Published in final edited form as:

Genet Epidemiol. 2011 July ; 35(5): 389–397. doi:10.1002/gepi.20587.

Assessment of Rare BRCA1 and BRCA2 Variants of Unknown Significance Using Hierarchical Modeling

Marinela Capanu¹, Patrick Concannon², Robert W. Haile³, Leslie Bernstein⁴, Kathleen E. Malone⁵, Charles F. Lynch⁶, Xiaolin Liang¹, Sharon N. Teraoka², Anh T. Diep³, Duncan C. Thomas³, Jonine L. Bernstein¹, The WECARE Study Collaborative Group, and Colin B. Begg^{1,*}

¹Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY

²Center for Public Health Genomics and Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA

³Department of Preventive Medicine, University of Southern California, Los Angeles, CA

⁴Department of Population Sciences, Division of Cancer Etiology, City of Hope Comprehensive Cancer Center, Duarte, CA

⁵Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA

⁶Department of Epidemiology, University of Iowa, Iowa City, IA

Abstract

Current evidence suggests that the genetic risk of breast cancer may be caused primarily by rare variants. However, while classification of protein-truncating mutations as deleterious is relatively straightforward, distinguishing as deleterious or neutral the large number of rare missense variants is a difficult on-going task. In this article we present one approach to this problem, hierarchical statistical modeling of data observed in a case-control study of contralateral breast cancer in which all participants were genotyped for variants in *BRCA1* and *BRCA2*. Hierarchical modeling permits leverage of information from observed correlations of characteristics of groups of variants with case-control status to infer with greater precision the risks of individual rare variants. A total of 181 distinct rare missense variants were identified among the 705 cases with contralateral breast cancer and the 1398 controls with unilateral breast cancer. The model identified 3 bioinformatic hierarchical covariates, align-GV, align-GD and SIFT scores, each of which was modestly associated with risk. Collectively, the 11 variants that were classified as adverse on the basis of all three bioinformatic predictors demonstrated a stronger risk signal. This group included 5 of 6 missense variants that were classified as deleterious at the outset by conventional criteria. The remaining 6 variants can be considered as plausibly deleterious, and deserving of further investigation (*BRCA1* R866C; *BRCA2* G1529R, D2665G, W2626C, E2663V and R3052W). Hierarchical modeling is a strategy that has promise for interpreting the evidence from future association studies that involve sequencing of known or suspected cancer genes.

Keywords

Rare variants; cancer risk; logistic regression; pseudo-likelihood, Bayesian

*Corresponding Author: Colin B. Begg, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 307 East 63rd Street, 3rd floor, New York, NY 10065, Tel: (646) 735-8108, Fax: (646) 735-0009, beggc@mskcc.org.

INTRODUCTION

Women with germline variants in *BRCA1* or *BRCA2* that have been classified as deleterious are known to be at high risk of breast cancer. Evidence about individual variants has accumulated over the past two decades from information gathered primarily from individual families with clusters of breast cancers who share the same variant. Thousands of variants have been identified in this way and detailed information is available from the Breast Cancer Information Core (<http://research.nhgri.nih.gov/bic>). Many population-based case-control studies of breast cancer have also been conducted in which specimens have been subjected to mutation testing. However, these studies have been used primarily to provide global estimates of the relative risk and absolute risk (penetrance) among carriers in aggregate, rather than as a source for estimating the distinctive risks conferred by individual variants.

There are two reasons why case-control studies have not been used commonly as a data source for assessing the risks conferred by individual variants. First, mutation carriers are very infrequent in the population at large, and so control groups comprising individuals unaffected with breast cancer will typically comprise very few subjects who are carriers. Second, the actual mutations that confer risk are rare: that is, the common polymorphic variants in *BRCA1* and *BRCA2* appear to have little impact on risk. Thus, even among variants that are observed, an individual rare variant will typically occur only once or twice in a given study, providing no meaningful direct evidence about risk by itself. However, two recent developments have encouraged us to pursue the goal of estimating the influence of individual rare variants on disease risk from case-control data. First, it has been recognized that risk factors can be evaluated in novel case-control designs whereby the “cases” are individuals who experience two independent primaries, and the “controls” are individuals with a single primary [Begg and Berwick, 1997; Fletcher et al., 2009; Imyanitov et al., 2007; Kuligina et al., 2010]. The participants in these designs are greatly enriched for risk variants, and this enrichment is especially valuable for rare variants with substantial relative risks. The WECARE Study, from which the data in the present article are derived, is an important prototype of this kind of investigation, using women with contralateral breast cancer as cases and women with unilateral breast cancer as controls [Bernstein et al., 2006]. The second development is the adaptation of hierarchical statistical modeling to the evaluation of genetic risk variants [Aragaki et al., 1997; Capanu et al., 2008; Conti and Witte, 2003; DeRoos et al., 2003; Hung et al., 2004; Hung et al. 2007; Lewinger et al., 2007]. This technique allows the modeled aggregation of variants that share higher level molecular characteristics, such as those captured by modern bioinformatic tools. If these bioinformatic characteristics are observed to be strongly associated with risk, then variants that share the high risk characteristic can be inferred to share the high risks. In earlier work we have adapted hierarchical modeling for use specifically in the setting of rare variants [Capanu et al., 2008], and have further studied the statistical properties of the method [Capanu and Begg, 2010].

