Title: A universal translator for AI scores: Providing context using error

Authors & Affiliations: Maggie Chung, MD[1], Micheal H. Bernstein, PhD[2], Adam Yala, PhD[3], and Grayson L. Baird[2], PhD

1. Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA
2. Brown Radiology Human Factors Lab, Department of Diagnostic Imaging, The Warren Alpert Medical School, Brown University, and Brown University Health, Providence, RI
3. Computational Precision Health, University of California, Berkeley and University of California, San Francisco, CA

Abstract

Artificial intelligence (AI) programs in radiology typically provide a numeric score for each case that correlates with the underlying pathology. However, these scores are not readily interpretable by themselves. To address this, we propose improving score interpretability by providing the False Discovery Rate (FDR) and False Omission Rate (FOR) corresponding with each score threshold. Using an open-source AI program for breast cancer, we estimated FDR and FOR across a range of AI scores using data from 130,712 digital screening mammograms, of which 907 were positive and 129,805 were negative. FDR and FOR ranged from 99.27% and 0.03%, respectively, at the low end of the score distribution to 60.98% and 0.65%, respectively, at the high end of the distribution. Providing these error rates alongside AI scores allows clinicians to consider the balance of trade-offs between false positive and false negative interpretations.

Introduction.

Artificial intelligence (AI) in medicine has expanded rapidly in recent years [1,2], particularly in radiology. AI outputs almost always include a numeric score for the radiologist interpreting images with AI's aid [3-7]. These scores come in various forms, including risk scores, severity scores, etc. All are designed to provide a numeric value correlating with possible pathology for a particular image. However, these scores have several limitations, which we have discussed at length [8]. Below, we briefly review a few of these key limitations. We then discuss a solution for how scores can be presented and provide an empirical example.

Limitations of AI Scores

First, AI scores are not readily interpretable, especially when they fall within the mid-range rather than at the extreme ends of the scale. For instance, if scores can range from 0.0 to 1.0, it is

unclear what a score of 0.50 or 0.70 represents in a clinical context. In other words, AI-generated scores lack intrinsic interpretability— that is, AI scores are not directly interpretable by themselves. This ambiguity may paradoxically increase a clinician's uncertainty when evaluating a given case rather than providing useful guidance to inform a radiologist's interpretation.

Second, each AI algorithm has its own, often proprietary, scoring system [8]. As a result, a score of 0.90 from one AI model may not correspond to the same level of risk as a score of 0.90 in another model, even when assessing the same pathology. Moreover, different algorithms may use entirely different scoring scales, further complicating direct comparisons between models. Finally, the relationship between scores with and without pathology is unknown. For example, we cannot assume that the relative increase in risk between a score of 0.20 and 0.30 is the same as the relative increase between a score of 0.60 and 0.70. Without a clear understanding of how scores relate to pathology, clinicians may struggle to determine what constitutes a meaningful change in risk. Moreover, because a score is not inherently interpretable, clinicians will likely not agree with each other about the meaning of a score, placing the inter-rater reliability of the interpretation of scores into question. These factors markedly limit the utility of these scores in clinical decision-making.

How to Make AI Scores Interpretable

Since AI scores alone cannot resolve these ambiguities, information needs to accompany these scores to provide context. Traditional accuracy metrics, like sensitivity and specificity, are commonly used when validating AI systems, but they are not good options for aiding score interpretation. This is because radiologists only know the AI score, not if the case has pathology. Specifically, the question relevant to a radiologist is what is the proportion of cases for a given AI score or higher that have pathology (i.e., the denominator being all cases with a given AI

score or higher) instead of sensitivity, which is the proportion of cases with pathology that have a given AI score or higher (i.e., the denominator being all cases with pathology).

