Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 2015 February 21; 9416: . doi:10.1117/12.2081954.

Developing a clinical utility framework to evaluate prediction models in radiogenomics

Yirong Wu^a, Jie Liu^b, Alejandro Munoz del Rio^{a,c}, David C. Page^b, Oguzhan Alagoz^d, Peggy Peissig^e, Adedayo A. Onitilo^{e,f}, and Elizabeth S. Burnside^a

^aDept. of Radiology, UW Madison, WI, USA

^bDept. of Computer Science, UW Madison, WI, USA

^cDept. Of Medical Physics, UW Madison, WI, USA

^dDept. of Industrial and Systems Engineering, UW Madison, WI, USA

^eMarshfield Clinic Research Foundation, Marshfield, WI, USA

^fDept. of Hematology/Oncology, Marshfield Clinic Weston Center, Weston, WI, USA

Abstract

Combining imaging and genetic information to predict disease presence and behavior is being codified into an emerging discipline called “radiogenomics.” Optimal evaluation methodologies for radiogenomics techniques have not been established. We aim to develop a clinical decision framework based on utility analysis to assess prediction models for breast cancer. Our data comes from a retrospective case-control study, collecting Gail model risk factors, genetic variants (single nucleotide polymorphisms-SNPs), and mammographic features in Breast Imaging Reporting and Data System (BI-RADS) lexicon. We first constructed three logistic regression models built on different sets of predictive features: (1) Gail, (2) Gail+SNP, and (3) Gail+SNP+BI-RADS. Then, we generated ROC curves for three models. After we assigned utility values for each category of findings (true negative, false positive, false negative and true positive), we pursued optimal operating points on ROC curves to achieve maximum expected utility (MEU) of breast cancer diagnosis. We used McNemar’s test to compare the predictive performance of the three models. We found that SNPs and BI-RADS features augmented the baseline Gail model in terms of the area under ROC curve (AUC) and MEU. SNPs improved sensitivity of the Gail model (0.276 *vs.* 0.147) and reduced specificity (0.855 *vs.* 0.912). When additional mammographic features were added, sensitivity increased to 0.457 and specificity to 0.872. SNPs and mammographic features played a significant role in breast cancer risk estimation (*p*-value < 0.001). Our decision framework comprising utility analysis and McNemar’s test provides a novel framework to evaluate prediction models in the realm of radiogenomics.

Keywords

ROC methodology; expected utility; breast imaging; genetics; mammography

1. INTRODUCTION

Effective clinical decision making about screening, diagnosis, surgery, and preventive intervention for cancer relies on accurate assessment of a patient's cancer risk, which has prompted the development of a number of cancer risk prediction models¹⁻⁹. The "Breast Cancer Risk Assessment Tool" (the Gail model) is a prominent risk prediction model based on self-reported demographic risk factors including age, age at menarche, age at first live birth, the number of first-degree relatives with a diagnosis of breast cancer, and number of previous breast biopsies², 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 prediction models to potentially estimate breast cancer risk on the basis of both demographic risk factors and genetic variants. However, early attempts to use genetic variants to predict breast cancer risk have demonstrated only modest improvements over conventional demographic risk factors^{2, 8, 10}. On the other hand, there is a long history of risk estimation for breast cancer by using imaging findings garnered from mammograms¹¹⁻¹⁴. However, it has proven a difficult task to characterize the molecular properties of malignancies using imaging findings. Now it is widely agreed that imaging findings, in concert with genetic variants will likely be necessary for accurate assessment of a patient's cancer risk. A promising new paradigm, radiogenomics delves into the analysis of the interaction of imaging findings, genetic variants, and cancer risk¹⁵⁻¹⁸.

The performance of prediction models in radiogenomics has typically been evaluated with the area under the receiver operating characteristic (ROC) curve (AUC)¹⁵. Although AUC is a popular statistical measure, the technique has several weaknesses¹⁹⁻²¹. AUC does not account for the prevalence of disease or the consequence of decisions, which heavily influences the ultimate outcomes of medical decisions. AUC considers the entire ROC curve while in reality just a single threshold point matters in decision making. A radiologist consciously or subconsciously chooses one threshold point of sensitivity/specificity for recommending further management. Of note, prior studies demonstrate that the increment of AUC was only moderate when some genetic variants that were strongly associated with breast cancer were added to models possessing reasonable good discrimination^{3, 22, 23}.