There has been some recent attention to the methodology for evaluating rare variants in the setting of a case-control study. Li and Leal [2008] advocate the informal use of information on functional status of rare variants at a locus of interest to select a candidate group of deleterious rare variants, and then to construct an omnibus test that the locus is associated with disease. Our approach differs from this in two important ways. First, we use the information pertaining to functionality formally in the model, as represented by hierarchical covariates. Second, our approach is targeted at determining which variants confer risk at a locus, rather than testing the locus overall. Li et al. [2010] have also studied this issue, comparing various proposed statistical tests for association of an individual rare variant with disease. However, their simulations deal with minor allele frequencies that are likely to result in the occurrence of several participants with the variant in the study, and these minor

allele frequencies (MAFs) are notably higher than apply to the preponderance of the *BRCA1* and *BRCA2* variants that we evaluate in our study. Consequently, the tests proposed by Li et al. are not sensitive enough for our setting. Morris and Zeggini [2010] have examined two strategies for testing the significance of rare variants collectively at a locus, one based on examining the association of the number of variant alleles with the phenotype, and the other examining the presence or absence of any allele with the phenotype. Again, our goal is to identify which specific rare variants are associated with the phenotype, rather than establishing the overall significance of the locus.

In this article we present a hierarchical model analysis of all missense variants in *BRCA1* and *BRCA2* observed in the WECARE Study. The vast majority of these are variants of unknown significance (VUS), and indeed several were observed for the first time in this study (Borg et al., 2010). VUS present problems for genetic counselors in that they may indicate high breast cancer risk but we have no definitive evidence to classify them as deleterious or neutral. We show that hierarchical modeling, adapted to the setting of rare variants, has the potential to shed light on which VUS are plausibly deleterious, and deserving of further laboratory investigation regarding their functional status.

METHODS

The WECARE Study population has been reported in several previous articles [Begg et al., 2008; Bernstein et al., 2004; Borg et al., 2010; Malone et al., 2010]. Briefly the study included 705 women with contralateral breast cancer (CBC) and 1398 women with unilateral breast cancer (UBC) who were individually matched to the CBCs, all of whom were ascertained through five population-based cancer registries (Iowa, Los Angeles, Orange County CA, Seattle WA and Denmark). Epidemiologic risk factor information for all participants was ascertained through a structured questionnaire. The study protocol was approved by the institutional review boards at each center. All participants were diagnosed with their first primary breast cancer at age 54 years or younger and the mean interval from first to second breast cancer in the CBCs was 5 years (range 1–16 years). Controls, women with UBC, were matched individually in a ratio of 2:1 to cases on date of birth, date of diagnosis, race and reporting registry.

The complete coding sequences of *BRCA1* (5,589 bp split into 22 coding exons) and *BRCA2* (10,254 bp and 26 coding exons) were screened for variations by denaturing high-performance liquid chromatography (DHPLC), using leukocyte genomic DNA as template. Three laboratories (two in the US, one in Sweden) performed all of the screening using a fixed set of primers and DHPLC protocols. Further details of the screening protocol are available in previous publications [Begg et al., 2008; Borg et al., 2010; Malone et al., 2010]. In order to compare results of this study with previous publications and mutation databases, all sequence variants were named and are referred to in the text according to the nomenclature used by the BIC (<http://research.nhgri.nih.gov/bic>), with nucleotide numbering starting at the first transcribed base of *BRCA1* (GenBank U14680.1) and *BRCA2* (NM_000059.1).

Sequence variants were categorized based on their predicted effect on the mRNA and amino acid level and defined as deleterious mutations according to the following established (BIC) criteria: all frameshift and nonsense variants with the exception of the neutral stop codon *BRCA2* K3326X [Mazoyer et al., 1996] and other variants located 3' thereof; all intronic/non-coding intervening sequence (IVS) variants occurring in the consensus splice acceptor or donor sequence sites, either within 2 bp of exon-intron junctions or when experimentally demonstrated to result in abnormal mRNA transcript processing; missense variants that have been conclusively demonstrated, on the basis of data from linkage analysis of high risk

families, functional assays or biochemical evidence, to have a deleterious effect on known functional regions.