Therefore, we propose providing radiologists with the probability of pathology conditioned on the AI score, which is the only piece of information known to the radiologist—that is, the rates of error corresponding with a score as a threshold. Presenting the predicted probabilities (PP) and corresponding false discovery rate (FDR) and false omission rate (FOR) with AI scores provides radiologists with clinically relevant information. For example, the PP allows a radiologist to compare a given AI score's likelihood of pathology relative to the base rate of pathology. Likewise, the FDR represents the probability that the AI score or higher is actually negative for pathology (1-positive predictive value, PPV), while the FOR represents the probability of the values under the AI score being actually positive for pathology (1-negative predictive value, NPV).

Calculating the PP, FDR, and FOR is straightforward. Practices using a given AI system must first run that algorithm on their local historical data—a local validation. From this validation, they need only to regress outcomes with the AI scores using the generalized linear model (mixed, if applicable). If that is not possible, practices can also use their local prevalence rate along with published sensitivities and specificities of the AI algorithm to calculate the FDR and FOR for their local practice. We will demonstrate both applications using screening mammography as a case study. By pairing AI scores with PPs, FDRs, and FORs, we provide a framework for clinicians to make more informed, probability-based decisions using local prevalence.

Methods

University of California, San Francisco (UCSF) Institutional Review Board gave ethical approval for this Health Insurance Portability and Accountability Act–compliant study and waived the requirement for written informed consent.

Study sample. We conducted a retrospective review using a single-institution radiology database to identify 130,712 digital screening mammograms acquired between January 2006 and January 2023. All cases included at least one year of mammographic and/or clinical follow-up. Positive exams were defined as those with a histopathologic diagnosis of invasive breast carcinoma or ductal carcinoma in situ (DCIS) within 12 months of imaging. Negative exams were defined as those with at least 12 months of follow-up without a breast cancer diagnosis. Among the 129,805 cases, there were 907 positive exams.

AI Model. To promote reproducibility and open science, we applied Mirai, an open-source AI model for mammogram-based risk prediction, to estimate 1-year breast cancer risk using 2D digital mammograms [9]. No additional image processing was performed before the model application.

Statistics. The logistic function was fit using the LOGISTIC and GLIMMIX procedures in SAS 9.4 (SAS Cary, NC). Sensitivities and specificities were estimated using the %ROCPLOT macro and PPV and NPV (and their complements) were calculated using Bayes' Theorem. Presence of cancer was regressed on AI scores from Mirai. The base rate of cancer was 0.69%. We also use a base of 0.57% as an example when local historical data are unavailable (Table 1).

Results

Providing Context with Scores. As demonstrated in Table 1 (and partially illustrated in Supplemental Video 1), each selected AI score is presented with its corresponding PP, FDR, and FOR values as a reference. For brevity, these values are provided at 0.01-unit increments from 0.01 to 0.85 (inclusive). Also included in Table 1, for reference, are sensitivity, specificity, NPV, PPV, and prevalence.

For example, for a score of 0.50, the PP of cancer is 2.62% (corresponding PP in Table 1), representing a 3.7-fold increase in the risk of cancer compared to the overall prevalence of 0.69%. The FDR at this threshold is 95%, meaning that out of all cases being 0.50 or higher, 95% will be negative, and 5% will be positive for cancer. Thus, among 100 cases with scores ≥0.50, 95 would be negative and 5 positive for cancer. The FOR at this threshold is 0.2%, meaning that out of all cases under 0.50, 99.8% will be negative, and 0.2% will be positive or cancer. Thus, among 1,000 cases with scores <0.50, 998 will be negative and 2 will be positive for cancer. Now, consider a higher score threshold of 0.70. The PP of cancer is 7.9%, corresponding to an 11.3-fold increase in the risk of cancer relative to the prevalence rate. The FDR is 84%, and the FOR is 0.5%.

