Leveraging Interaction between Genetic Variants and Mammographic Findings for Personalized Breast Cancer Diagnosis

Jie Liu, PhD[1], Yirong Wu, PhD[1], Irene Ong, PhD[1], David Page, PhD[1], Peggy Peissig, PhD[2], Catherine McCarty, PhD[3], Adedayo A. Onitilo, MD, MSCR, FACP[2,4] and Elizabeth Burnside, MD, MPH, MS[1]

[1] University of Wisconsin, Madison, WI, US

[2] Marshfield Clinic Research Foundation, Marshfield, WI, US

[3] Essentia Institute of Rural Health, Duluth, MN, US

[4]Department of Hematology/Oncology, Marshfield Clinic Weston Center, Weston, WI, US

Abstract

Recent large-scale genome-wide association studies (GWAS) have identified a number of genetic variants associated with breast cancer which showed great potential for clinical translation, especially in breast cancer diagnosis via mammograms. However, the amount of interaction between these genetic variants and mammographic features that can be leveraged for personalized diagnosis remains unknown. Our study utilizes germline genetic variants and mammographic features that we collected in a breast cancer case-control study. By computing the conditional mutual information between the genetic variants and mammographic features given the breast cancer status, we identified six interaction pairs which elevate breast cancer risk and five interaction pairs which reduce breast cancer risk.

1. Introduction

In the last couple of years, the field of genome-wide association studies has made tremendous progress, and a number of genetic variants (single-nucleotide polymorphisms, or SNPs) have been identified to be associated with breast cancer [1], providing the opportunity to use patients' genetic information for personalized medicine. However, the rapid progress within GWAS has been both an opportunity and a challenge. The large number of variants found to be associated with breast cancer provide low signal, and their contribution over and above other conventional risk factors for breast cancer risk prediction and clinical diagnosis remains difficult to evaluate in terms of clinical significance.

Perhaps the most important question is whether germline genetic variants can further improve the prediction of future breast cancer evolvement in order to influence care. Many in the scientific community admit that genotype manifests itself as phenotype based on a m d range of environmental factors; hence breast cancer risk determination should involve a combination of genetic and phenotypic ("imaging") traits. Understanding the interaction effect is extremely critical to connect genetic variants with mammographic features, especially as we are just beginning to understand the biological mechanism of cancers [2]. On one hand, the germline genetic variants are the genetic information shared by all the normal cells in the patient's body, and the information is static all through the life of the patient but provides the genetic background of abnormalities within the developed tumors. On the other hand, the features we observe on the mammograms provide a closer portrait of the tumor at the time of diagnosis, but the tumor is dynamic over time and the mammographic information is on the tissue level rather than the molecular level. In the perfect situation, we would like to keep track of the somatic genetic information within the tumor over the time, but the current biotechnology and computation facility do not support it yet. Therefore, the true value of combining the germline DNA and mammographic features is from leveraging the interaction effect between the genetic variants and environmental exposures, the two key determinants in the development and prognosis of breast cancer. Imaging features (like mammographic findings), can provide a window (or "intermediate phenotype") into the complex interactions between genetics and the environment in order to predict individual disease risk. Thus a specific combination of a genetic variant and a mammographic feature may increase or decrease breast cancer risk due to these two converging etiologies. We hypothesize that the combined features can be used to predict the likelihood of breast cancer as well as the prognosis to inform future prevention, early detection, and treatment.

In this paper, we explore the interaction effect between the breast-cancer-associated genetic variants and the mammographic features. We calculate conditional mutual information between SNPs and mammographic features given breast cancer status variable. In total, we identified six interaction pairs that increase breast cancer risk when they present together. We also identified five interaction pairs that decrease breast cancer risk when they present together.

2. Data

[Subjects] The subjects were sampled through the Personalized Medicine Research Project [3] at the Marshfield Clinic. The project was reviewed and approved by the Marshfield Clinic IRB. The subjects were from a retrospective case-control design, and used in our previous study [4]. In our study, each subject must have a plasma sample from which we can genotype the genetic variants, a diagnostic mammogram, and a follow-up breast biopsy within 12 months after the mammogram. Cases were defined as women having a confirmed diagnosis of breast cancer, which was obtained from the institutional cancer registry. Controls were confirmed through the Marshfield Clinic electronic medical records as never having had a breast cancer diagnosis by ICD-9 diagnosis code. Cases included both invasive breast cancer and ductal carcinoma in situ. We used an age matching strategy to construct case and control groups that were similar in age distribution. Specifically, we selected a control whose age was within five years of the age of each case. We decided to focus on high-frequency/low-penetrance SNPs that affect breast cancer risk as opposed to low frequency SNPs with high penetrance or intermediate penetrance. Individuals with a known high-penetrance genetic mutation, including the BRCA1 and BRCA2 mutations, were excluded. In total, there are 336 cases and 375 controls.