We employed a hierarchical statistical model that has been adapted for use specifically in the setting of rare variants, that is where each random effect represents an individual missense variant and where the random effects may be observed very infrequently in the study, in many cases only once [Capanu et al., 2008]. Although conventional logistic regression will inevitably fail to converge if any of the variants does not occur in at least one case and one control the hierarchical modeling approach circumvents this problem using pseudo-likelihood estimation of the parameters. The first stage of the hierarchical model uses a mixed model logistic regression to relate the dependent variable (case versus control status, i.e. CBC versus UBC) to the individual rare missense variants ($MAF \leq 2.5\%$); the individual variants are included as random effects, while known patient-specific risk factors are included as fixed effects. To ensure that our model identified the distinct effects of the individual variants we also included common polymorphisms ($MAF > 2.5\%$) in *BRCA1* and *BRCA2* as fixed effects, and terms to adjust for the presence of a mutation other than a missense mutation (i.e. splicing, truncation, frameshift, nonsense). The second stage of the hierarchical model involves a linear regression of the log relative risks of the missense variants on bioinformatic characteristics of the variants, in addition to a term that distinguished *BRCA1* from *BRCA2* variants. This approach is designed to combine the data on CBC and UBC frequencies of each specific variant with bioinformatics and other characteristics of the variants to infer the relative risks of the individual variants. This is accomplished through the mediating effects of the second level covariates, primarily the bioinformatics predictors.

The parameters of the model were estimated using a method that involves pseudo likelihood estimation of the relative risk parameters with Bayesian estimation of the variance components, a strategy that has been shown to have the best statistical properties among competitor estimation techniques studied in this setting [Capanu and Begg, 2010]. This approach approximates the mixed logistic regression model with a linear mixed model using the SAS %glimmix macro. The SAS Proc Mixed with the Prior option is employed to estimate the variance of the random effects using a random walk Metropolis-Hastings algorithm with a non-informative Inverse-Gamma distribution. Once the Bayesian estimate of the variance is obtained, the %glimmix macro is employed again to estimate the regression parameters using the pseudo-likelihood method but with the random effects' variance pre-specified at the Bayesian estimate. For purposes of benchmarking, we also employed ordinary logistic regression including in the model all variants that were observed in at least one case and one control and adjusting for the same confounders as the first stage model of the hierarchical model. We note that two variants (V2138F and F486L) co-occurred with other variants and were included on their own in a logistic regression model adjusted for the same confounders as described above. We also compared these results with estimates obtained from a hierarchical model in which the second-stage model included no covariates besides an intercept (i.e. assuming exchangeability of the random effects).

As a benchmarking tool to characterize the distribution of relative risks we would expect to observe if all of the rare variants have no effect on risk we built upon a simulation strategy employed previously in Capanu and Begg (2010). We generated datasets that mirrored the observed dataset (i.e. number of participants, number of variants and their carriers were kept fixed), while each time generating the case-control status based on the probability that a subject with a given configuration of genetic variants and a specific set of confounders is a case. The hierarchical model at the basis of these simulations involved relative risks for all first and second stage covariates as estimated from the data, but with all rare missense variants assumed to have no effect on risk, i.e. true relative risks of 1.0.

We used three bioinformatic tools as potential second stage covariates: Align-GVGD, Polyphen and SIFT. Align-GVGD (<http://agvgd.iarc.fr/>) is a program that combines the biophysical characteristics (side-chain composition, polarity and volume) of amino acids and protein multiple sequence alignments (Grantham Variation and Grantham Deviation scores) to predict where amino acid substitutions fall in a spectrum from enriched deleterious to enriched neutral, based on GV and GD scores (0 to >200) [Tavtigian et al., 2006]. Alignments were done with up to 13 *BRCA1* sequences and 12 *BRCA2* sequences to the depth of *Xenopus laevis* (frog), *Tetraodon nigroviridis* (pufferfish - green spotted) or *Strongylocentrotus purpuratus* (purple sea urchin). PolyPhen (Polymorphism Phenotyping) (<http://genetics.bwh.harvard.edu/pph>) predicts the possible impact of an amino acid substitution based on empirical rules applied to the sequence, phylogenetic and structural information characterizing the substitution. SIFT (Sorting Intolerant from Tolerant) (<http://blocks.fhcrc.org/sift/SIFT.html>) predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. Normalized probabilities of substitutions are calculated under default settings and probabilities ≤ 0.05 are predicted to be deleterious [Tong et al., 2004].

RESULTS

In previous reports from this study we have shown that overall the presence of a deleterious variant (as defined by BIC criteria) demonstrated an approximately four fold increase in the risk of CBC [Malone et al., 2010] and that in aggregate the unclassified VUS appear to have a neutral effect on risk, regardless of their MAFs [Borg et al., 2010]. However, 181 distinct rare missense variants ($\leq 2.5\%$ frequency) were observed in the study (including the 6 previously classified as deleterious by BIC criteria), and so it is possible that a few additional variants may truly be deleterious. [We note that in previous publications we classified *BRCA2* M1I as a missense variant, but on reflection we have elected to remove it from this classification.] The purpose of our hierarchical modeling was to see if evidence exists in the data to identify additional plausibly deleterious missense variants.

