half (ANN AUC, 0.949; radiologists AUC, 0.926; $P < .001$) or when trained on the second half of the dataset and tested on the first half (ANN AUC, 0.966; radiologists AUC, 0.951; $P < .001$). At the individual radiologists level, 4 of 8 comparisons were not statistically significant (Table 3). Of the 4 significant differences, our ANN outperformed the radiologists in all except a single, low-volume reader (Radiologist 8, Table 3).

At a specificity of 90%, the sensitivity of our ANN was significantly better (90.7% vs 82.2%; $P < .001$) than that of the radiologists (in aggregate; Table 4). Our ANN identified 44 more cancers when compared with the radiologists at this level of specificity (Table 5, part A.). At a fixed sensitivity of 85%, the specificity of our ANN was also significantly better (94.5% vs 88.2%, $P < .001$) than that of the radiologists (in aggregate; Table 4). Our ANN decreased the number of false positives by 3941 when compared with the radiologists' performance at this level of sensitivity (Table 5, part B). In terms of specificity, all statistically significant comparisons revealed the ANN to be superior with the exception of 1 low-volume reader (Radiologist 8 in Table 4). In terms of sensitivity, all statistically significant comparisons revealed the ANN to be

Table 3. Comparison of Radiologist and ANN AUCs

| Rad No. | No. of Benign Findings | No. of Malignant Findings | Cancer Prevalence | Radiologist AUC | ANN AUC | P |
|------------|------------------------|---------------------------|-------------------|-----------------|---------|-------|
| 1 | 3312 | 77 | 0.0227 | 0.954 | 0.956 | .607 |
| 2 | 47 | 1 | 0.0208 | 0.777 | 1 | <.001 |
| 3 | 18953 | 180 | 0.0094 | 0.928 | 0.969 | <.001 |
| 4 | 26690 | 171 | 0.0064 | 0.936 | 0.969 | <.001 |
| 5 | 82 | 0 | 0 | ND | ND | ND |
| 6 | 6796 | 36 | 0.0053 | 0.954 | 0.955 | .903 |
| 7 | 3637 | 29 | 0.0079 | 0.931 | 0.941 | .416 |
| 8 | 1695 | 9 | 0.0053 | 0.940 | 0.873 | .005 |
| Unassigned | 497 | 7 | 0.0139 | 0.995 | 0.998 | .305 |
| Total | 61709 | 510 | 0.0081 | 0.939 | 0.965 | <.001 |

AUC indicates area under the curve; ANN AUC, artificial neural network area under the curve; ND, not defined.

Table 4. Comparison of Radiologist and ANN Sensitivity and Specificity

| Rad No. | No. of Benign Findings | No. of Malignant Findings | Radiologist Sensitivity ^{a,b} | ANN Sensitivity ^{a,b} | P ^c | Radiologist Specificity ^{a,d} | ANN Specificity ^{a,d} | P ^c |
|-------------------------|------------------------|---------------------------|--|--------------------------------|----------------|--|--------------------------------|----------------|
| 1 | 3312 | 77 | 93.5 (84.8, 97.6) | 88.4 (78.4,94.1) | .0625 | 94.4 (93.6, 95.2) | 96.9 (96.4, 97.5) | <.001 |
| 2 | 47 | 1 | NC | NC | NC | NC | NC | NC |
| 3 | 18953 | 180 | 78.3 (71.4, 83.9) | 90.0 (84.5, 93.8) | <.001 | 85.0 (84.4, 85.5) | 95.0 (94.7, 95.3) | <.001 |
| 4 | 26690 | 171 | 82.4 (75.7, 87.6) | 93.0 (87.8, 96.1) | <.001 | 85.6 (85.1, 86.0) | 96.4 (96.1, 96.5) | <.001 |
| 5 | 82 | 0 | NC | NC | NC | NC | NC | NC |
| 6 | 6796 | 36 | 83.3 (66.5, 93.0) | 86.1 (69.7, 94.7) | .999 | 88.4 (87.6, 89.1) | 94.5 (93.9, 95.0) | <.001 |
| 7 | 3637 | 29 | 75.8 (56.0, 88.9) | 72.5 (52.5, 86.5) | .999 | 79.9 (78.6, 81.2) | 86.2 (85.0, 87.2) | <.001 |
| 8 | 1695 | 9 | 77.7 (40.1, 96.0) | 66.7 (30.9, 90.9) | .999 | 86.7 (85.0, 88.3) | 80.7 (78.7, 82.5) | <.001 |
| Unassigned ^e | 497 | 7 | 100.0 (56.1, 100.0) | 100.0 (56.1, 100.0) | ND | 98.3 (96.7, 99.2) | 99.6 (98.4, 99.9) | 0.015 |
| Total | 61709 | 510 | 82.2 (78.5, 85.3) | 90.7 (87.8, 93.0) | <.001 | 88.2 (87.9, 88.5) | 94.5 (94.3, 94.6) | <.001 |