[Genetic Variants] Our study included the genetic variants which were identified by the recent large-scale genome-wide association studies. In the previous study of Liu et al. (2013) [4], we performed the same interaction analysis for the 22 SNPs identified before 2010; hence, in this paper we focus on the 55 new SNPs which have been identified since 2010. Among the 55 SNPs, 41 were identified by COGS [5] and 14 SNPs were included based on several other recent studies [6-12].

[Mammography Features] Mammography is the most common breast cancer screening test, and the only one supported by multiple randomized trials demonstrating reduction in mortality rate [13]. There is a long history of development and codification of features observed by radiologists on mammograms. The American College of Radiology developed the BI-RADS lexicon to standardize mammographic findings and recommendations. The BI-RADS lexicon consists of 49 descriptors, including the characteristics of masses and microcalcifications, background breast density and other associated findings. Mammography data was historically recorded as free text reports in the electronic health records, and thus it was difficult to directly access the information contained therein. We used a parser to extract these mammography features from the text reports; the parser was shown to outperform manual extraction [14, 15]. After extraction, each mammography feature took the value "present" or "not present" except that the variable *mass size* was discretized into three values, "not present", "small" and "large", depending on whether there was a reported mass size and whether any dimension was larger than 30mm.

3. Methods

In this paper, we focus on the interaction effect between the breast cancer associated genetic variants and the mammographic features. We use conditional mutual information between SNPs and mammographic features given breast cancer status variable. Conditional mutual information (CMI) between a discrete feature X and a discrete feature Y given a discrete response Z is

$$CMI(X,Y|Z) = \sum_z P(Z=z) \sum_x \sum_y P(X=x, Y=y|Z=z) \log \frac{P(X=x, Y=y|Z=z)}{P(X=x|Z=z)P(Y=y|Z=z)}.$$

We also calculate the 95% confidence intervals for the CMI between each SNP and each mammography feature via bootstrapping. We randomly draw samples with replacement from the subjects, and calculate the conditional mutual information from the samples. We bootstrap for 1,000 times and calculate the corresponding 1,000 CMI values. We sort the 1,000 CMI values from the smallest to the largest, and report the 26-th smallest value and the 26-th largest value as the boundaries of the 95% confidence interval.

One subtlety during the calculation is how to code the genetic variants. Ideally, it is desirable to code each individual SNP as the three genotypes values, namely the risk allele homozygous carrier, the risk allele heterozygous carrier and the non-risk allele homozygous carrier. However, due to the limited number of samples in our cohort and that we usually need sufficient samples for each configuration (a combination of genetic variable, mammographic feature and case/control status) for a reliable estimate of the conditional mutual information, we code each SNP as a binary variable, namely whether the subject carries the risk allele.

4. Results

For each SNP, we find the top three mammographic features which have the greatest conditional mutual information. We also calculated the 95% confidence intervals for these pairs. There are in total 11 pairs with CMI significantly greater than zero. The 11 interaction pairs are summarized in Table 1. Among them, there are six interaction pairs which increase the breast cancer risk and five interaction pairs which decrease breast cancer risk when the specific allele of the genetic variant and the specific mammographic feature present at the same time. The six imaging-genetic pairs that increase risk are summarized in Table 2. The five imaging-genetic pairs that decrease breast cancer risk are summarized in Table 3.

Table 1. The interaction between SNPs and imaging features.

SNP ID	Imaging features	CMI	95% CI
rs9790517	heterogeneous breast composition	0.008	(0.001, 0.023)
rs10472076	indistinct mass margin	0.012	(0.004, 0.026)
rs10472076	linear distribution of calcifications	0.006	(0.001, 0.023)
rs11242675	grouped distribution of calcifications	0.007	(0.001, 0.021)
rs13281615	irregular mass shape	0.008	(0.002, 0.024)
rs17817449	large mass size	0.005	(0.001, 0.020)
rs11552449	dystrophic calcifications	0.010	(0.004, 0.021)
rs12493607	heterogeneous breast composition	0.009	(0.001, 0.027)
rs4973768	indistinct mass margin	0.006	(0.001, 0.021)
rs10759243	linear distribution of calcifications	0.007	(0.001, 0.021)
rs17356907	lobular mass shape	0.005	(0.001, 0.020)

Table 2. The contingency tables for the six imaging-genetic pairs that increase breast cancer risk.