And finally, let us consider the extremes. At a score of 0.01, the PP of cancer is 0.15%, representing a 0.22-fold decrease in risk relative to the prevalence rate of 0.69%. At the 0.01 threshold, FDR is 99.27% and the FOR is 0.03%. A score of 0.90 corresponds to a PP of cancer of 21.5%, representing a relative increase of 31-fold compared to a prevalence rate of 0.69%. At that threshold, the FDR is 56%, and the FOR is 0.7%. Note, at the highest levels of the score (0.90), the PP of cancer is roughly 20%, and about half of the cases at or above 0.90 are actually negative.

Context with change in scores. This demonstration allows us to examine how a particular change in scores should be interpreted across the range of all possible scores. For example, a 0.10 increase from an AI score of 0.20 to 0.30 means an increase in PP from 0.46% to 0.82%, a small decrease in FDR from 97.9% to 97.2%, and an increase in FOR from 0.2% to 0.3%. In contrast, an identical 0.10 increase in the score from 0.60 to 0.70 results in an increase in PP from 4.5% to 7.9%, a larger decrease in FDR from 91.6% to 84.4%, and no change in FOR (remaining at 0.5%). These examples highlight that changes in error rates are not uniform across the score scale.

Comparison across AI systems. Providing scores with FDR and FOR allows for a more direct comparison across AI systems in clinical settings. Different AI algorithms or systems may produce similar raw scores, but the trade-offs between false positives and false negatives may vary significantly. For example (hypothetically), imagine comparing two AI algorithms for the same pathology: a score of “5” using one algorithm may translate into an FDR of 95% and FOR of 0.1%, while a score of “5” using another algorithm may translate into an FDR of 67% and FOR of 0.3%. By reporting both FDR and FOR for each score across AI systems, clinicians are informed of how each algorithm behaves across various thresholds. This allows them to make more informed clinical decisions when selecting an AI model by allowing for cross-comparison of AI models and assessing the real-world implications of each model in their specific patient population.

Bridging scores and clinical practice. FDR and FOR can easily be converted to number of patients, making the clinical interpretability of scores straightforward for clinicians and patients. As seen in Table 1, out of 1,000 mammograms and a 0.69% prevalence, a score of 0.50 or higher corresponds with 54 false positives and 4 false negatives.

When historical data are not available. When it is not possible to run an AI system on historical data, published sensitivities and specificities can be used to calculate the local FDR and FOR if the local prevalence of pathology is known, assuming these sensitivities and specificity estimates are for the same population. This is demonstrated in Table 1 using a common prevalence of 0.57% for breast cancer, though any prevalence could be used.

Discussion

By reporting the error rates corresponding to each score, scores can now be contextualized using a common language—the language of probability and error. This is achieved because AI scores can be regressed with outcomes. When done with local historical data, this provides the estimates of PP, FDR, and FOR (and sensitivity, specificity, PPV, NPV) for each score for future reference. As mentioned, if a practice cannot derive these estimates with their own historical data, they can use published sensitivity and specificity estimates, along with their local prevalence, to estimate FDR and FOR for each score. Both approaches can be done with every new AI algorithm or AI algorithm update.

Providing PP, FDR, and FOR in conjunction with scores enables clinicians not just to interpret scores in a general sense but also to interpret these scores within the specific context of their patient population. Disease prevalence serves a critical role in how scores should be interpreted [10]. Commercial AI models often come with thresholds pre-set by the vendor based on their own training data and performance metrics. These thresholds determine the cut-off at which the AI classifies cases as positive or negative and are usually based on the model of the development datasets. However, it may not necessarily reflect the local population where the model will be

deployed. Reporting FDR and FOR helps clinicians apply generalized AI models, consider their local population and prevalence, and assess if the trade-off between false positives and false negatives is aligned with their practice patterns.

As AI tools become integrated into clinical practice, providing error rates can also assist with patient education and the shared decision-making process between clinicians and patients. For instance, if an AI system predicts a high risk of cancer but the FDR at that threshold is 95%, this means that while the AI identifies cases as positive, the vast majority (95%) of those cases are actually negative. Educating patients about this error rate can help manage anxiety and set appropriate expectations [11].