ANN indicates artificial neural network; ND, not defined (McNemar test cannot be computed when the ratio is 1); NC, not calculated as numbers were too small to obtain reliable performance.

^aData in parentheses are 95% confidence intervals.

^bSensitivity calculated at a specificity of 90%.

^cCalculated with McNemar test.

^dSpecificity calculated at a sensitivity of 85%.

^eUnassigned mammographic studies resulting from inability to match studies with radiologists when merging mammography reporting system and institutional radiology information system.

Table 5. Sensitivity and Specificity Results

A. Performance at 90% Specificity

| | True Positive | False Negative |
|--------------|---------------|----------------|
| Radiologists | 419 (400-435) | 91 (75-110) |
| ANN | 463 (449-475) | 47 (36-62) |

B. Performance at 85% Sensitivity

| | False Negative | True Positive |
|--------------|------------------|------------------------|
| Radiologists | 7282 (7126-7441) | 54,427 (54,268-54,583) |
| ANN | 3341 (3232-3454) | 58,368 (58,256-58,477) |

Data are numbers (95% CIs) of cases.

ANN indicates artificial neural network.

superior; however, 1 low-volume reading radiologist demonstrated the opposite trend (Radiologist 1 in Table 4).

The H-L statistic for our ANN was 12.46 ($P = .13$, $df = 8$). The precision of the predicted probabilities is shown graphically in Figure 2. Although the calibration curve of our ANN does not perfectly match the line of identity (the line at a 45° angle), the deviation is pictorially minimal.

DISCUSSION

We have demonstrated that our ANN can accurately estimate the risk of breast cancer by using a dataset that contains demographic data and prospectively collected mammographic findings. To our knowledge, this study uses 1 of the largest datasets of mammography findings to develop a CADx model. Our results demonstrate that

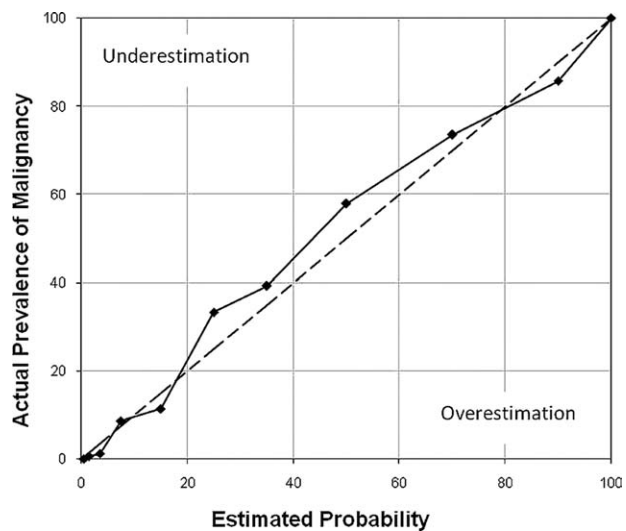


Figure 2. Calibration curve of our artificial neural network ([ANN] solid line) and the curve representing perfect calibration (dashed line) are shown. This is the actual prevalence of malignancy versus estimated risk of malignancy for each decile of the probability scale.

ANNs may have the potential to aid radiologists in discriminating between benign and malignant breast diseases. When we compare discriminative accuracy by using AUC, sensitivity, and specificity, our ANN performs significantly better than all radiologists in aggregate. Although the difference between the AUCs of the radiologists and our ANN may appear to be small (0.026), this difference is both statistically ($P < .001$) and clinically significant because our ANN identified 44 more cancers and

decreased the number of false positives by 3941 when compared with the radiologists at the specified sensitivity and specificity values. Note that these results would be similar for any other specified sensitivity and specificity values because the ROC curve of our ANN outperforms that of the radiologists at all threshold levels. On the other hand, the reason for obtaining a numerically small difference between the AUCs relates to the disproportionate number of benign findings (61,709) compared to malignant findings (510) in our dataset resulting in very high specificity at baseline and little room for improvement in this parameter.