By definition, the 181 rare variants observed occur infrequently in the dataset and so any additional evidence that can be used to sort them in terms of risk must be contained in the hierarchical covariates that characterize their plausible functionality. The breakdown of cases and controls with respect to these covariates is displayed in Table I. For align-GV and align-GD we set the criteria as $GV=0$ and $GD \geq 65$ respectively, following the classification at <http://agvgd.iarc.fr/classifiers.php>. For SIFT we used the recommended ≤ 0.05 as the criterion, while for PolyPhen variants were classified as benign, possibly damaging, or probably damaging, under default settings. The table indicates that both align-GV and align-GD are significantly associated with risk, SIFT is borderline significant, but that PolyPhen has no apparent association. Examination of the aggregated GVGD scale proposed by Tavtigian et al. (2006) indicates that essentially all of the excess risk is focused on the highest risk category (C65). Based on these results we included these 3 covariates in the hierarchical model, along with various other first and second stage covariates.

The adjusted relative risks of all covariates from the model are listed in Table II. These include the known hormonal risk factors for breast cancer (age at menarche, childbearing history and menopausal status) used in the first, subject-specific stage of the hierarchical model. The first stage of the model also included fixed effects parameters for the 24 common polymorphic variants ($MAF > 2.5\%$), and their relative risks are shown in Table III. Only one of the 18 SNPs has a significantly elevated risk (D693N), but the increase in risk is modest, and we expect one false positive when studying about 20 independent comparisons. One of the IVS polymorphic variants is also significant. Table II also shows the variables used in the second, variant-specific stage of the model. In addition to the

bioinformatic predictors, family history of breast cancer was accommodated as a hierarchical covariate using a continuous standardized ratio of the frequency of breast cancer in first degree female relatives of all the probands with the specific variant relative to the total number of female relatives for these probands. This covariate did not seem to have a significant effect, though others have noted that family history is associated with a subset of missense variants [Scott et al., 2003]. The average age at diagnosis of the carriers of a specific variant was also included as a hierarchical covariate dichotomized at age 45 (<45 vs. ≥45), and showed a slightly elevated risk among carriers younger than 45 years at diagnosis, though this is not statistically significant.

The relative risk estimates from the hierarchical model along with confidence intervals and case control frequencies for the entire set of 181 rare missense variants are contained in the Supplementary Table. These relative risks are plotted in the red histogram in Figure I. The blue histogram displays the expected distribution if all of the 181 rare variants are unrelated to risk. The Figure shows that there is a small upper tail to the distribution that is consistent with the hypothesis that there is a small subset of harmful variants. The individual relative risk estimates of the 14 variants whose estimates exceeded 2.0 are listed in Table IV. These variants are evenly distributed between *BRCA1* and *BRCA2*. We see that 5 of the 6 missense variants classified at the outset as deleterious by BIC criteria are in this high risk category (denoted with a superscript of “1”). The 6th “deleterious” variant, *BRCA1* M1775R, was observed in a single control participant and the modeled relative risk estimate is 1.6 (95% CI 0.5–5.5). Of the 14 variants in Table IV, the top 6 achieve a nominal statistical significance at the 5% level. The features of the data that are driving these results are the bioinformatic predictors listed in the right-hand columns of the table. Average age at onset and family history of breast cancer have very modest impact but are included in the table for completeness. We see that the top 11 variants are all characterized by the simultaneous presence of adverse bioinformatics classifications for all 3 bioinformatic predictors. Indeed these are the sole 11 variants in the data with these characteristics. Among these, the model assigns a somewhat higher relative risk estimate if the variant was observed in cases rather than controls. Somewhat lower relative risk estimates are assigned to the variants near the bottom of the list that are seen to possess two of the three adverse bioinformatic classifications. In essence, the model is observing a strong relationship between this combination of characteristics and risk, and is inferring that any rare variant that shares this combination of characteristics is likely to share the high risk.

Relative risk estimates for individual variants based solely on a conventional, non-hierarchical logistic regression model are shown in the 3rd column of Table IV. Estimates are available only for the 2 variants observed in at least one case and one control, and we see that the confidence intervals for these are much wider than for the corresponding hierarchical estimates. In the Supplementary Table we see that in general the logistic regression estimates are much more variable, with correspondingly larger confidence intervals, while the hierarchical estimates are shrunk substantially into a much narrower range. The Supplementary Table also reports the relative risk estimates from an “exchangeable” hierarchical model with no bioinformatic covariates. These estimates exhibit the same shrinkage behavior as noted above and show that the information in the bioinformatic covariates contributes strongly to the hierarchical model estimates.

