Journal of
Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Developing a utility decision framework to evaluate predictive models in breast cancer risk estimation

Yirong Wu
Craig K. Abbey
Xianqiao Chen
Jie Liu
David C. Page
Oguzhan Alagoz
Peggy Peissig
Adedayo A. Onitilo
Elizabeth S. Burnside

SPIE.

Developing a utility decision framework to evaluate predictive models in breast cancer risk estimation

Yirong Wu,[a] Craig K. Abbey,[b] Xianqiao Chen,[c] Jie Liu,[d] David C. Page,[e] Oguzhan Alagoz,[f] Peggy Peissig,[g] Adedayo A. Onitilo,[g,h] and Elizabeth S. Burnside[a,*]
[a]University of Wisconsin-Madison, Department of Radiology, 600 Highland Avenue, Madison, Wisconsin 53792, United States
[b]University of California-Santa Barbara, Department of Psychological and Brain Sciences, 251 UCEN Road, Santa Barbara, California 93106, United States
[c]Wuhan University of Technology, School of Computer Science and Technology, 1178 Heping Avenue, Wuhan, Hubei 430070, China
[d]University of Washington-Seattle, Department of Genome Sciences, 3720 15th Avenue, Seattle, Washington 98105, United States
[e]University of Wisconsin-Madison, Department of Biostatistics and Medical Informatics, 600 Highland Avenue, Madison, Wisconsin 53706, United States
[f]University of Wisconsin-Madison, Department of Industrial and Systems Engineering, 1513 University Avenue, Madison, Wisconsin 53706, United States
[g]Marshfield Clinic Research Foundation, 1000 North Oak Avenue, Marshfield, Wisconsin 54449, United States
[h]Marshfield Clinic Weston Center, Department of Hematology/Oncology, 3501 Cranberry Boulevard, Weston, Wisconsin 54476, United States

Abstract. Combining imaging and genetic information to predict disease presence and progression is being codified into an emerging discipline called "radiogenomics." Optimal evaluation methodologies for radiogenomics have not been well established. We aim to develop a decision framework based on utility analysis to assess predictive models for breast cancer diagnosis. We garnered Gail risk factors, single nucleotide polymorphisms (SNPs), and mammographic features from a retrospective case-control study. We constructed three logistic regression models built on different sets of predictive features: (1) Gail, (2) Gail + Mammo, and (3) Gail + Mammo + SNP. Then we generated receiver operating characteristic (ROC) curves for three models. After we assigned utility values for each category of outcomes (true negatives, false positives, false negatives, and true positives), we pursued optimal operating points on ROC curves to achieve maximum expected utility of breast cancer diagnosis. We performed McNemar's test based on threshold levels at optimal operating points, and found that SNPs and mammographic features played a significant role in breast cancer risk estimation. Our study comprising utility analysis and McNemar's test provides a decision framework to evaluate predictive models in breast cancer risk estimation. *© 2015 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JMI.2.4.041005]

Keywords: receiver operating characteristic methodology; expected utility; breast imaging; genomics; mammography.

Paper 15069SSPRR received Mar. 31, 2015; accepted for publication Jul. 20, 2015; published online Aug. 17, 2015.

1 Introduction

Effective clinical decision making about screening, diagnosis, surgery, and preventive intervention for breast cancer relies on accurate assessment of a patient's cancer risk, which has prompted the development of a number of cancer risk predictive models.[1–9] The "Breast Cancer Risk Assessment Tool" (the Gail model) is a prominent risk predictive model based on self-reported demographic risk factors including age, age at menarche, age at first live birth, number of first-degree relatives with a diagnosis of breast cancer, and number of previous breast biopsies,[2] which has limited discriminatory power. Recent advances in genome-wide association studies (GWAS) and successes with cost reduction in genome-sequencing have paved the road for developing predictive models to potentially estimate breast cancer risk on the basis of both demographic risk factors and genetic variants. On the other hand, there is a long history of risk estimation for breast cancer by using imaging findings.[10–13] Now, it is widely agreed that imaging findings, in concert with genetic variants will likely be necessary for accurate assessment of a patient's breast cancer risk. A promising new paradigm, "radiogenomics," delves into the analysis of the interaction of imaging findings and genetic variants for estimating cancer risk.[14–17]