Among statistically significant comparisons, our ANN demonstrates superior AUC, sensitivity, and specificity versus all but 1 radiologist, including the 2 highest-volume readers. Therefore, similar to other ANN models presented in the literature, our ANN has the potential to aid radiologists in classifying (discriminating) findings on mammograms by predicting the risk of malignancy. When compared with the previous CADx models developed by our research team (a logistic regression and a Bayesian network), the discrimination performance of our ANN was slightly higher (ANN AUC, 0.965; logistic regression AUC, 0.963; Bayesian network AUC, 0.960). On the other hand, no statistically significant difference was found between the ANN and the logistic regression ($P = .57$), or the ANN and the Bayesian network ($P = .13$).

However, our model is unique in several ways. In contrast to prior ANN models, which used a relatively small selected population of suspicious findings undergoing tissue sampling with biopsy as the reference standard,^{30,31,35-37,39} we use a large consecutive dataset of mammography findings with tumor registry outcomes as the reference standard to train our ANN. Furthermore, contrary to previously developed CADx models in breast cancer-risk prediction, we expand the evaluation of CADx models beyond discrimination by measuring the accuracy of the estimated probabilities themselves by using calibration metrics.

Although discrimination or accurate classification is of primary interest for CADx models,^{11,60} calibration is also crucial, especially when clinical decisions are being made for individual patients.^{11,61} Individual decisions are made under uncertainty and, therefore, aided more effectively by accurate risk estimates. Because there is a trade off between discrimination and calibration,¹⁰ the selection of the primary performance measure should be based on the intended purpose of the model.¹¹ In this study, similar to previous

CADx models, we designed our ANN primarily for optimizing the discrimination ability. However, contrary to previous CADx studies, we also measured the calibration as the secondary objective. We showed that our ANN is well calibrated, as demonstrated by the low value of the H-L statistic, the corresponding high P value, and the favorable calibration curve; and, thus, our ANN can accurately estimate the risk of malignancy for individual patients. The ability of our ANN to assign accurate numeric probabilities is an important complement to its ability to discriminate between ultimate outcomes.⁶¹

We posit that the good calibration of our ANN is attributable to both the characteristics of our training set and attributes of our model. For example, the consecutive nature of our dataset of mammography findings and the use of a tumor registry match as a reference standard, which reflects a real-world population, may lead to accurate calibration. In addition, the use of a large number of hidden nodes in concert with training with a validation set to prevent overfitting may have enhanced calibration. In future work, we plan to analyze which parameters most profoundly influence calibration.

CADx models for breast cancer risk estimation have ignored calibration and have typically been developed and evaluated on the basis of their discrimination ability.³¹⁻³⁹ Although calibration has not been formally assessed in previous CADx models, there is some evidence that these models are not well calibrated.^{31,35,42} Poor calibration may indicate that these models are not optimized for individual cases, ie, the predicted breast cancer risk for a single patient may be incorrect.

From a clinical standpoint, our ANN may be valuable because it provides an accurate post-test probability for malignancy. This post-test probability may be useful to communication among the radiologist, patient, and referring physician, which, in turn, may encourage making shared decisions.⁵⁻⁷ Each individual patient has a unique risk tolerance and comorbidities, and these factors should be considered when making decisions involving mammographic abnormalities. Risk assessments based on individual characteristics may also help promote the concept of personalized care in the diagnosis of breast cancer. Furthermore, our ANN is designed to increase the effectiveness of mammography by aiding radiologists and not by acting as a substitute. Our ANN quantifies the risk of breast cancer by using mammographic features assessed by the radiologist, so the ANN's performance depends largely on the radiologist's accurate observations and overall assessment (BI-RADS category).