	Case		Ctrl	
rs9790517	Carry T	Not Carry T	Carry T	Not Carry T
heterogeneous breast composition present	41	34	21	46
heterogeneous breast composition not present	99	162	122	186
rs10472076	Carry C	Not Carry C	Carry C	Not Carry C
indistinct mass margin present	39	7	23	17
indistinct mass margin not present	176	114	206	129
rs10472076	Carry C	Not Carry C	Carry C	Not Carry C
linear distribution of calcifications present	13	3	6	10
linear distribution of calcifications not present	202	118	223	136
rs11242675	Carry C	Not Carry C	Carry C	Not Carry C
grouped distribution of calcifications present	40	22	52	58
grouped distribution of calcifications not present	177	97	165	100
rs13281615	Carry G	Not Carry G	Carry G	Not Carry G
irregular mass shape present	69	18	20	8
irregular mass shape not present	162	87	207	140
rs17817449	Carry T	Not Carry T	Carry T	Not Carry T
large mass size present	21	2	7	5
large mass size not present	263	50	306	57

Table 3. The contingency tables for the five imaging-genetic pairs that decrease breast cancer risk.

	Case		Ctrl	
rs11552449	Carry T	Not Carry T	Carry T	Not Carry T
dystrophic calcifications present	0	10	5	4
dystrophic calcifications not present	96	230	102	264
rs12493607	Carry C	Not Carry C	Carry C	Not Carry C
heterogeneous breast composition present	33	42	36	31
heterogeneous breast composition not present	166	95	169	139
rs4973768	Carry T	Not Carry T	Carry T	Not Carry T
indistinct mass margin present	28	18	32	8
indistinct mass margin not present	224	66	255	80
rs10759243	Carry A	Not Carry A	Carry A	Not Carry A
linear distribution of calcifications present	6	10	12	4
linear distribution of calcifications not present	164	156	163	196
rs17356907	Carry G	Not Carry G	Carry G	Not Carry G
lobular mass shape present	28	31	32	23
lobular mass shape not present	154	121	144	176

5. Discussion

The primary contribution of our study is to show that there exist, in our cohort, a number of interaction pairs between the genetic variants and mammographic features. These interaction pairs, if can be further validated in a larger cohort, are potentially useful for personalized breast cancer diagnosis. For example, when radiologists read mammograms for breast cancer diagnosis, they can also take into account the genetic variants of the patient if the information is available. If the interaction pairs are protective, successful adoption of them can help alleviate the problem of overdiagnosis. If the interaction pairs confer elevated breast cancer risk, successfully identifying them may increase the stratification power and allow for early detection of breast cancer. Our study differs from the previous study of Wacholder et al. (2010) [16] which added ten genetic variants to the Gail model, a risk model based on self-reported demographic and personal risk factors. Therefore, our study investigates the potential clinical impact of translating the exciting discoveries from GWAS to the patient experience at diagnosis. Unlike our previous study [4], which focused on the additional stratification power from these genetic variants in breast cancer risk prediction models, our current study focuses on the interaction effect between the genetic variants and mammographic features.

One methodological limitation of our study is that we only look into the two-way interaction between the genetic variants and the mammographic features. It is quite likely that the interaction comes from more than two risk variables. On the genomics side, it is likely that several genetic variants interact with each other and confer an elevated breast cancer risk. On the mammography side, radiologists usually make medical diagnosis and decisions based on a combination of features rather than a single one. However, detecting high-order interaction effect requires more samples.

We are also aware of other methods for identifying the interaction pairs such as hypothesis testing and Bonferroni correction. However, these tests are dependent on each other and the conservativeness of Bonferroni correction may reduce the power of detection. Therefore, we decided to use conditional mutual information as the measure and report the contingency tables as we elaborate these interaction pairs.

Limitations of our study include small sample size and the pitfalls of data extraction from text reports. We are aware of the limited power of detecting such interactions due to the limited number of samples. We understand that parsing mammography features from text reports may introduce noise into the data. Especially, we may have failed to extract some of the features from the text. Therefore, a number of the interaction pairs we identified may be false positives. We investigated the literature for evidence that can support these interactions, however almost all the new

SNPs were first identified by COGS in 2013 and there is no existing literature about them. Nevertheless, we believe our results will be useful to other researchers and warrant further investigation.