The performance of predictive models in radiogenomics has typically been evaluated with the area under the receiver operating characteristic (ROC) curve (AUC).[14] Although AUC is a popular statistical measure, the technique has several weaknesses.[18–20] AUC does not take into account the prevalence of disease or the consequence of decisions, which heavily influences the ultimate outcomes of medical decisions. In addition, AUC considers the entire ROC curve while in reality, just a single threshold point matters in decision making. A physician consciously or subconsciously chooses one threshold level of sensitivity/specificity for recommending further management. The recent emphasis on cost-effective medical practice has also strengthened the need to seek the optimal threshold level in ROC curve analysis. Moreover, prior studies have demonstrated that the incremental improvement in AUC is only moderate when some genetic variants that are strongly associated with disease are added to models possessing good discrimination.[3,21,22]

Utility analysis, a fundamentally complementary component of ROC analysis, offers a solution to address weaknesses of

*Address all correspondence to: Elizabeth S. Burnside, E-mail: eBurnside@uwhealth.org

2329-4302/2015/$25.00 © 2015 SPIE

AUC analysis. Utility analysis explicitly considers the clinical consequences of decisions by summing the utility of each possible outcome [true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN)] weighted by the probability of that outcome. The maximization of expected utility occurs at the operating point where a rational physician should make a clinical decision.[23] However, utility analysis has received relatively little attention in practical settings because it requires agreement upon the utility of each outcome.[24,25] In order to sidestep this difficulty, some prior studies have defined a ratio of utilities and estimated the ratio from clinical studies.[25–27] Recent efforts have specified the utilities of different outcomes in breast cancer research,[28] which has engendered the enthusiasm that utility analysis will contribute to the evaluation of predictive models in radiogenomics.

In this study, we aim to develop a decision framework by employing utility analysis to identify optimal operating points and optimally balancing sensitivity and specificity, which allows us to accurately assess predictive models in breast cancer risk estimation. We demonstrate the framework by using an example of imaging findings from mammography and germline genetic variants.

2 Materials and Methods

The Marshfield Clinic Institutional Review Board approved the use of Marshfield Clinic's Personalized Medicine Research Project (PMRP) cohort in the study.

2.1 *Subjects*

We used data from a retrospective case-control study from Marshfield Clinic, the details of which have been previously published.[29] Women of western European heritage with an available plasma sample, a mammogram, and a breast biopsy within 12 months after the mammogram were included. Subjects having no mammography reports were excluded from the study. Subjects having BRCA1 or BRCA2 mutations were also excluded. Cases were defined as women having a confirmed diagnosis of breast cancer obtained from the Institutional Cancer Registry. In our case cohort, we included both invasive breast cancer and ductal carcinoma in situ. Controls were confirmed through the electronic medical records (and absence from the cancer registry) as never having had a breast cancer diagnosis. We employed an age matching strategy, selecting a control whose age was within 5 years of the age of each case in order to ensure similarity in age distribution in the case and control cohorts.

2.2 *Risk Variables*

2.2.1 *Demographic risk factors*

For each subject, we collected demographic risk factors (Gail risk factors): age (at biopsy), age at menarche, number of previous biopsies, and family history of breast cancer. Age at first live birth was not available in our cohort so parity (number of pregnancies) was used instead in our predictive models because of its known association with breast cancer risk and correlation with age at first birth.[30]

2.2.2 *Genetic variants*

For germline genetic variants, we collected 10 commonly used single nucleotide polymorphisms (SNPs) in line with prior large GWAS,[31,32] and used them to predict breast cancer risk (Table 1). We focused on high-frequency/low-penetrance genes that affect breast cancer risk (minor allele frequency >25%) as opposed to low frequency genes with high penetrance (BRCA1 and BRCA2) or intermediate penetrance (CHEK-2). For each SNP, we quantified how many risk alleles were present (0, 1, or 2 risk alleles) as the value.

Table 1 Common genetic variants associated with breast cancer.

Single nucleotide polymorphisms (SNPs)	Chromosome	Gene	Risk allele
RS1045485	2	CASP8	C
RS13281615	8	Unknown	G
RS13387042	2	Unknown	G
RS2981582	10	FGFR2	T
RS3803662	16	TOX3	T
RS3817198	11	LSP1	C
RS889312	5	MAP3K1	C
RS10941679	5	Unknown	G
RS999737	14	RAD51L1	T
RS11249433	1	Unknown	C

2.2.3 *Mammographic features*

At the Marshfield Clinic, mammography results were recorded as free text reports in the electronic health record. We used a parser to extract Breast Imaging-Reporting and Data System (BI-RADS)[33] mammographic features from free text reports.[34] After extraction, every mammographic feature takes the value "present" or "not present." From these features, we selected the most predictive abnormality descriptors based on the literature:[13] mass margin, microcalcification shape, microcalcification morphology, and architectural distortion. For microcalcification features, we consolidated the suspicious morphology descriptors (linear, amorphous, and pleomorphic) and suspicious distribution descriptors (clustered, segmental, linear) into the "present" category; cases lacking any of these descriptors in their records were assigned to the "not present" category. Breast composition was discretized into the four values defined by BI-RADS: predominantly fatty, scattered fibroglandular, heterogeneously dense, or extremely dense. This is regularly reported in mammogram reports, and we consider it as a mammographic feature in this study for predicting breast cancer risk.