Expected utility (EU) analysis is a fundamentally complementary component of ROC analysis, which explicitly considers the clinical consequences of decisions by summing the utility of each possible outcome (true positive, false positive, true negative, and false negative) weighted by the probability of that outcome. The maximization of EU occurs at the operating point where a rational radiologist should make a clinical decision²⁴. In this study, we aim to develop a decision framework by employing utility analysis to induce sensitivity and specificity, which allows us accurately assess prediction models in radiogenomics.

2. MATERIALS and METHODS

2.1 Subjects

We used data from a retrospective case-control study from Marshfield Clinic's Personalized Medicine Research Project (PMRP)²⁵. Women of western European heritage with a plasma sample available, a mammogram, and a breast biopsy within 12 months after the mammogram were included. Subjects who have no mammography reports were excluded from the study. Subjects who have 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 five years of the age of each case in order to ensure similarity in age distribution in the case and control cohorts. We succeeded in identifying 373 cases and 395 controls. The age range for these subjects was 29 to 90 years of age (mean = 62, standard deviation = 12.8).

2.2 Risk variables

For each subject, we collected Gail risk factors: age (at biopsy), age at menarche, number of previous biopsies, and family history of breast cancer (Table 1). Age at first live birth was not available in our cohort so parity (number of pregnancies) was instead used in our prediction models because of its known association with breast cancer risk and correlation with age at first birth²⁶. For genetic variants, we collected 10 commonly used single nucleotide polymorphisms (SNPs) in line with prior studies^{27, 28}. 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 risky alleles were present (0, 1, or 2 risky alleles) as the value. For mammography findings described in Breast Imaging-Reporting and Data System (BI-RADS) lexicon²⁹, we collected 5 most informative mammographic features¹⁴ (Table 1).

2.3 Utility based decision framework

We first constructed three logistic regression models built on different sets of predictive features: (1) Gail, (2) Gail + SNP, and (3) Gail + SNP + BI-RADS. We generated receiver operator characteristic (ROC) curves using ROCKIT software^{30, 31} based on the probabilities of malignancy predicted by each of the three models.

Then, we assigned utility values for each category of findings (True Negative (TN), False Positive (FP), False Negative (FN) and True Positive (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^{32, 33}.
- We used the University of Wisconsin Breast Cancer Simulation (UWBCS) model³⁴ to estimate the utility of FN as a loss of 2.52 years³⁵,

- For TP, we assumed that its utility was $U_{FN} \times (1-\alpha)$, $0 \leq \alpha \leq 1$, where α is an unknown parameter representing the overall effectiveness of breast cancer treatment. In this study, we chose α as 0.86, the five-year survival rate from Surveillance Epidemiology and End Results (SEER)³⁶ program for breast cancer.

The expected utility of a prediction model f is defined as follows.

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

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 typical value of 4 breast cancers per 1000 women screened³⁷. The maximum expected utility (MEU) is defined as expected utility at the optimal operating point where the line with slope S is tangent to the ROC curve.

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

We pursued finding optimal operating points on ROC curves to achieve the maximum expected utility of breast cancer diagnosis. We obtained sensitivity, specificity, and threshold level at the optimal operating point. McNemar's test was used to evaluate the difference of predictive performance for the three models, demonstrating the effects of SNPs and mammographic features in breast cancer risk estimation.

3. RESULTS

We found that SNPs augmented the baseline Gail model in terms of AUC (0.638 vs. 0.597, p -value = 0.0124) (Figure 1). When additional mammographic features were added, AUC increased to 0.733. Both Gail + SNP and Gail + SNP + BI-RADS models provided improved MEU compared with the Gail model (Figure 2).

For the Gail model, the threshold at the optimal operating point was 0.644 when MEU was achieved, its sensitivity was 0.147, and its specificity was 0.912. When SNPs were added for risk prediction, at the optimal threshold of 0.659, the sensitivity of Gail + SNP model improved to 0.276 while its specificity decreased to 0.855. When additional mammographic features were added, the sensitivity of Gail + SNP + BI-RADS model improved to 0.457 and its specificity became 0.870 at the optimal threshold level of 0.643 (Figure 3).