DISCUSSION

The search for genetic risk factors for cancer and countless other diseases has been a major focus of investigation in recent years. Many genome-wide association studies have been conducted, including several in breast cancer [Easton et al., 2007; Gold et al., 2008; Hunter et al., 2007; Ruiz-Narvaez et al., 2010; Thomas et al., 2009]. These studies, which involve

the use of relatively common SNPs, have been based on the premise that the major influences on the risk of common diseases such as cancer are likely to involve relatively common variants, on the grounds that only common variants can deliver high attributable risks. Although there has been some success in identifying cancer genes through this process, in general the relative risks conferred by the identified SNPs have been small. This is leading to a reconsideration of the premise that genetic risk of cancer is caused primarily by common variants [Gorlov et al., 2008; Shork et al., 2009]. Indeed the *BRCA1* and *BRCA2* genes are good examples of why this hypothesis may be false, in that selected rare variants in these genes confer very high risk, while the common variants appear to convey little or no risk. Even more perplexing is the fact that among the thousands of observed rare variants only some appear to confer risk, while others, including the vast majority of single nucleotide changes, appear to be harmless. This would seem to be a paradigm that is likely to apply to other cancer genes that are identified in the future. Faced with this possibility, the challenge to find strategies that distinguish risk bearing from harmless variants in a setting where the information on each individual variant is necessarily sparse becomes increasingly important. Consequently resolution of these ambiguities is an important public health issue.

Our approach has been to use sequence variant data from a case-control study to accomplish the evaluation of individual rare variants in a highly efficient manner. In our study the cases are women with double breast malignancies while the controls have unilateral breast cancer. This ensures that the subject pool is as gene rich as possible, thereby optimizing the number of variants observed in the study. We then used hierarchical statistical modeling to leverage the information in the data as efficiently as possible. This strategy assesses the observed case-control ratios of the individual variants, while simultaneously evaluating the broader case-control ratios of groups of variants, grouped on the basis of pre-defined categories defined by bioinformatic or other classifications of the variants. The study produced an interesting finding. It showed that among the rare missense VUS, which collectively demonstrated very little increased risk [Borg et al., 2010], there is a small subset of variants characterized by adverse predictions on all three bioinformatic classifications that has a strong collective association with breast cancer risk. The model then instructs us to conclude that membership of a variant in this group indicates that the variant is likely to possess high risk, even though there is no direct evidence, on the basis of the frequent occurrence of the variant itself in cases, that the variant has high risk.

How do these results compare with previous modeled strategies to assign risk to individual rare variants? In a comprehensive analysis that involved combining data from multiple sources Easton et al. [2007] examined all of the *BRCA1* and *BRCA2* variants that have been observed to date. The authors' method is not directly comparable with ours, and the data sets are non-overlapping. Nonetheless the results show some overlap and some disagreements on the classification of the variants. Of special interest are the 11 variants that we identify as being high risk variants on the basis of possessing adverse predictions on all three bioinformatic classifications (Table III). Of these 11 variants, 6 were considered to be VUS at the outset of our study (the other 5 are deleterious by BIC criteria). Three of these (*BRCA1* R866C, *BRCA2* G1529R and D2665G) were classified as neutral by Easton et al., while two are in agreement (W2626C and E2663V). Interestingly, the sixth variant, R3052W, was confirmed as functional in a recent study using a mouse embryonic stem cell assay [Kuznetsov et al., 2008]. Tavtigian et al. (2008) have also addressed this issue using data from the Myriad Genetics Laboratory database, and a variety of approaches for mapping risk to the GV and GD indices. However, none of these methods involve direct comparison of frequencies of observed cases and controls, either for individual variants or for groups defined by GVGD criteria. Their results indicate a monotonic increase in risk as a function of their summary score, ranging from C0 to C65, but our results seem to suggest,

albeit with small sample sizes, that the elevated risk may be limited to variants in the highest (C65) category (see Table I).

Of the 181 rare missense variants that we examined in the model, 6 emerged as having nominally statistically significant association with risk (the top 6 variants listed in Table III). Conventionally we expect 1 statistically significant result for every 20 independent tests performed, and so at first glance the identification of 6 significant results from the 181 variants seems unremarkable. However, the use of conventional methods to adjust for multiple comparisons is problematic in this setting. For one thing, we knew at the outset that the case-control ratio overall for subjects with rare variants was similar to that of the comparison group of subjects with no rare variants [Borg et al., 2010], and so the vast majority of the variants had to have neutral risk. Indeed the purpose of the exercise was to see if any signals emerged from this overall pattern of null association. Second, each individual variant occurs in very few subjects. Indeed the majority of these rare variants occurred only once. So there is a fundamental problem of low statistical power per variant, and the fact that any of these variants emerged with “significant” statistical evidence is itself remarkable. Third, the fact that estimates and confidence intervals for the individual variants are essentially derived by grouping them on the basis of common bioinformatic categories means that these estimates and tests are highly correlated, and thus are not amenable to the assumptions of conventional multiple testing algorithms such as the Bonferroni correction [Bonferroni, 1935] and false discovery rate techniques [Benjamini and Hochberg, 1995]. Nonetheless the issue of multiple comparisons is highly relevant in this context, and further research is needed to interpret the results of such analyses of large numbers of individual rare variants. Finally, the statistical properties of the hierarchical model itself are sub-optimal in the context of studying rare variants. This is an inevitable consequence of using asymptotic statistical methods when the data are sparse. However, a detailed study of this issue has provided evidence that the relative risk estimates have low bias, and the confidence intervals have coverage that exceeds 90% for a nominal 95% interval for most data configurations [Capanu and Begg, 2010].