2.3 *Utility-Based Decision Framework*

We first constructed three logistic regression models built on different sets of risk variables: (1) Gail model constructed with demographic risk factors only, (2) Gail + Mammo model constructed with demographic risk factors and mammographic features, and (3) Gail + Mammo + SNP model constructed with demographic risk factors, mammographic features, and SNPs. We employed a 10-fold cross-validation to help confirm the validity of predictions. We generated ROC curves, and obtained the AUC as a measure of predictive performance based on the probabilities of malignancy predicted by each of the three models. The AUCs of the models were compared by using the DeLong method.[35] We used a P-value of 0.05 as the threshold for statistical significance testing.

Then we assigned utility values for each category for the outcomes of TN, FP, FN, and TP) as follows:

- We chose TN outcomes as our baseline and assigned a utility of zero.
- We assigned a loss of 4.7 days to the utility of FP, U_{FP} based on the literature.[36,37]
- We used the University of Wisconsin Breast Cancer Simulation (UWBCS) model[38] to estimate the utility of FN as a loss of 2.52 years.[28]
- For TP, we assumed that its utility was $U_{\mathrm{FN}} \times (1 - \alpha)$, $0 \le \alpha \le 1$, where α is an unknown parameter representing the overall effectiveness of breast cancer treatment. In this study, we chose α as 0.86, the 5-year survival rate from Surveillance Epidemiology and End Results[39] program for breast cancer.

The expected utility of a predictive model f is defined as follows:

$$E[U(f)] = p \times [U_{\mathrm{TP}} \times \mathrm{TPR} + U_{\mathrm{FN}} \times (1 - \mathrm{TPR})] + (1 - p) \times [U_{\mathrm{FP}} \times \mathrm{FPR} + U_{\mathrm{TN}} \times (1 - \mathrm{FPR})],$$

where $E[]$ is the expected value of $U(f)$. FPR (false positive rate) and TPR (true positive rate) are the coordinates of a point in ROC space for a given threshold level and p is the prevalence of breast cancer. We considered p to be fixed with a typical value of four breast cancers per 1000 women screened.[40] The maximum expected utility (MEU) is defined as the expected utility at the optimal operating point where the line with slope S is tangent to the ROC curve

$$S = \frac{U_{\mathrm{TN}} - U_{\mathrm{FP}}}{U_{\mathrm{TP}} - U_{\mathrm{FN}}} \times \frac{1 - p}{p}.$$

After binormal ROC curves were generated using ROCKIT software,[41,42] we pursued finding optimal operating points on ROC curves to achieve the MEU of breast cancer diagnosis. We obtained sensitivity, specificity, and threshold level at the optimal operating point. For comparison, we also found sensitivity, specificity, and threshold level when the sum of sensitivity and specificity was maximized. After threshold levels were specified, we used McNemar's test to determine the effects of SNPs and mammographic features in breast cancer risk estimation.

3 Results

We succeeded in identifying 373 cases and 395 controls. The age range (at biopsy) for the subjects in this study was 29 to 90 years of age (mean = 62, standard deviation = 12.8). There were more young people (age < 50) in the case group than in the control group, and the proportion of elderly people (age ≥ 60) was roughly the same in the case group and in the control group (Table 2).

Table 2 Distribution of the subjects by demographic risk factors.