References

1. Haug, C.J. and Drazen, J.M., 2023. Artificial intelligence and machine learning in clinical medicine, 2023. New England Journal of Medicine, 388(13), pp.1201-1208.
2. Reddy, S., 2022. Explainability and artificial intelligence in medicine. The Lancet Digital Health, 4(4), pp.e214-e215.
3. Li, M.D., Little, B.P., Alkasab, T.K., Mendoza, D.P., Succi, M.D., Shepard, J.A.O., Lev, M.H. and Kalpathy-Cramer, J., 2021. Multi-radiologist user study for artificial intelligence-guided grading of COVID-19 lung disease severity on chest radiographs. Academic radiology, 28(4), pp.572-576.
4. Lessmann, N., Sánchez, C.I., Beenen, L., Boulogne, L.H., Brink, M., Calli, E., Charbonnier, J.P., Dofferhoff, T., van Everdingen, W.M., Gerke, P.K. and Geurts, B., 2021. Automated assessment of COVID-19 reporting and data system and chest CT severity scores in patients suspected of having COVID-19 using artificial intelligence. Radiology, 298(1), pp.E18-E28.
5. Van Assen, M., Zandehshahvar, M., Maleki, H., Kiarashi, Y., Arleo, T., Stillman, A.E., Filev, P., Davarpanah, A.H., Berkowitz, E.A., Tigges, S. and Lee, S.J., 2022. COVID-19 pneumonia chest radiographic severity score: variability assessment among experienced and in-training radiologists and creation of a multireader composite score database for artificial intelligence algorithm development. The British Journal of Radiology, 95(1134), p.20211028.
6. Pacilè, S., Lopez, J., Chone, P., Bertinotti, T., Grouin, J.M. and Fillard, P., 2020. Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool. Radiology: Artificial Intelligence, 2(6), p.e190208.

7. Ahn, J.S., Ebrahimian, S., McDermott, S., Lee, S., Naccarato, L., Di Capua, J.F., Wu, M.Y., Zhang, E.W., Muse, V., Miller, B. and Sabzalipour, F., 2022. Association of artificial intelligence–aided chest radiograph interpretation with reader performance and efficiency. JAMA Network Open, 5(8), pp.e2229289-e2229289.
8. Is a score enough? Pitfalls and Solutions for AI Severity Scores (under review).
9. Yala, A., Mikhael, P.G., Strand, F., Lin, G., Smith, K., Wan, Y.L., Lamb, L., Hughes, K., Lehman, C. and Barzilay, R., 2021. Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine*, *13*(578).
10. Scaringi, J.A., McTaggart, R.A., Alvin, M.D., Atalay, M., Bernstein, M.H., Jayaraman, M.V., Jindal, G., Movson, J.S., Swenson, D.W. and Baird, G.L., 2024. Implementing an AI algorithm in the clinical setting: a case study for the accuracy paradox. European Radiology, pp.1-7.
11. Song, E.C., Bernstein, M.H., Lay, P.S., Druart, L., Dibble, E.H., Lourenco, A.P. and Baird, G.L., 2024. Accessing AI mammography reports impacts patient interest in pursuing a medical malpractice claim: The unintended consequences of including AI in patient portals. medRxiv, pp.2024-12.

Table 1. Mirai Scores corresponding with Diagnostic Performance Metrics and outcomes (false positive and negative counts)