Our ANN has the potential to be used as a decision-support tool, although it may face similar challenges that have, in the past, prevented the implementation of effective decision-support algorithms in clinical practice. To be used in the clinic, a decision-support tool must be seamlessly integrated into the clinical workflow, which can be challenging. We believe in the case of mammography, a decision-support tool would be most useful if directly linked to structured reporting software that radiologists use in daily practice, which would enable immediate feedback. On the other hand, the good performance of our ANN may not be preserved after the integration into clinical practice. Before clinical integration, it is important to consider the ways our ANN could fail, due to both inherent theoretical limitations and errors that may occur during the process of integration.⁶² In fact, numerous computer-aided diagnostic models that have performed well in evaluation studies have not made an impact on clinical practice.⁶³⁻⁶⁸ Furthermore, the optimal performance of our ANN would be required to gain the trust of clinicians to influence clinical practice. Unfortunately, the parameters of ANNs do not carry any real-life interpretation, and clinicians have trouble trusting decision-support algorithms that represent a “black box” without explanation capabilities. Although there is rule extraction software that converts a trained ANN to a more humanly understandable representation,⁶⁹⁻⁷¹ integration of these various software programs with the ANN requires extra effort. Therefore, we recognize that substantial challenges remain in the implementation of ANNs for decision support at the point of care, and we emphasize the importance of these issues for future research and implementation.

There are 3 important implementation considerations. First, determining the number of effective hidden nodes in an ANN is crucial and may significantly affect its output performance. Unfortunately, there is no general rule to determine the effective number of hidden nodes that maximizes the network performance when presented with an unseen dataset (generalizability).⁴⁷ Although some researchers have said that conventional wisdom suggests that when neural networks have excess hidden nodes they generalize poorly,⁴⁸ several recent studies in the machine-learning literature have shown that ANNs with excess capacity (ie, with a large number of hidden nodes) generalize better than small networks (ie, networks with a small number of hidden nodes) when trained with back-propagation and early stopping.⁴⁷⁻⁴⁹ Therefore, we built an ANN with excess capacity and did not optimize the number of hidden nodes. Also, note that if we had opti-

mized the number of hidden nodes to maximize the AUC, as other researches have, we would have achieved an even higher AUC than described here.

Second, selection of the primary performance measure is also crucial when building an ANN model. In our study, we built our ANN principally to maximize the discrimination accuracy because discrimination is of primary interest to optimize accurate diagnosis.^{11,60} On the other hand, ANNs could also be trained for maximizing the calibration when the primary purpose is to stratify individuals into higher or lower risk categories of clinical importance. However, it should be noted that for a direct maximization of calibration, the estimated probabilities by the ANN should be compared with the true underlying probabilities,⁷² which are seldom explicitly known. Alternatively, it is possible to adjust, if not maximize, the model calibration by using some advanced methods, called “recalibration”.⁷² The use of recalibration methods is beyond the scope of this article.

Third, the way we handled BI-RADS 0 findings deserves attention. We generated risk estimates for BI-RADS 0 instead of combining these results with subsequent imaging findings. In clinical practice, it is important to know the probability of cancer for BI-RADS 0 cases to make appropriate patient management decisions. Furthermore, the risk estimate for a BI-RADS 0 finding contains much more uncertainty than when the additional imaging information is available. We hope to model this uncertainty in future work to understand the value of additional information. For this reason, we consider BI-RADS 0 to be a positive mammogram^{24,45,73} and an appropriate time for risk estimation because we have the potential to 1) ameliorate anxiety that naturally arises at this juncture and 2) to improve performance in making decisions.

Our study has limitations. First, our ANN is built on a mix of screening and diagnostic examinations that cannot be reliably separated. Specifically, we do not know the proportion of diagnostic examinations in our dataset, which may influence performance.⁷⁴ However, this concern is mitigated by the finding that we are using a consecutive dataset that likely reflects an American College of Radiology-accredited mammography practice that is comparable to similar settings. Second, our dataset contains a substantial amount of unpopulated fields. Radiologists do not systematically attempt to note the absence of all findings on mammography—they simply leave them blank. Obviously, our results depend entirely on the

quality of the structured data. We labeled all missing descriptors as “not present.” Our approach is appropriate for mammography data where radiologists often leave the descriptors blank when nothing is observed on the mammogram. Third, we assume that our cancer registry is a completely accurate reference standard. Although cancer registries are generally accurate with respect to cancer diagnosis^{75,76} and represent the best method to identify cancer cases,⁷⁶ our cancer registry may not be perfect. However, the WCRS is an accredited cancer registry making it unlikely that expected errors would substantively affect our results or conclusions. Finally, our comparison of ROC curves between the radiologists and our ANN is suboptimal. Specifically, we constructed the radiologists’ ROC curve by using BI-RADS categories and our ANN’s ROC curve by using probabilities. While similar analysis has been presented in the literature,²⁴ this is not a perfectly equal comparison. These concerns are partly ameliorated in our work by the finding that the performance of our ANN is superior to that of the radiologists at all threshold levels.