Variables	Controls ($N = 395$)	Cases ($N = 373$)	All subjects ($N = 768$)	Odds ratio
Age (years)				
39 and below	8 (2.03%)	16 (4.29%)	24 (3.12%)	Reference
40 to 49	54 (13.67%)	71 (19.03%)	125 (16.28%)	0.66
50 to 59	128 (32.41%)	82 (21.98%)	210 (27.34%)	0.32
60 to 69	91 (23.04%)	85 (22.79%)	176 (22.92%)	0.47
70 and above	114 (28.86%)	119 (31.90%)	233 (30.34%)	0.52
Age at menarche				
≥14	32 (8.1%)	89 (23.9%)	121 (15.8%)	Reference
12 to 13	98 (24.8%)	157 (42.1%)	255 (33.2%)	0.58
7 to 11	26 (6.6%)	63 (16.9%)	89 (11.6%)	0.87
Missing	239 (60.5%)	64 (17.2%)	303 (39.5%)	NA
No. of biopsies				
0	337 (85.62%)	303 (81.23%)	640 (83.33%)	Reference
1	52 (13.16%)	60 (16.09%)	112 (14.58%)	1.28
≥2	6 (1.52%)	10 (2.68%)	16 (2.08%)	1.85
No. of pregnancies				
0	42 (10.63%)	31 (8.31%)	73 (9.51%)	0.62
1 to 2	125 (31.65%)	126 (33.78%)	251(32.68%)	0.85
3 to 5	168 (42.53%)	163 (43.70%)	331 (43.10%)	0.82
≥6	44 (11.14%)	52 (13.94%)	96 (12.50%)	Reference
Missing	16 (4.05%)	1 (0.27%)	17 (2.21%)	NA
No. of first-degree relatives with breast cancer				
0	325 (82.28%)	268 (71.85%)	593 (77.21%)	Reference
1	57 (14.43%)	91 (24.40%)	148 (19.27%)	1.93
≥2	13 (3.29%)	14 (3.75%)	27 (3.52%)	1.30

To better demonstrate the effects of different risk factors on breast cancer, some exploratory analysis was provided. We summarized the distribution of the subjects by demographic risk factors (Table 2), genetic variants (Table 3), and mammographic features (Table 4).

We found that mammographic features augmented the baseline Gail model in terms of AUC (0.713 versus 0.597) and the P-value was less than 0.001 based on DeLong method (Fig. 1). With threshold levels at optimal operating points when MEU was achieved, subjects were reclassified according to their risk of breast cancer. Using McNemar's test, we found that a statistically significant change in proportions from reclassification occurred between the Gail model and the Gail + Mammo

Table 3 Distribution of the subjects by individual genetic variants.

SNPs	Controls ($N = 395$)	Cases ($N = 373$)	All subjects ($N = 768$)	Odds ratio
RS1045485				
CC	7 (1.77%)	4 (1.07%)	11 (1.43%)	Reference
CG	86 (21.77%)	79 (21.18%)	165 (21.48%)	1.61
GG	302 (76.46%)	290 (77.75%)	592 (77.08%)	1.68
RS13281615				
AA	154 (39.0%)	121 (32.4%)	275 (35.8%)	Reference
AG	184 (46.6%)	181 (48.5%)	365 (47.5%)	1.25
GG	57 (14.4%)	71 (19.0%)	128 (16.7%)	1.59
RS13387042				
AA	89 (22.5%)	126 (33.8%)	215 (28.0%)	2.08
AG	206 (52.2%)	179 (48.0%)	385 (50.1%)	1.28
GG	100 (25.3%)	68 (18.2%)	168 (21.9%)	Reference
RS2981582				
CC	151 (38.2%)	134 (35.9%)	285 (37.1%)	Reference
CT	192 (48.6%)	173 (46.4%)	365 (47.5%)	1.02
TT	52 (13.2%)	66 (17.7%)	118 (15.4%)	1.43
RS3803662				
CC	209 (52.91%)	176 (47.18%)	385 (50.13%)	Reference
CT	156 (39.49%)	169 (45.31%)	325 (42.32%)	1.29
TT	30 (7.59%)	28 (7.51%)	58 (7.55%)	1.11
RS3817198				
CC	31 (7.85%)	36 (9.65%)	67 (8.72%)	1.35
CT	170 (43.04%)	170 (45.58%)	340 (44.27%)	1.16
TT	194 (49.11%)	167 (44.77%)	361 (47.01%)	Reference
RS889312				
AA	196 (49.62%)	175 (46.92%)	371 (48.31%)	Reference
AC	179 (45.32%)	160 (42.90%)	339 (44.14%)	1.00
CC	20 (5.06%)	38 (10.19%)	58 (7.55%)	2.13
RS10941679				
AA	232 (58.73%)	182 (48.79%)	414 (53.91%)	Reference
AG	141 (35.70%)	164 (43.97%)	305 (39.71%)	1.48
GG	22 (5.57%)	27 (7.24%)	49 (6.38%)	1.56
RS999737				
CC	243 (61.52%)	230 (61.66%)	473 (61.59%)	2.18
CT	129 (32.66%)	133 (35.66%)	262 (34.11%)	2.37
TT	23 (5.82%)	10 (2.68%)	33 (4.30%)	Reference
RS11249433				
CC	69 (17.5%)	62 (16.6%)	131 (17.1%)	0.97
CT	187 (47.3%)	182 (48.8%)	369 (48.0%)	1.05
TT	139 (35.2%)	129 (34.6%)	268 (34.9%)	Reference

model (P-value <0.001), which was in concert with the results using the DeLong method. When additional SNPs were added to the Gail + Mammo model, AUC increased to 0.733 and the P-value was 0.071 based on the DeLong method. However, using McNemar's test after optimal operating points were specified with utility analysis, we found the P-value was 0.045, which indicated that SNPs might play a significant role in breast cancer risk estimation.