	Values After running on Local Historical Data											Values After using Local Prevalence							
Mirai	PP	Sen.	Spec.	Prev.	PPV	FDR	NPV	FOR	Rate #	#FN	#FP	Local Prev.	Local PPV	Local FDR	Local NPV	Local FOR	Rate	Local #FN	Local #FP
	%	%	%	%	%	%	%	%	Count	Count	Count	%	%	%	%	%	Count	Count	Count
0.01	0.152	99.78	4.70	0.69	0.73	99.27	99.97	0.03	1000	0	946	0.57	0.60	99.40	99.97	0.03	1000	0	948
0.02	0.162	98.46	17.09	0.69	0.82	99.18	99.94	0.06	1000	0	823	0.57	0.68	99.32	99.95	0.05	1000	0	824
0.03	0.172	96.80	26.88	0.69	0.92	99.08	99.92	0.08	1000	0	726	0.57	0.75	99.25	99.93	0.07	1000	0	727
0.04	0.177	96.14	30.39	0.69	0.96	99.04	99.91	0.09	1000	0	691	0.57	0.79	99.21	99.93	0.07	1000	0	692
0.05	0.194	92.94	39.98	0.69	1.07	98.93	99.88	0.12	1000	0	596	0.57	0.88	99.12	99.90	0.10	1000	0	597
0.06	0.207	90.74	44.98	0.69	1.14	98.86	99.86	0.14	1000	1	546	0.57	0.94	99.06	99.88	0.12	1000	1	547
0.07	0.211	90.41	46.43	0.69	1.17	98.83	99.86	0.14	1000	1	532	0.57	0.96	99.04	99.88	0.12	1000	1	533
0.08	0.224	88.42	50.20	0.69	1.23	98.77	99.84	0.16	1000	1	495	0.57	1.01	98.99	99.87	0.13	1000	1	495
0.09	0.238	86.88	53.89	0.69	1.30	98.70	99.83	0.17	1000	1	458	0.57	1.07	98.93	99.86	0.14	1000	1	459
0.10	0.261	84.67	58.16	0.69	1.39	98.61	99.82	0.18	1000	1	416	0.57	1.15	98.85	99.85	0.15	1000	1	416
0.11	0.275	83.02	60.48	0.69	1.45	98.55	99.80	0.20	1000	1	392	0.57	1.19	98.81	99.84	0.16	1000	1	393
0.12	0.294	81.59	63.07	0.69	1.52	98.48	99.80	0.20	1000	1	367	0.57	1.25	98.75	99.83	0.17	1000	1	367
0.13	0.301	81.48	63.89	0.69	1.55	98.45	99.80	0.20	1000	1	359	0.57	1.28	98.72	99.83	0.17	1000	1	359
0.14	0.327	79.60	66.75	0.69	1.65	98.35	99.79	0.21	1000	1	330	0.57	1.35	98.65	99.83	0.17	1000	1	331
0.15	0.345	78.28	68.40	0.69	1.70	98.30	99.78	0.22	1000	2	314	0.57	1.40	98.60	99.82	0.18	1000	1	314
0.16	0.354	78.06	69.17	0.69	1.74	98.26	99.78	0.22	1000	2	306	0.57	1.43	98.57	99.82	0.18	1000	1	307
0.17	0.387	76.30	71.54	0.69	1.84	98.16	99.77	0.23	1000	2	283	0.57	1.51	98.49	99.81	0.19	1000	1	283
0.18	0.398	76.07	72.20	0.69	1.88	98.12	99.77	0.23	1000	2	276	0.57	1.54	98.46	99.81	0.19	1000	1	276
0.19	0.431	75.19	74.01	0.69	1.98	98.02	99.77	0.23	1000	2	258	0.57	1.63	98.37	99.81	0.19	1000	1	258
0.20	0.449	74.64	74.90	0.69	2.04	97.96	99.76	0.24	1000	2	249	0.57	1.68	98.32	99.81	0.19	1000	1	250
0.21	0.495	72.22	76.87	0.69	2.14	97.86	99.75	0.25	1000	2	230	0.57	1.76	98.24	99.79	0.21	1000	2	230
0.22	0.503	72.00	77.18	0.69	2.16	97.84	99.75	0.25	1000	2	227	0.57	1.78	98.22	99.79	0.21	1000	2	227
0.23	0.549	70.45	78.69	0.69	2.26	97.74	99.74	0.26	1000	2	212	0.57	1.86	98.14	99.79	0.21	1000	2	212
0.24	0.590	69.02	79.96	0.69	2.35	97.65	99.73	0.27	1000	2	199	0.57	1.94	98.06	99.78	0.22	1000	2	199
0.25	0.628	68.14	80.98	0.69	2.44	97.56	99.73	0.27	1000	2	189	0.57	2.01	97.99	99.77	0.23	1000	2	189
0.26	0.637	67.92	81.18	0.69	2.46	97.54	99.72	0.28	1000	2	187	0.57	2.03	97.97	99.77	0.23	1000	2	187
0.27	0.692	66.59	82.36	0.69	2.57	97.43	99.72	0.28	1000	2	175	0.57	2.12	97.88	99.77	0.23	1000	2	175