As a cautionary note it would seem inappropriate to over-interpret the distinct results for individual variants. Clearly the power of this method derives from the grouping of variants into classes with high collective risk. Any variant that belongs to such a group is inevitably allocated a high risk by the model. Consequently the results for some individual variants will be highly dependent on how we choose to group the data. We elected to use the align-GVGD and SIFT scores with pre-specified cut-offs for this purpose. If we were to use different cut-offs, or if we focused instead on grouping variants on the basis of, say, known functional domains in *BRCA1* and *BRCA2* then the allocation of some variants to the high risk category could change. We did explore the empirical association of 5 functional domains with risk of CBC: *BRCA1* Ring; *BRCA1* BRCT; *BRCA1* NLS, *BRCA2* DNA Binding; *BRCA2* Transactivation. Of these, only the *BRCA1* Ring domain exhibited a significant association with risk. Of the 3 observed variants in this domain, 2 (C44S and C61G) were classified as “deleterious” at the outset (see Table IV), and the third (K45T) was observed in one control subject. This variant is neutral for all 3 bioinformatic predictors.

In summary, we conducted a hierarchical modeling analysis of rare VUS in *BRCA1* and *BRCA2* using data from the WECARE Study. The results show that the vast majority of rare missense variants are neutral, but the study supports the growing evidence that there exists a small group of variants that are deleterious. This group includes a few variants classified as deleterious by conventional (BIC) criteria, but it also includes a few additional VUS. These additional variants are worthy of further investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The research was supported by the National Cancer Institute, award CA131010, CA097397 and CA083178. The WECARE Study Collaborative Group involves the following participants: Memorial Sloan Kettering Cancer Center (New York, NY): Jonine L. Bernstein Ph.D. (WECARE Study P.I.), Colin B. Begg, Ph.D., Marinela Capanu Ph.D., Xiaolin Liang M.D., Anne S. Reiner M.P.H., Tracy M. Layne M.P.H.; City of Hope (Duarte, CA) (some work performed at University of Southern California, Los Angeles CA): Leslie Bernstein Ph.D., Laura Donnelly-Allen; Danish Cancer Society (Copenhagen, Denmark): Jørgen H. Olsen M.D., D.M.Sc., Michael Andersson M.D., D.M.Sc., Lisbeth Bertelsen M.D.Ph.D., Per Guldberg Ph.D., Lene Mellemkjær Ph.D.; Fred Hutchinson Cancer Research Center (Seattle, WA): Kathleen E. Malone Ph.D.; International Epidemiology Institute (Rockville, MD) and Vanderbilt University (Nashville, TN): John D. Boice Jr. Sc.D.; Lund University (Lund, Sweden): Åke Borg Ph.D., Therese Törngren M.Sc., Lina Tellhed, B.Sc.; National Cancer Institute (Bethesda, MD): Daniela Seminara Ph.D. M.P.H.; New York University (New York, NY): Roy E. Shore Ph.D., Dr.PH.; Norwegian Radium Hospital (Oslo, Norway): Laila Jansen, Anne-Lise Børresen-Dale Ph.D. (also University of Oslo, Norway); University of California at Irvine (Irvine, CA): Hoda Anton-Culver, Ph.D., Joan Largent Ph.D. M.P.H.; University of Iowa (Iowa City, IA): Charles F. Lynch M.D., Ph.D., Jeanne DeWall M.A.; University of Southern California (Los Angeles, CA): Robert W. Haile DrPH., Graham Casey, Ph.D., Bryan Langholz Ph.D., Duncan C. Thomas Ph.D., Shanyan Xue M.D., Nianmin Zhou, M.D., Anh T. Diep, Evgenia Ter-Karapetova; University of Southern Maine (Portland, ME): W. Douglas Thompson Ph.D.; University of Texas, M.D. Anderson Cancer Center (Houston, TX): Marilyn Stovall Ph.D., Susan Smith M.P.H.; University of Virginia (Charlottesville, VA) (some work performed at Benaroya Research Institute, Seattle WA): Patrick Concannon, Ph.D., Sharon N. Teraoka, Ph.D., Eric R. Olson, Ph.D., Nirasha Ramchuren, Ph.D.