We also identified operating points on ROC curves when the sum of sensitivity and specificity was maximized. With threshold levels at these operating points, we found that reclassification resulted in a statistically significant change in proportions between the Gail model and the Gail + Mammo model by using McNemar's test ($P\text{-value} = 0.0127$). For the Gail + Mammo

Table 4 Distribution of the subjects by mammographic features.

Variables	Controls ($N = 395$)	Cases ($N = 373$)	All subjects ($N = 768$)	Odds ratio
Breast composition				
Fatty	11 (2.78%)	12 (3.22%)	23 (2.99%)	Reference
Scattered	29 (7.34%)	18 (4.83%)	47 (6.12%)	0.57
Heterogeneous	171 (43.29%)	173 (46.38%)	344 (44.79%)	0.93
Extremely dense	5 (1.27%)	13 (3.49%)	18 (2.34%)	2.38
Missing	179 (45.32%)	157 (42.09%)	336 (43.75%)	NA
Mass margin				
Circumscribed	36 (9.11%)	16 (4.29%)	52 (6.77%)	0.59
Obscured	14 (3.54%)	8 (2.14%)	22 (2.86%)	0.76
Ill-defined	48 (12.15%)	47 (12.60%)	95 (12.37%)	1.30
Spiculated	4 (1.01%)	82 (21.98%)	86 (11.20%)	27.27
Calcification shape	79 (20.00%)	63 (16.89%)	142 (18.49%)	0.81
Calcification distribution	117 (29.62%)	79 (21.18%)	196 (25.52%)	0.64
Architectural distortion	21 (5.32%)	50 (13.40%)	71 (9.24%)	2.76

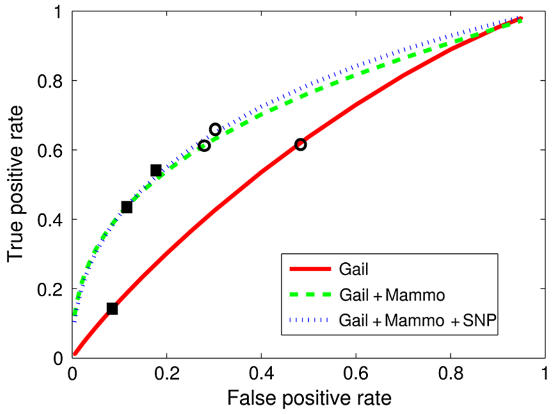

Fig. 1 Receiver operating characteristic curves for the three predictive models. Solid curve, the Gail model; dashed curve, the Gail + Mammo model; dotted curve, the Gail + Mammo + SNP model. Square data points, optimal operating points by maximizing expected utility; round data points, operating points by maximizing the sum of sensitivity and specificity.

Table 5 Comparison of sensitivity and specificity between the method maximizing expected utility and the method maximizing sensitivity and specificity.

	Maximizing expected utility		Maximizing the sum of sensitivity and specificity	
Models	Sensitivity	Specificity	Sensitivity	Specificity
Gail	0.147	0.912	0.610	0.525
Gail + Mammo	0.432	0.887	0.564	0.775
Gail + Mammo + SNP	0.467	0.865	0.603	0.750

model and the Gail + Mammo + SNP model, the P-value was 0.0265 from McNemar's test, which provided evidence that SNPs had a significant predictive effect. These results harmonized with our findings when the expected utility was maximized to pursue optimal operating points.

We specified operating points on ROC curves when MEU was achieved or when the sum of sensitivity and specificity was maximized (Fig. 1), at which we found sensitivities and specificities for the three predictive models (Table 5). Sensitivities generated by utility analysis were lower than those by the method maximizing the sum of sensitivity and specificity. For specificities, utility analysis produced higher values than the method maximizing the sum of sensitivity and specificity.

4 Discussion

We have developed a decision framework combining utility analysis and McNemar's test to evaluate predictive models in breast cancer risk estimation. With traditional ROC analysis and the DeLong method, we found that SNPs augmented the Gail + Mammo model in terms of AUC (0.733 versus 0.713, $P\text{-value} = 0.071$), but the improvement was nonstatistically significant. With our proposed framework, including ROC analysis, utility analysis, and McNemar's test, we found SNPs might play a significant role in breast cancer risk estimation ($P\text{-value} = 0.045$). The difference of the results between the two approaches indicates that the utility framework may have some merits in assessing predictive models.