In conclusion, we built our ANN by using standardized demographic risk factors and mammographic findings, which performed well in terms of both discrimination and calibration. Although future work will be required to determine whether this performance is primarily attributable to the characteristics of our training set or attributes of our model itself, we are encouraged that an ANN can achieve good performance in terms of both discrimination and calibration. Whereas ROC curve analysis is valuable to evaluate the discriminative ability of CADx models, calibration would be another important measure for evaluating models that predict risks for individual patients. These promising results may indicate that ANNs have the potential to help radiologists improve mammography interpretation.

CONFLICT OF INTEREST DISCLOSURES

This research is funded by National Institutes of Health grants R01CA127379, K07CA114181, and R21CA129393.

REFERENCES

- Giger ML. Computer-aided diagnosis in radiology. *Acad Radiol.* 2002;9:18-25.
- Kahneman D, Slovic P, Tversky A. *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge, UK: Cambridge University; 2001.
- Smith-Bindman R, Chu PW, Miglioretti DL, et al. Comparison of screening mammography in the United States and the United Kingdom. *JAMA.* 2003;290:2129-2137.
- Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists’ interpretations of mammograms. *N Engl J Med.* 1994;331:1493-1499.
- Chan EC. Promoting an ethical approach to unproven screening imaging tests. *J Am Coll Radiol.* 2005;2:311-320.
- Hillman BJ. Informed and shared decision making: an alternative to the debate over unproven screening tests. *J Am Coll Radiol.* 2005;2:297-298.
- Picano E. Informed consent and communication of risk from radiological and nuclear medicine examinations: how to escape from a communication inferno. *BMJ.* 2004;329:849-851.
- Harrell FE Jr. *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression and Survival Analysis.* 1st ed. New York, NY: Springer; 2001.
- Ikeda M, Ishigaki T, Yamauchi K. Relationship between Brier score and area under the binormal ROC curve. *Comput Methods Programs Biomed.* 2002;67:187-194.
- Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation.* 2007;115:928-935.
- Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem.* 2008;54:17-23.
- Birdwell R, Bandodkar P, Ikeda D. Computer-aided detection with screening mammography in a university hospital setting. *Radiology.* 2005;236:451-457.
- Cupples TE, Cunningham JE, Reynolds JC. Impact of computer-aided detection in a regional screening mammography program. *AJR Am J Roentgenol.* 2005;185:944-950.
- Dean JC, Ilvento CC. Improved cancer detection using computer-aided detection with diagnostic and screening mammography: prospective study of 104 cancers. *AJR Am J Roentgenol.* 2006;187:20-28.
- Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med.* 2007;356:1399-1409.
- Hadjiiski L, Sahiner B, Helvie MA, et al. Breast masses: computer-aided diagnosis with serial mammograms. *Radiology.* 2006;240:343-356.
- Chan HP, Sahiner B, Helvie MA, et al. Improvement of radiologists’ characterization of mammographic masses by using computer-aided diagnosis: an ROC study. *Radiology.* 1999;212:817-827.
- Huo Z, Giger ML, Vyborny CJ, Metz CE. Breast cancer: effectiveness of computer-aided diagnosis observer study with independent database of mammograms. *Radiology.* 2002;224:560-568.
- Kallergi M. Computer-aided diagnosis of mammographic microcalcification clusters. *Med Phys.* 2004;31:314-326.
- Jiang Y, Nishikawa RM, Schmidt RA, Metz CE. Comparison of independent double readings and computer-aided diagnosis (CAD) for the diagnosis of breast calcifications. *Acad Radiol.* 2006;13:84-94.
- Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst.* 1989;81:1879-1886.
- Claus EB, Risch N, Thompson WD. Genetic analysis of breast cancer in the cancer and steroid hormone study. *Am J Hum Genet.* 1991;48:232-242.
- Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. *Cancer.* 1994;73:643-651.