0.28	0.726	65.60	83.07	0.69	2.64	97.36	99.71	0.29	1000	2	168
0.29	0.753	65.05	83.55	0.69	2.69	97.31	99.71	0.29	1000	2	163
0.30	0.804	64.06	84.39	0.69	2.79	97.21	99.70	0.30	1000	2	155
0.31	0.868	63.07	85.26	0.69	2.90	97.10	99.70	0.30	1000	3	146
0.32	0.909	62.18	85.76	0.69	2.96	97.04	99.69	0.31	1000	3	141
0.33	0.983	60.75	86.60	0.69	3.07	96.93	99.68	0.32	1000	3	133
0.34	1.021	60.42	86.96	0.69	3.13	96.87	99.68	0.32	1000	3	130
0.35	1.093	59.65	87.64	0.69	3.26	96.74	99.68	0.32	1000	3	123
0.36	1.157	58.32	88.21	0.69	3.34	96.66	99.67	0.33	1000	3	117
0.37	1.205	57.99	88.56	0.69	3.42	96.58	99.67	0.33	1000	3	114
0.38	1.274	57.22	89.07	0.69	3.53	96.47	99.67	0.33	1000	3	109
0.39	1.416	54.91	89.95	0.69	3.68	96.32	99.65	0.35	1000	3	100
0.40	1.499	54.24	90.48	0.69	3.83	96.17	99.65	0.35	1000	3	95
0.41	1.520	54.13	90.60	0.69	3.87	96.13	99.65	0.35	1000	3	93
0.42	1.619	53.36	91.13	0.69	4.04	95.96	99.64	0.36	1000	3	88
0.43	1.699	52.59	91.53	0.69	4.16	95.84	99.64	0.36	1000	3	84
0.44	1.849	51.38	92.19	0.69	4.40	95.60	99.63	0.37	1000	3	78
0.45	1.944	49.94	92.57	0.69	4.48	95.52	99.62	0.38	1000	3	74
0.46	2.040	49.39	92.91	0.69	4.64	95.36	99.62	0.38	1000	4	70
0.47	2.135	48.73	93.23	0.69	4.79	95.21	99.62	0.38	1000	4	67
0.48	2.248	47.74	93.57	0.69	4.94	95.06	99.61	0.39	1000	4	64
0.49	2.406	45.87	94.03	0.69	5.09	94.91	99.60	0.40	1000	4	59
0.50	2.616	44.21	94.55	0.69	5.36	94.64	99.59	0.41	1000	4	54
0.51	2.682	43.77	94.69	0.69	5.44	94.56	99.59	0.41	1000	4	53
0.52	2.893	42.34	95.12	0.69	5.72	94.28	99.58	0.42	1000	4	48
0.53	3.135	41.01	95.60	0.69	6.11	93.89	99.57	0.43	1000	4	44
0.54	3.217	40.90	95.72	0.69	6.27	93.73	99.57	0.43	1000	4	42
0.55	3.497	39.25	96.14	0.69	6.63	93.37	99.56	0.44	1000	4	38
0.56	3.688	38.26	96.40	0.69	6.91	93.09	99.55	0.45	1000	4	36
0.57	3.891	37.16	96.68	0.69	7.25	92.75	99.55	0.45	1000	4	33
0.58	3.963	37.05	96.78	0.69	7.44	92.56	99.55	0.45	1000	4	32