Our decision framework could be utilized to achieve two important goals in breast cancer risk prediction. One goal is to identify novel biomarkers to improve the accuracy of breast cancer diagnosis in clinical practice. The other goal is to specify optimal operating points in decision making since a physician consciously or subconsciously chooses one threshold point for recommending an operation. There are many methods of identifying the optimal operating points. However, most of them are short of a theoretical foundation.[43] In practice, maximizing the sum of sensitivity and specificity is widely used to identify an operating point in ROC space. Breast cancer is a low prevalence disease which typically results in more FP than TP. In clinics, physicians should select an operating point that yields fewer FP. In ROC space, such an operating point should be chosen from the lower-left quadrant. As we can see in Table 5, specificities generated by using our utility decision framework are higher than those by the method maximizing the sum of sensitivity and specificity, which is in concert with clinic intuition. For identification of optimal operating points, we prefer utility analysis to the method that maximizes the sum of sensitivity and specificity. Utility analysis in our framework leads us to identify optimal operating points by considering different clinical outcomes with scientific justification.

The AUC is a summary of an ROC curve, representing the overall performance of all possible FP fractions, and it is simple for implementation. We believe that AUC analysis will still play an important role in assessing predictive models despite some limitations demonstrated in this study. Our decision framework is not the intent to replace AUC analysis, but rather to augment AUC analysis. Our decision framework provides a new approach for the assessment of predictive models by identifying optimal operating points from a decision analytic standpoint, which creates the opportunity to validate and demonstrate the value of novel and effective biomarkers in breast cancer risk estimation.

The ongoing discovery of new risk factors presents opportunities and challenges to evaluate these risk factors and incorporate them into predictive models. Each SNP will likely contribute a small increase in the predictive ability of these models. Many SNPs with this low-level information will need to substantially improve risk prediction.[20] Prior studies have identified the challenges of using AUC to evaluate the added predictive ability of a new biomarker, and have proposed net reclassification improvement (NRI) analysis to assess the improvement in model performance offered by the new biomarker.[21,22,44] NRI analysis treats each outcome equally but it is rare that different outcomes have the same effect on a patients' quality of life in clinic. Our framework improves NRI analysis by explicitly considering the utility of each outcome to specify optimal operating points. We determine threshold levels at optimal operating points to assess breast cancer predictive models with McNemar's test.

There are several limitations in our study. First, due to the inherent difficulty of collecting a rich multimodality data set, the sample size is small compared with large-scale GWAS. Second, we use logistic regression models to estimate breast cancer risk. A possible line of future research is to employ other predictive models such as Bayesian network, artificial neural network, or support vector machine for validating our results.

Third, we employed 10-fold cross-validation to help confirm the validity of predictions. The Delong method might not be appropriate for comparing AUCs here.[45] We will explore the possibilities of using other statistical tests to compare AUCs. Finally, we obtained the utility of FP from the literature and the utility of TP from the domain knowledge. We plan to use the UWBCS model to obtain both utilities. We also plan to implement sensitivity analysis to demonstrate the robustness of our decision framework to variations in utility specification.

5 Conclusion

Genetic variants and mammographic features have the potential to lead to substantial improvements in breast cancer risk prediction. Our proposed decision framework could be used as a general technique to characterize optimal thresholds and to quantify the potential predictive power of different imaging modalities and biomarkers.

Acknowledgments

The authors gratefully acknowledge the support of the Wisconsin Genomics Initiative, NCI Grant R01CA127379-01 and its ARRA supplement 3R01CA127379-03S1, NIGMS Grant R01GM097618-01, NLM Grant R01LM011028-01, NIEHS Grant 5R01ES017400-03, NIH Grant 1U54AI117924-01, the UW Institute for Clinical and Translational Research, and the UW Carbone Cancer Center.