24. Barlow WE, White E, Ballard-Barbash R, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. *J Natl Cancer Inst.* 2006;98:1204-1214.
25. Freedman AN, Seminara D, Gail MH, et al. Cancer risk prediction models: a workshop on development, evaluation, and application. *J Natl Cancer Inst.* 2005;97:715-723.
26. Freer T, Ulissey M. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology.* 2001;220:781-786.
27. Morton MJ, Whaley DH, Brandt KR, Amrami KK. Screening mammograms: interpretation with computer-aided detection—prospective evaluation. *Radiology.* 2006;239:375-383.
28. Gur D, Wallace LP, Klym AH, et al. Trends in recall, biopsy, and positive biopsy rates for screening mammography in an academic practice. *Radiology.* 2005;235:396-401.
29. Vyborny CJ, Giger ML, Nishikawa RM. Computer-aided detection and diagnosis of breast cancer. *Radiol Clin North Am.* 2000;38:725-740.
30. Baker J, Kornguth P, Lo J, Floyd C Jr. Artificial neural network: improving the quality of breast biopsy recommendations. *Radiology.* 1996;198:131-135.
31. Baker J, Kornguth P, Lo J, Williford M, Floyd C Jr. Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. *Radiology.* 1995;196:817-822.
32. Burnside ES. Bayesian networks: computer-assisted diagnosis support in radiology. *Acad Radiol.* 2005;12:422-430.
33. Burnside ES, Davis J, Chhatwal J, et al. A probabilistic computer model developed from clinical data in the national mammography database format to classify mammography findings. *Radiology.* 2009;251:663-672.
34. Chhatwal J, Alagoz O, Lindstrom MJ, Kahn CE, Shaffer KA, Burnside ES. A logistic regression model based on the national mammography database format to aid breast cancer diagnosis. *AJR Am J Roentgenol.* 2009;192:1117-1127.
35. Floyd C Jr, Lo J, Yun A, Sullivan D, Kornguth P. Prediction of breast cancer malignancy using an artificial neural network. *Cancer.* 1994;74:2944-2948.
36. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. *Acad Radiol.* 1999;6:22-33.
37. Lo JY, Baker JA, Kornguth PJ, Floyd CE Jr. Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features. *Acad Radiol.* 1995;2:841-850.
38. Orr RK. Use of an artificial neural network to quantitate risk of malignancy for abnormal mammograms. *Surgery.* 2001;129:459-466.
39. Wu Y, Giger M, Doi K, Vyborny C, Schmidt R, Metz C. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. *Radiology.* 1993;187:81-87.
40. Giger ML, Huo Z, Kupinski MA, Vyborny CJ. Computer-aided diagnosis in mammography. In: Fitzpatrick JM, Milan S. *Handbook of Medical Imaging.* Vol. 2. Bellingham WA: SPIE; 2000:917-986.
41. Gurney JW. Neural networks at the crossroads: caution ahead. *Radiology.* 1994;193:28-30.
42. Burnside ES, Rubin DL, Fine JP, Shachter RD, Sisney GA, Leung WK. Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience. *Radiology.* 2006;240:666-673.
43. Sickles EA. Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases. *Radiology.* 1991;179:463-468.
44. Osuch J, Anthony M, Bassett L, et al. A proposal for a national mammography database: content, purpose, and value. *AJR Am J Roentgenol.* 1995;164:1329-1334.
45. Breast Imaging Reporting And Data System (BI-RADS), 4th ed. Reston, VA: American College of Radiology; 2004.
46. Rumelhart DE, Hinton EE, Williams RJ. Learning representations by back-propagating errors. *Nature.* 1986;323:533.
47. Weigend A. On overfitting and the effective number of hidden units. In: Mozer A, Smolensky P, Touretzky DS, Elman JL, Weigend AS, eds. *Proceedings of the 1993 Connectionist Models Summer School.* Hillsdale, NJ: Lawrence Erlbaum; 1994:335-342.
48. Caruana R, Lawrence S, Giles CL. Overfitting in neural networks: backpropagation, conjugate gradient, and early stopping. In: Leen TK, Dietterich TG, Tresp V, eds. *Advances in Neural Information Processing Systems 13.* Cambridge: Mass: MIT Press; 2001.
49. Lawrence S, Giles CL, Tsoi AC. Lessons in neural network training: overfitting may be harder than expected. In: American Association for Artificial Intelligence. *Proceedings of the Fourteenth National Conference on Artificial Intelligence.* Menlo Park, Calif: AAAI Press; 1997:540-545.
50. Bishop CM. *Neural Networks For Pattern Recognition.* New York, NY: Oxford; 1995.
51. Mitchell TM. *Machine Learning.* Burr Ridge, Ill: McGraw Hill; 1997.
52. Haykin S. *Neural Networks: A Comprehensive Foundation.* Upper Saddle River, NJ: Prentice Hall; 1998.
53. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44:837-845.
54. Bassett LW, Hendrick RE, Bassford TL, Butler PF, Carter D, DeBor M; Agency for Health Care Policy and Research. *Quality Determinants of Mammography. Clinical Practice Guideline No. 13.* Washington, DC: Public Health Service, US Department of Health and Human Services; 1994.
55. Rosner B. Two-sample test for binomial proportions for matched-pair data (McNemar's Test). In: Rosner B, ed. *Fundamentals of Biostatistics.* Pacific Grove, Calif: Duxbury; 2000:376-384.
56. Agresti A, Coull BA. Approximate is better than exact for interval estimation of binomial proportions. *Am Stat.* 1998;52:119-126.
57. Hosmer DW, Lemeshow S. *Applied Logistic Regression.* New York, NY: John Wiley; 2000.
58. Poses RM, Cebul RD, Centor RM. Evaluating physicians' probabilistic judgments. *Med Decis Making.* 1988;8:233-240.
59. Diamond G. What price perfection? Calibration and discrimination of clinical prediction models. *J Clin Epidemiol.* 1992;45:85-89.
60. Ohno-Machado L. *Medical Applications of Artificial Neural Networks: Connectionist Models of Survival [dissertation].* Stanford, Calif: Stanford University; 1996.
61. Harrell FE Jr, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems,