With threshold levels at the optimal operating points, subjects could be reclassified according to their risk of breast cancer (Table 2 & 3). Using McNemar's test, we found that SNPs and mammographic features played a significant role in breast cancer risk estimation (p -value < 0.001). For this case/control study, when SNPs were added to the Gail model, 59 subjects were upgraded to high risk. Similarly, when both SNPs and BI-RADS features were added to the Gail model, 121 subjects were upgraded.

4. DISCUSSION

We have developed a novel decision framework combining utility analysis and McNemar's test to efficiently evaluate prediction models in the new discipline of radiogenomics. Specifically, we determine threshold levels at optimal operating points to measure the performance of breast cancer prediction models in terms of sensitivity and specificity. Understanding optimal operating points from a decision analytic standpoint creates the opportunity for developing prediction models to look for novel and effective biomarkers in clinical practice to ultimately improve care.

The quest for novel and effective risk factors has been revitalized with the emergence of new technologies that generate a large amount of multi-modality datasets in the era of "Big Data". Our decision framework could be utilized to achieve two important goals in breast cancer risk prediction. The first goal is to identify effective biomarkers. In order to use risk factors in predictive models effectively, it is necessary to quantify the value of these risk factors and select important variables when a large amount of risk factors available. In breast cancer risk prediction, accuracy is a high priority and efficiency is an important second goal. Identifying important variables may help improve the efficiency of breast cancer risk interpretation directly in order to enable physicians to focus on the most important features. The other goal is to specify optimal operating points in decision making, since a radiologist consciously or subconsciously chooses one threshold point for recommending operation. There are a lot of methods to identify the optimal operating points. However, most of them are short of a theoretical foundation³⁸. Utility analysis in our framework leads us to identify optimal operating points by considering different clinical outcomes with scientific justification.

Expected utility analysis leverages ROC analysis by explicitly considering the clinical consequences of decisions. However, EU analysis has received relatively little attention in practical settings because it requires the agreement upon the utility of each outcome. There are two potential approaches to expected utility analysis. One is to specify the utility of each outcome directly and implement the analysis. The other is to define a ratio of utilities and to estimate the ratio from clinical studies that involve patient management^{39–41}. In this study, we follow the first approach to implement utility analysis by specifying the utilities of outcomes based on a simulation model and the literature, which demonstrates that it is very promising to use utility analysis to evaluate prediction models even though the challenges of utility specification exist.

The ongoing discovery of new risk factors presents opportunities and challenges for researchers to evaluate these risk factors and incorporate them into prediction models. In genetics studies, each SNP is expected to contribute a small amount to the capability of prediction models. Many more SNPs with this low-level information will need to be discovered to substantially improve risk prediction²¹. Prior studies have identified the challenges of using AUC to evaluate the added predictive ability of a new biomarker, and proposed net reclassification improvement (NRI) analysis to assess the improvement in mode performance offered by a new biomarker^{22, 23, 42}. NRI analysis would accurately capture the low utility of a biomarker in the new discipline of radiogenomics but it treats

each outcome equally. However, it is rare that different outcomes have the same effect on a patients' quality of life in clinic. Our framework improves NRI analysis by considering the utility of each possible outcome explicitly to specify an optimal operating point. We determine threshold levels at optimal operating points to measure the performance of breast cancer prediction models in terms of sensitivity and specificity.

There are several limitations to our study. First, the sample size is small compared with large-scale genome-wide association studies, due to the inherent difficulty of collecting a rich multi-modality dataset. Second, our study used logistic regression models to estimate breast cancer risk. A possible line of future research is to employ other prediction models such as Bayesian network, artificial neural network, or support vector machine for validating our results. Third, we obtained the utility of false positive from the literature and the utility of true positive from domain knowledge. We plan to use UWBCS model to obtain both utilities. Finally, we need to validate our decision framework using other study populations.

5. CONCLUSIONS

Genetic variants and mammographic features have the potential to lead to substantial improvements in breast cancer risk prediction. In addition, our proposed decision framework could be used as a general technique to characterize optimal thresholds and to quantify 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, the UW Institute for Clinical and Translational Research (ICTR) and the UW Carbone Cancer Center.