References

1. J. Dai et al., "Breast cancer risk assessment with five independent genetic variants and two risk factors in Chinese women," *Breast Cancer Res.* **14**(1), R17 (2012).
2. M. H. Gail, "Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model," *JNCI J. Natl. Cancer Inst.* **101**(13), 959–963 (2009).
3. C. Lee et al., "Breast cancer risk assessment using genetic variants and risk factors in a Singapore Chinese population," *Breast Cancer Res.* **16**(3), R64 (2014).
4. J. Liu et al., "Genetic variants improve breast cancer risk prediction on mammograms," in *American Medical Informatics Association Symposium (AMIA)*, Washington, DC (2013).
5. J. Liu et al., "New genetic variants improve personalized breast cancer diagnosis," in *AMIA Summit on Translational Bioinformatics (AMIA-TBI)*, San Francisco, California (2014).
6. C. Meads, I. Ahmed, and R. Riley, "A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance," *Breast Cancer Res. Treat.* **132**(2), 365–377 (2012).
7. A. Quante et al., "Breast cancer risk assessment across the risk continuum: genetic and nongenetic risk factors contributing to differential model performance," *Breast Cancer Res.* **14**(6), R144 (2012).
8. S. Wacholder et al., "Performance of common genetic variants in breast-cancer risk models," *N. Engl. J. Med.* **362**(11), 986–993 (2010).
9. Y. Wu et al., "Comparing the value of mammographic features and genetic variants in breast cancer risk prediction," in *American Medical Informatics Association Symposium (AMIA)*, Washington, DC (2014).
10. E. S. Burnside, D. Rubin, and R. Shachter, "A Bayesian network for mammography," in *American Medical Informatics Association Symposium (AMIA)*, Los Angeles, California (2000).
11. E. S. Burnside et al., "Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings," *Radiology* **251**(3), 663–672 (2009).
12. L. Liberman et al., "The breast imaging reporting and data system: positive predictive value of mammographic features and final assessment categories," *Am. J. Roentgenol.* **171**(1), 35–40 (1998).
13. Y. Wu et al., "A comprehensive methodology for determining the most informative mammographic features," *J. Digital Imaging* **26**(5), 941–947 (2013).
14. M. Kuo and N. Jamshidi, "Behind the numbers: decoding molecular phenotypes with radiogenomics-guiding principles and technical considerations," *Radiology* **270**(2), 320–325 (2014).
15. A. Rutman and M. Kuo, "Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging," *Eur. J. Radiol.* **70**(2), 232–241 (2009).
16. S. Yamamoto et al., "Radiogenomic analysis of breast cancer using MRI: a preliminary study to define the landscape," *Am. J. Roentgenol.* **199**(3), 654–663 (2012).
17. S. Kerns et al., "Radiogenomics: the search for genetic predictors of radiotherapy response," *Future Oncol.* **10**(15), 2391–2406 (2014).
18. J. Eng, "Receiver operating characteristic analysis: utility, reality, covariates, and the future," *Acad. Radiol.* **20**, 795–797 (2013).
19. N. Obuchowski, "ROC analysis," *Am. J. Roentgenol.* **184**, 364–372 (2005).
20. M. S. Pepe and H. E. Janes, "Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer," *JNCI J. Natl. Cancer Inst.* **100**(14), 978–979 (2008).
21. M. Pencina et al., "Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond," *Stat. Med.* **27**(2), 157–172 (2008).
22. M. Pencina, R. S. D'Agostino, and E. Steyerberg, "Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers," *Stat. Med.* **30**(1), 11–21 (2011).
23. H. Sox et al., *Medical Decision Making*, Butterworth-Heinemann, Philadelphia (1988).
24. C. Metz, "ROC analysis in medical imaging: a tutorial review of literature," *Radiol. Phys. Technol.* **1**, 2–12 (2008).
25. R. Wagner, C. Beam, and S. Beiden, "Reader variability in mammography and its implications for expected utility over the population of readers and cases," *Med. Decis. Making* **24**, 561–572 (2004).
26. C. Abbey, M. Eckstein, and J. Boone, "An equivalent relative utility metric for evaluating screening mammography," *Med. Decis. Making* **30**, 113–122 (2010).
27. C. Abbey, M. Eckstein, and J. Boone, "Estimating the relative utility of screening mammography," *Med. Decis. Making* **33**, 510–520 (2013).
28. Y. Wu et al., "Pursuing optimal thresholds to recommend breast biopsy by quantifying the value of tomosynthesis," *Proc. SPIE* **9037**, 90370U (2014).
29. C. A. McCarty et al., "Marshfield clinic personalized medicine research project (PMRP): design, methods and recruitment for a large population-based biobank," *Pers. Med.* **2**(1), 49–79 (2005).
30. C. A. McCarty et al., "The eMERGE network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies," *BMC Med. Genomics* **4**, 13 (2011).
31. D. F. Easton et al., "Genome-wide association study identifies novel breast cancer susceptibility loci," *Nature* **447**(7148), 1087–1093 (2007).
32. D. J. Hunter et al., "A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer," *Nat. Genet.* **39**(7), 870–874 (2007).
33. American College of Radiology, *Breast Imaging Reporting and Data System (BI-RADS®)*, Reston, Virginia (2003).
34. H. Nassif et al., "Information extraction for clinical data mining: a mammography case study," in *IEEE Int. Conf. on Data Mining Workshops* (2009).
35. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics* **44**(3), 837–845 (1988).
36. J. Schousboe et al., "Personalizing mammography by breast density and other risk factors for breast cancer: analysis of health benefits and cost-effectiveness," *Ann. Intern. Med.* **155**(1), 10–20 (2011).
37. J. Brett et al., "The psychological impact of mammographic screening: a systematic review," *Psycho-Oncol.* **14**(11), 917–938 (2005).
38. D. Fryback et al., "The Wisconsin breast cancer epidemiology simulation model," *JNCI Monogr.* **36**, 37–47 (2006).
39. R. L. Gloeckler et al., "Cancer survival and incidence from the Surveillance, Epidemiology, and End Results (SEER) program," *Oncologist* **8**(6), 541–542 (2003).
40. C. Abbey et al., "Statistical properties of a utility measure of observer performance compared to area under the ROC curve," *Proc. SPIE* **8673**, 86730D (2013).
41. C. Metz, "Basic principles of ROC analysis," *Semin. Nucl. Med.* **8**, 283–298 (1978).