- and suggested solutions. *Cancer Treat Rep.* 1985;69:1071-1077.
62. Miller RA. Why the standard view is standard: people, not machines, understand patients' problems. *J Med Philos.* 1990;15:581-591.
 63. Hickam DH, Shortliffe EH, Bischoff MB, Scott AC, Jacobs CD. The treatment advice of a computer-based cancer chemotherapy protocol advisor. *Ann Intern Med.* 1985; 103(6 pt 1):928-936.
 64. Masarie FE Jr, Miller RA, Myers JD. INTERNIST-I properties: representing common sense and good medical practice in a computerized medical knowledge base. *Comput Biomed Res.* 1985;18:458-479.
 65. Miller RA, Pople HE Jr, Myers JD. Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. *N Engl J Med.* 1982;307:468-476.
 66. Shortliffe EH. Medical expert systems—knowledge tools for physicians. *West J Med.* 1986;145:830-839.
 67. Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Comput Biomed Res.* 1975;8:303-320.
 68. Yu VL, Buchanan BG, Shortliffe EH, et al. Evaluating the performance of a computer-based consultant. *Comput Programs Biomed.* 1979;9:95-102.
 69. Craven MW, Shavlik JW. Extracting tree-structured representations of trained networks. In: Touretzky DS, Mozer MC, Hasselmo ME, eds. *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference.* Cambridge, Mass: MIT; 1996:24-30.
 70. Nauck D, Klawonn F, Kruse R. *Foundations of Neuro-Fuzzy Systems.* New York, NY: John Wiley; 1997.
 71. Setiono R, Liu H. Symbolic representation of neural networks. *Computer.* 1996;29:71-77.
 72. Vinterbo S. *Predictive models in medicine: some methods for construction and adaptation.* Trondheim, Norway: Norwegian University of Science and Technology, 1999.
 73. Taplin S, Abraham L, Barlow WE, et al. Mammography facility characteristics associated with interpretive accuracy of screening mammography. *J Natl Cancer Inst.* 2008;100:876-887.
 74. Sohlich RE, Sickles EA, Burnside ES, Dee KE. Interpreting data from audits when screening and diagnostic mammography outcomes are combined. *AJR Am J Roentgenol.* 2002; 178:681-686.
 75. Bickell NA, Chassin MR. Determining the quality of breast cancer care: do tumor registries measure up? *Ann Intern Med.* 2000;132:705-710.
 76. Malin JL, Kahn KL, Adams J, Kwan L, Laouri M, Ganz PA. Validity of cancer registry data for measuring the quality of breast cancer care. *J Natl Cancer Inst.* 2002;94:835-844.