0.57	2.17	97.83	99.76	0.24	1000	2	168
0.57	2.22	97.78	99.76	0.24	1000	2	164
0.57	2.30	97.70	99.76	0.24	1000	2	155
0.57	2.39	97.61	99.75	0.25	1000	2	147
0.57	2.44	97.56	99.75	0.25	1000	2	142
0.57	2.53	97.47	99.74	0.26	1000	2	133
0.57	2.59	97.41	99.74	0.26	1000	2	130
0.57	2.69	97.31	99.74	0.26	1000	2	123
0.57	2.76	97.24	99.73	0.27	1000	2	117
0.57	2.82	97.18	99.73	0.27	1000	2	114
0.57	2.91	97.09	99.73	0.27	1000	2	109
0.57	3.04	96.96	99.71	0.29	1000	3	100
0.57	3.16	96.84	99.71	0.29	1000	3	95
0.57	3.20	96.80	99.71	0.29	1000	3	93
0.57	3.34	96.66	99.71	0.29	1000	3	88
0.57	3.44	96.56	99.70	0.30	1000	3	84
0.57	3.64	96.36	99.70	0.30	1000	3	78
0.57	3.71	96.29	99.69	0.31	1000	3	74
0.57	3.84	96.16	99.69	0.31	1000	3	70
0.57	3.96	96.04	99.69	0.31	1000	3	67
0.57	4.09	95.91	99.68	0.32	1000	3	64
0.57	4.22	95.78	99.67	0.33	1000	3	59
0.57	4.44	95.56	99.66	0.34	1000	3	54
0.57	4.51	95.49	99.66	0.34	1000	3	53
0.57	4.74	95.26	99.65	0.35	1000	3	48
0.57	5.07	94.93	99.65	0.35	1000	3	44
0.57	5.20	94.80	99.65	0.35	1000	3	43
0.57	5.51	94.49	99.64	0.36	1000	3	38
0.57	5.74	94.26	99.63	0.37	1000	4	36
0.57	6.03	93.97	99.63	0.37	1000	4	33
0.57	6.19	93.81	99.63	0.37	1000	4	32