42. C. Metz, B. Herman, and J. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Stat. Med.* **17**, 1033–1053 (1998).
43. S. Cantor et al., "A comparison of C/B ratios from studies using receiver operating characteristic curve analysis," *J. Clin. Epidemiol.* **52**(9), 885–892 (1999).
44. M. Leening et al., "Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide," *Ann. Intern. Med.* **160**(2), 122–131 (2014).
45. W. Chen et al., "On the assessment of the added value of new predictive biomarkers," *BMC Med. Res. Methods* **13**, 98 (2013).

Yirong Wu is currently an associate scientist in the Department of Radiology at the University of Wisconsin-Madison. He works in the areas of machine learning, imaging informatics, medical image quality assessment and evaluation, genomics, personalized medicine, computer assisted diagnosis, image processing, and computer vision.

Craig K. Abbey is an applied mathematician in the Department of Psychological and Brain Sciences at the UC Santa Barbara. He works in the field of medical image quality assessment and evaluation. He has investigated a variety of imaging modalities including SPECT and PET, x-ray fluoroscopy, breast CT, and ultrasound. He has also been active in the development of ideal observer models for benchmarking the performance of human observers.

Xianqiao Chen is currently a professor in the School of Computer Science and Technology and the director in the Institute of Internet of Things at Wuhan University of Technology. His research areas are image processing, pattern recognition, simulation, communications and control, internet of things, GIS, and software engineering.

Jie Liu is currently a postdoctoral fellow in the eScience Institute and Department of Genome Sciences at the University of Washington, Seattle, USA. He received his PhD in computer science from University of Wisconsin, Madison, USA, in 2014. His research areas are machine learning, statistics, bioinformatics, and medical informatics.

David C. Page is now a Vilas Distinguished Achievement Professor in the School of Medicine and Public Health at the University of Wisconsin-Madison, Department of Biostatistics and Medical Informatics. He works on developing and applying machine learning algorithms to biomedical data, especially electronic health records, genomics (sequence and SNPs), gene expression (RNAseq and gene chips), and mass spectrometry proteomics and metabolomics data. He also works on privacy issues with such data.

Oguzhan Alagoz is currently an associate professor of Industrial and Systems Engineering and Population Health Sciences at the University of Wisconsin-Madison. His research interests include medical decision making, stochastic optimization, completely and partially observable Markov decision processes, simulation, risk-prediction modeling, personalized medicine, and health technology assessment. He is on the editorial boards of *Operations Research, Medical Decision Making, IIE Transactions,* and *IIE Transactions on Healthcare Engineering*.

Peggy Peissig is an associate research scientist with 25 years of experience in health care and research informatics. Her primary research areas include electronic health record phenotyping, pharmacogenetics, and adverse drug event prediction and surveillance. She holds the John Melski Endowed Distinguished Scientist Chair for biomedical informatics at the Marshfield Clinic where she is the director of the Biomedical Informatics Research Center.

Adedayo A. Onitilo is Marshfield Clinic's oncology service line director, department chair for East District Cancer Care, director of the East District Breast Cancer Program, coprincipal investigator of the Community Clinical Oncology Program and associate editor for *Clinical Medicine and Research*, a publication of original scientific medical research. He also holds a clinical faculty position at the University of Wisconsin-Madison.

Elizabeth S. Burnside is currently a professor and vice chair of the Radiology Department in the University of Wisconsin School of Medicine and Public Health. She is a subspecialty trained breast imager with an active clinical practice. Her research investigates the use of artificial intelligence methods to improve decision making in breast imaging. She is an Executive Leadership in Academic Medicine fellow.