REFERENCES

- Aragaki CC, Greenland S, Probst-Hensch N, Haile RW. Hierarchical modeling of gene-environment interactions: estimating NAT2 genotype-specific dietary effects on adenomatous polyps. *Cancer Epidemiol Biomarkers Prev.* 1997; 6(5):307–14. [PubMed: 9149889]
- Begg CB, Berwick M. A note on the estimation of relative risks of rare genetic susceptibility markers. *Cancer Epidemiol Biomarkers Prev.* 1997; 6(2):99–103. [PubMed: 9037560]
- Begg CB, Haile RW, Borg A, Malone KE, Concannon P, Thomas DC, Langholz B, Bernstein L, Olsen JH, Lynch CF, et al. Variation of breast cancer risk among BRCA1/2 carriers. *JAMA.* 2008; 299(2): 194–201. [PubMed: 18182601]
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc B.* 1995; 57(1):289–300.
- Bernstein JL, Langholz B, Haile RW, Bernstein L, Thomas DC, Stovall M, Malone KE, Lynch CF, Olsen JH, Anton-Culver H, et al. Study design: evaluating gene-environment interactions in the etiology of breast cancer - the WECARE study. *Breast Cancer Res.* 2004; 6(3):R199–214. [PubMed: 15084244]
- Bonferroni, CE. Studi in Onore del Professore Salvatore Ortu Carboni. Rome, Italy: 1935. Il calcolo delle assicurazioni su gruppi di teste; p. 13-60.
- Borg A, Haile RW, Malone KE, Capanu M, Diep A, Törngren T, Teraoka S, Begg CB, Thomas DC, Concannon P, et al. Characterization of BRCA1 and BRCA2 deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the WECARE study. *Hum Mutat.* 2010; 31(3):E1200–40. [PubMed: 20104584]
- Capanu M, Begg CB. Hierarchical model for estimating relative risks of rare genetic variants: properties of the pseudo-likelihood method. *Biometrics.* 2010 In press.
- Capanu M, Orlow I, Berwick M, Hummer AJ, Thomas DC, Begg CB. The use of hierarchical models for estimating relative risks of individual genetic variants: an application to a study of melanoma. *Stat Med.* 2008; 27(11):1973–92. [PubMed: 18335566]
- Conti DV, Witte JS. Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. *Am J Hum Genet.* 2003; 72(2):351–63. [PubMed: 12525994]

- De Roos AJ, Rothman N, Inskip PD, Linet MS, Shapiro WR, Selker RG, Fine HA, Black PM, Pittman GS, Bell DA. Genetic polymorphisms in GSTM1, -P1, -T1, and CYP2E1 and the risk of adult brain tumors. *Cancer Epidemiol Biomarkers Prev.* 2003; 12(1):14–22. [PubMed: 12540498]
- Easton DF, Deffenbaugh AM, Pruss D, Frye C, Wenstrup RJ, Allen-Brady K, Tavtigian SV, Monteiro AN, Iversen ES, Couch FJ, et al. A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. *Am J Hum Genet.* 2007; 81(5):873–83. [PubMed: 17924331]
- Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struwing JP, Morrison J, Field H, Luben R, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature.* 2007; 447(7148):1087–93. [PubMed: 17529967]
- Fletcher O, Johnson N, Dos Santos Silva I, Kilpivaara O, Aittomaki K, Blomqvist C, Nevanlinna H, Wasielewski M, Meijers-Heijerboer H, Broeks A, et al. Family history, genetic testing, and clinical risk prediction: pooled analysis of CHEK2 1100delC in 1,828 bilateral breast cancers and 7,030 controls. *Cancer Epidemiol Biomarkers Prev.* 2009; 18(1):230–4. [PubMed: 19124502]
- Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, Garber J, Friedman E, Narod S, Olshen AB, Gregersen P, et al. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc Natl Acad Sci U S A.* 2008; 105(11):4340–5. [PubMed: 18326623]
- Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet.* 2008; 82(1):100–12. [PubMed: 18179889]
- Hung RJ, Brennan P, Malaveille C, Porru S, Donato F, Boffetta P, Witte JS. Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. *Cancer Epidemiol Biomarkers Prev.* 2004; 13(6):1013–21. [PubMed: 15184258]
- Hung RJ, Thomas D, McKay J, Szeszenia-Dabrowska N, Zaridze D, Lissowska J, Rudnai P, Fabianova E, Mates D, Foretova L, Janout V, et al. Inherited predisposition of lung cancer: a hierarchical modeling approach to DNA repair and cell cycle control pathways. *Cancer Epidemiol Biomarkers Prev.* 2007; 16(12):2736–44. [PubMed: 18086781]
- Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet.* 2007; 39(7):870–4. [PubMed: 17529973]
- Imyanitov EN, Cornelisse CJ, Devilee P. Searching for susceptibility alleles: emphasis on bilateral breast cancer. *Int J Cancer.* 2007; 121(4):921–3. [PubMed: 17444500]
- Kuligina E, Reiner A, Imyanitov EN, Begg CB. Evaluating cancer epidemiological risk factors using multiple primary malignancies. *Epidemiology.* 2010; 21:366–72. [PubMed: 20299982]
- Kuznetsov SG, Liu P, Sharan SK. Mouse embryonic stem cell-based functional assay to evaluate mutations in BRCA2. *Nat Med.* 2008; 14(8):875–81. [PubMed: 18607349]
- Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet Epidemiol.* 2007; 31(8):871–82. [PubMed: 17654612]
- Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet.* 2008; 83:311–21. [PubMed: 18691683]
- Li Q, Zhang H, Yu K. Approaches for evaluating rare polymorphisms in genetic association studies. *Hum Hered.* 2010; 69:219–28. [PubMed: 20332646]
- Malone KE, Begg CB, Haile RW, Borg A, Concannon PJ, Tellhed L, Xue S, Teraoka S, Bernstein L, Capanu M, et al. A population-based study of the relative and absolute risks of contralateral breast cancer associated with carrying a mutation in BRCA1 or BRCA2: results from the WECARE Study. *J Clin Oncol.* 2010; 28(14):2404–10. [PubMed: 20368571]
- Mazoyer S, Dunning AM, Serova O, Dearden J, Puget N, Healey CS, Gayther SA, Mangian J, Stratton MR, Lynch HT, Goldgar DE, Ponder BA, Lenoir GM. A polymorphic stop codon in BRCA2. *Nat Genet.* 2006; 14:253–4. [PubMed: 8896551]
- Morris AR, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol.* 2010; 34:188–93. [PubMed: 19810025]