0.59	4.192	36.05	97.02	0.69	7.79	92.21	99.54	0.46	1000	4	30		0.57	6.48	93.52	99.62	0.38	1000	4	30
0.6	4.506	34.73	97.33	0.69	8.32	91.68	99.53	0.47	1000	5	27		0.57	6.93	93.07	99.62	0.38	1000	4	27
0.61	4.857	33.52	97.62	0.69	8.97	91.03	99.53	0.47	1000	5	24		0.57	7.48	92.52	99.61	0.39	1000	4	24
0.62	4.985	32.86	97.72	0.69	9.13	90.87	99.52	0.48	1000	5	23		0.57	7.62	92.38	99.61	0.39	1000	4	23
0.63	5.240	31.75	97.90	0.69	9.57	90.43	99.52	0.48	1000	5	21		0.57	7.99	92.01	99.60	0.40	1000	4	21
0.64	5.551	30.32	98.11	0.69	10.09	89.91	99.51	0.49	1000	5	19		0.57	8.43	91.57	99.59	0.41	1000	4	19
0.65	5.883	29.33	98.29	0.69	10.71	89.29	99.50	0.50	1000	5	17		0.57	8.96	91.04	99.59	0.41	1000	4	17
0.66	6.231	27.89	98.47	0.69	11.32	88.68	99.49	0.51	1000	5	15		0.57	9.48	90.52	99.58	0.42	1000	4	15
0.67	6.812	26.13	98.74	0.69	12.65	87.35	99.48	0.52	1000	5	13		0.57	10.62	89.38	99.57	0.43	1000	4	13
0.68	7.176	25.25	98.89	0.69	13.70	86.30	99.47	0.53	1000	5	11		0.57	11.53	88.47	99.57	0.43	1000	4	11
0.69	7.539	24.48	99.02	0.69	14.80	85.20	99.47	0.53	1000	5	10		0.57	12.47	87.53	99.56	0.44	1000	4	10
0.70	7.876	23.48	99.12	0.69	15.70	84.30	99.46	0.54	1000	5	9		0.57	13.25	86.75	99.56	0.44	1000	4	9
0.71	8.226	21.83	99.21	0.69	16.26	83.74	99.45	0.55	1000	5	8		0.57	13.74	86.26	99.55	0.45	1000	4	8
0.72	8.862	20.51	99.37	0.69	18.60	81.40	99.44	0.56	1000	6	6		0.57	15.79	84.21	99.54	0.46	1000	5	6
0.73	9.200	20.29	99.43	0.69	19.91	80.09	99.44	0.56	1000	6	6		0.57	16.94	83.06	99.54	0.46	1000	5	6
0.74	9.714	18.74	99.51	0.69	20.96	79.04	99.43	0.57	1000	6	5		0.57	17.87	82.13	99.53	0.47	1000	5	5
0.75	10.208	16.87	99.56	0.69	21.31	78.69	99.42	0.58	1000	6	4		0.57	18.18	81.82	99.52	0.48	1000	5	4
0.76	10.575	15.88	99.60	0.69	21.92	78.08	99.41	0.59	1000	6	4		0.57	18.72	81.28	99.52	0.48	1000	5	4
0.77	11.388	14.66	99.68	0.69	24.18	75.82	99.41	0.59	1000	6	3		0.57	20.74	79.26	99.51	0.49	1000	5	3
0.78	11.971	12.90	99.72	0.69	24.53	75.47	99.39	0.61	1000	6	3		0.57	21.05	78.95	99.50	0.50	1000	5	3
0.79	12.311	12.57	99.76	0.69	26.51	73.49	99.39	0.61	1000	6	2		0.57	22.84	77.16	99.50	0.50	1000	5	2
0.80	13.404	10.92	99.82	0.69	29.64	70.36	99.38	0.62	1000	6	2		0.57	25.69	74.31	99.49	0.51	1000	5	2
0.81	14.239	10.03	99.85	0.69	32.38	67.62	99.37	0.63	1000	6	1		0.57	28.21	71.79	99.49	0.51	1000	5	1
0.82	14.978	9.37	99.88	0.69	34.41	65.59	99.37	0.63	1000	6	1		0.57	30.09	69.91	99.48	0.52	1000	5	1
0.83	15.141	8.82	99.88	0.69	34.04	65.96	99.37	0.63	1000	6	1		0.57	29.75	70.25	99.48	0.52	1000	5	1
0.84	16.505	7.06	99.92	0.69	37.65	62.35	99.35	0.65	1000	6	1		0.57	33.13	66.87	99.47	0.53	1000	5	1
0.85	16.636	7.06	99.92	0.69	39.02	60.98	99.35	0.65	1000	6	1		0.57	34.43	65.57	99.47	0.53	1000	5	1

Table Notes. AI scores ranging from 0.01 to 0.85 are provided along with the corresponding diagnostic performance metrics and outcomes (false positive and negative counts). PP= Predicted Probability (likelihood), Sen=Sensitivity, Spec=Specificity, PPV=Positive Predictive Value, FDR=False Discovery Rate (1-PPV), NPV=Negative Predictive Value, FOR=False Omission Rate (1-NPV), Prev.=prevalence, Rate (hypothetical number of mammograms read), #FN (of these 1,000, number of misses (false negatives)), #FP (of these 1,000, number of false alarms (false positives)).