REFERENCES

1. Dai J, Hu Z, Jiang Y, et al. Breast cancer risk assessment with five independent genetic variants and two risk factors in Chinese women. *Breast Cancer Res.* 2012; 14(1):R17. [PubMed: 22269215]
2. Gail MH. Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. *J Natl Cancer Inst.* 2009; 101(13):959–963. [PubMed: 19535781]
3. Lee C, Irwanto A, Salim A, et al. Breast cancer risk assessment using genetic variants and risk factors in a Singapore Chinese population. *Breast Cancer Res.* 2014; 16(3):1–13.
4. Liu J, Page D, Nassif H, et al. Genetic variants improve breast cancer risk prediction on mammograms. *American Medical Informatics Association Symposium (AMIA).* 2013
5. Liu J, Page D, Peissig P, et al. New genetic variants improve personalized breast cancer diagnosis. *AMIA Summit on Translational Bioinformatics (AMIA-TBI).* 2014
6. Meads C, Ahmed I, Riley R. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. *Breast Cancer Res Treat.* 2012; 132(2):365–377. [PubMed: 22037780]
7. Quante A, Whittemore A, Shriver T, et al. Breast cancer risk assessment across the risk continuum: genetic and nongenetic risk factors contributing to differential model performance. *Breast Cancer Res.* 2012; 14(6):R144. [PubMed: 23127309]
8. Wacholder S, Hartge P, Prentice R, et al. Performance of common genetic variants in breast-cancer risk models. *N Engl J Med.* 2010; 362(11):986–993. [PubMed: 20237344]

9. Wu Y, Liu J, Page D, et al. Comparing the value of mammographic features and genetic variants in breast cancer risk prediction. *Proceedings of the American Medical Informatics Association (AMIA)*. 2014
10. Gail MH. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *J Natl Cancer Inst*. 2008; 100(14):1037–1041. [PubMed: 18612136]
11. Burnside ES, Rubin D, Shachter R. A Bayesian network for mammography. *American Medical Informatics Association Symposium (AMIA)*. 2000:106–110.
12. Burnside ES, Davis J, Chhatwal J, et al. Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings. *Radiology*. 2009; 251(3):663–672. [PubMed: 19366902]
13. Liberman L, Abramson AF, Squires FB, et al. The breast imaging reporting and data system: positive predictive value of mammographic features and final assessment categories. *AJR Am J Roentgenol*. 1998; 171(1):35–40. [PubMed: 9648759]
14. Wu Y, Alagoz O, Ayvaci MU, et al. A comprehensive methodology for determining the most informative mammographic features. *J Digit Imaging*. 2013; 26(5):941–947. [PubMed: 23503987]
15. Kuo M, Jamshidi N. Behind the numbers: Decoding molecular phenotypes with radiogenomics--guiding principles and technical considerations. *Radiology*. 2014; 270(2):320–325. [PubMed: 24471381]
16. Rutman A, Kuo M. Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging. *Eur J Radiol*. 2009; 70(2):232–241. [PubMed: 19303233]
17. Yamamoto S, Maki D, Korn R, et al. Radiogenomic analysis of breast cancer using MRI: a preliminary study to define the landscape. *AJR Am J Roentgenol*. 2012; 199(3):654–663. [PubMed: 22915408]
18. Kerns S, West CL, Andreassen C, et al. Radiogenomics: the search for genetic predictors of radiotherapy response. *Future Oncol*. 2014; 10(15):2391–2406. [PubMed: 25525847]
19. Eng J. Receiver operating characteristic analysis: utility, reality, covariates, and the future. *Acad Radiol*. 2013; 20:795–797. [PubMed: 23747151]
20. Obuchowski N. ROC analysis. *AJR Am J Roentgenol*. 2005; 184:364–372. [PubMed: 15671347]
21. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *J Natl Cancer Inst*. 2008; 100(14):978–979. [PubMed: 18612128]
22. Pencina M, D'Agostino RS, D'Agostino RJ, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008; 27(2): 157–172. [PubMed: 17569110]
23. Pencina M, D'Agostino RS, Steyerberg E. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2011; 30(1):11–21. [PubMed: 21204120]
24. Sox, H.; Blatt, M.; Higgins, M., et al. *Medical Decision Making*. Philadelphia: Butterworth-Heinemann; 1988.
25. McCarty CA, Wilke RA, Giampietro PF, et al. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. *Personalized Med*. 2005; 2(1):49–79.
26. McCarty CA, Chisholm RL, Chute CG, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics*. 2011; 4:13. [PubMed: 21269473]
27. Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*. 2007; 447(7148):1087–1093. [PubMed: 17529967]
28. Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet*. 2007; 39(7):870–874. [PubMed: 17529973]
29. Breast Imaging Reporting And Data System (BI-RADS®). Reston VA: American College of Radiology; 2003.
30. Metz C. Basic principles of ROC analysis. *Semin Nucl Med*. 1978; 8:283–298. [PubMed: 112681]