To sum up, our study connects mammographic features with germline genetic variants and explores the interaction effect between them. Mammography features represent richer phenotypic data directly relevant to breast cancer diagnosis and thus provide high signal. The germline variants contain the genetic information present in all the normal cells of the patient's body; this provides the genetic background from which abnormalities can arise, leading to the development of tumors. In order to fully investigate the susceptibility of genetic variants that might lead to mutations that develop into tumors, the DNA from tumor cells (to identify somatic mutations) should also be studied using emerging single-cell technologies. Analysis of germline variants and somatic mutations in individual patients and combining such data from cohort studies can help to identify germline predispositions and environmental effects related to cancer, which can in turn lead to more informed diagnosis and treatment. We hope that our work can move forward and eventually bring radiogenomic imaging into breast care, understanding the correlation between gene expression profiling of solid tumors and noninvasive cancer imaging features to provide new insights into human cancers. In a future study, we plan to utilize additional genomic and transcriptomic data in the hopes of linking specific radiological tumor phenotypes from routine clinical imaging to treatment-response gene expression patterns in order to predict the likely response to specific chemotherapeutics.

Acknowledgements

The authors acknowledge the support of the Wisconsin Genomics Initiative from the state of Wisconsin and support from the National Institutes of Health (grants: R01CA127379, R01CA127379-03S1, R01GM097618, R01LM011028, R01ES017400). We also acknowledge support from the eMERGE Network (U01HG004608), the UW Institute for Clinical and Translational Research (UL1TR000427) and the UW Carbone Comprehensive Cancer Center (P30CA014520).

Reference

1. Maxwell, K.N. and K.L. Nathanson, *Common breast cancer risk variants in the post-COGS era: a comprehensive review.* Breast Cancer Res, 2013. **15**(6): p. 212.
2. Green, E.D. and M.S. Guyer, *Charting a course for genomic medicine from base pairs to bedside.* Nature, 2011. **470**(7333): p. 204-13.
3. McCarty, C., et al., *Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank.* Personalized Med, 2005. **2**: p. 49-79.
4. Liu, J., et al. *Genetic Variants Improve Breast Cancer Risk Prediction on Mammograms*. in *American Medical Informatics Association Symposium*. 2013.
5. Michailidou, K., et al., *Large-scale genotyping identifies 41 new loci associated with breast cancer risk.* Nat Genet, 2013. **45**(4): p. 353-61, 361e1-2.
6. Warren, H., et al., *9q31.2-rs865686 as a susceptibility locus for estrogen receptor-positive breast cancer: evidence from the Breast Cancer Association Consortium.* Cancer Epidemiol Biomarkers Prev, 2012. **21**(10): p. 1783-91.
7. Stevens, K.N., et al., *19p13.1 is a triple-negative-specific breast cancer susceptibility locus.* Cancer Res, 2012. **72**(7): p. 1795-803.
8. Siddiq, A., et al., *A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11.* Hum Mol Genet, 2012. **21**(24): p. 5373-84.
9. Ghoussaini, M., et al., *Genome-wide association analysis identifies three new breast cancer susceptibility loci.* Nat Genet, 2012. **44**(3): p. 312-8.
10. Fletcher, O., et al., *Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study.* J Natl Cancer Inst, 2011. **103**(5): p. 425-35.
11. Turnbull, C., et al., *Genome-wide association study identifies five new breast cancer susceptibility loci.* Nat Genet, 2010. **42**(6): p. 504-7.
12. Antoniou, A.C., et al., *A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population.* Nat Genet, 2010. **42**(10): p. 885-92.
13. Marmot, M., et al., *The benefits and harms of breast cancer screening: an independent review.* British Journal of Cancer, 2013. **108**(11): p. 2205--2240.
14. Houssam, N., et al., *Information Extraction for Clinical Data Mining: A Mammography Case Study*, in *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*. 2009, IEEE Computer Society.
15. Percha, B., et al., *Automatic classification of mammography reports by BI-RADS breast tissue composition class.* J Am Med Inform Assoc, 2012. **19**(5): p. 913-6.
16. Wacholder, S., et al., *Performance of common genetic variants in breast-cancer risk models.* N Engl J Med, 2010. **362**(11): p. 986-993.