31. Metz C, Herman B, Shen J. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Stat Med*. 1998; 17:1033–1053. [PubMed: 9612889]
32. Schousboe J, Kerlikowske K, Loh A, et al. Personalizing mammography by breast density and other risk factors for breast cancer: analysis of health benefits and cost-effectiveness. *Ann Intern Med*. 2011; 155(1):10–20. [PubMed: 21727289]
33. Brett J, Bankhead C, Henderson B, et al. The psychological impact of mammographic screening. A systematic review. *Psychoncology*. 2005; 14(11):917–938.
34. Fryback D, Stout N, Rosenberg M, et al. The Wisconsin breast cancer epidemiology simulation model. *J Natl Cancer Inst Monogr*. 2006; 36:37–47. [PubMed: 17032893]
35. Wu Y, Alagoz O, Vanness DV, et al. Pursuing optimal thresholds to recommend breast biopsy by quantifying the value of tomosynthesis. *SPIE Medical Imaging*. 2014
36. Gloeckler RL, Reichman M, Lewis D, et al. Cancer survival and incidence from the Surveillance, Epidemiology, and End Results (SEER) program. *Oncologist*. 2003; 8(6):541–552. [PubMed: 14657533]
37. Abbey C, Samuelson F, Gallas B, et al. Statistical properties of a utility measure of observer performance compared to area under the ROC curve. *SPIE Medical Imaging*. 2013
38. Cantor S, Sun C, Tortolero-Luna G, et al. A comparison of C/B ratios from studies using receiver operating characteristic curve analysis. *J Clin Epidemiol*. 1999; 52(9):885–892. [PubMed: 10529029]
39. Abbey C, Eckstein M, Boone J. An equivalent relative utility metric for evaluating screening mammography. *Medical Decision Making*. 2010; 30:113–122. [PubMed: 19706880]
40. Abbey C, Eckstein M, Boone J. Estimating the relative utility of screening mammography. *Medical Decision Making*. 2013; 33:510–520. [PubMed: 23295543]
41. Wagner R, Beam C, Beiden S. Reader variability in mammography and its implications for expected utility over the population of readers and cases. *Medical Decision Making*. 2004; 24:561–572. [PubMed: 15534338]
42. Leening M, Vedder M, Witteman J, et al. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Ann Intern Med*. 2014; 160(2):122–131. [PubMed: 24592497]

Figure 1. ROC curves for three models. Solid curve, Gail; dotted curve, Gail + SNP; dashed curve, Gail + SNP + BI-RADS

Figure 2.
Expected utility curves of the three models. Solid curve, Gail; dotted curve, Gail + SNP;
dashed curve, Gail + SNP + BI-RADS

Figure 3.

Optimal operating points on ROC curves at maximum expected utility. Solid curve, Gail model; dotted curve, Gail + SNP model; dashed curve, Gail + SNP + BI-RADS model

Table 1

Predictive variables used in three models

Gail risk factors	SNPs	Mammographic features
Age of biopsies	rs1045485	Mass margins
Age of menarche	rs13281615	Calcification shape
Number of previous biopsies	rs13387042	Calcification distribution
Number of first-degree relative with breast cancer	rs2981582	Architectural distortion
Parity	rs3803662	Breast composition
	rs3817198	
	rs889312	
	rs7716600	
	rs11249433	
	rs999737	

Table 2

Reclassification table based on threshold levels at optimal operating points for Gail model and Gail+SNP model

		Gail + SNP		Total
		0.659	< 0.659	
Gail model	0.644	64	29	93
	< 0.644	88	587	675
Total		152	616	768

Table 3

Reclassification table based on threshold levels at optimal operating points for Gail model and Gail+SNP+BI-RADS model

		Gail + SNP + BI-RADS		Total
		0.643	< 0.643	
Gail model	0.644	55	38	93
	< 0.644	159	516	675
Total		214	554	768