- Ruiz-Narvaez EA, Rosenberg L, Rotimi CN, Cupples LA, Boggs DA, Adeyemo A, Cozier YC, Adams-Campbell LL, Palmer JR. Genetic variants on chromosome 5p12 are associated with risk of breast cancer in African American women: the Black Women's Health Study. *Breast Cancer Res Treat.* 2010; 123(2):525–30. [PubMed: 20140701]
- Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev.* 2009; 19(3):212–9. [PubMed: 19481926]
- Scott CL, Jenkins MA, Southey MC, Davis TA, Leary JA, Easton DF, Phillips KA, Hopper JL. Average age-specific cumulative risk of breast cancer according to type and site of germline mutations in BRCA1 and BRCA2 estimated from multiple-case breast cancer families attending Australian family cancer clinics. *Hum Genet.* 2003; 112:542–51. [PubMed: 12601471]
- Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet.* 2006; 43(4):295–305. [PubMed: 16014699]
- Tavtigian SV, Byrnes GB, Glodgar DE, Thomas A. Classification of rare missense substitutions using risk surfaces with genetic and molecular applications. *Hum Mutat.* 2008; 29(11):1342–1354. [PubMed: 18951461]
- Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K, et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet.* 2009; 41(5):579–84. [PubMed: 19330030]
- Tong XI, Jones IM, Mohrenweiser HW. Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. *Genomics.* 2004; 83:970–9. [PubMed: 15177551]

Figure I. Hierarchical modeling estimates for 181 rare missense variants (red histogram) versus expected distribution if no rare variants affect risk (blue histogram).

Table I

Frequencies and Relative Risks Observed for Higher Level Covariates

Risk Category¹	Controls	Cases	Relative Risk (95% CI)²
Referent	1275	577	
GV = 0	10	15	3.3 (1.4, 8.3)
Referent	1263	569	
GD ≥ 65	22	23	2.3 (1.2, 4.4)
Referent	1188	534	
SIFT ≤ 0.05	97	56	1.3 (0.9, 1.7)
Referent	1184	536	
GV = 0 or GD ≥ 65 or SIFT ≤ 0.05	101	56	1.2 (0.9, 1.7)
Referent	1277	577	
GV = 0 & GD ≥ 65 & SIFT ≤ 0.05	8	15	4.1 (1.6, 11.4)
Referent	1262	570	
C15–C55	15	7	1.0 (0.4, 2.7)
C65	8	15	4.2 (1.6, 11.4)
Referent	1092	498	
PolyPhen (possibly damaging)	75	40	1.2 (0.8, 1.8)
Polyphen (probably damaging)	118	54	1.0 (0.7, 1.4)

¹The referent categories exclude 96 controls and 107 cases with frameshift, nonsense or splice mutations.

²Unadjusted relative risks and confidence intervals.

Table II**Influence of Covariates in First and Second Stage Models**

	Risk Factor	Relative Risk (95% CI)
First Stage Model ¹	Age at menarche <12 vs. ≥ 12	1.2 (1.0 – 1.5)
	# Pregnancies 1–3 vs. 0	0.8 (0.7 – 1.1)
	# Pregnancies ≥ 4 vs. 0	0.6 (0.4 – 0.8)
	Age at menopause > 45 vs. ≤ 45	1.4 (1.1 – 1.7)
	Premenopausal vs. postmen. ≤ 45	0.9 (0.7 – 1.2)
	Truncation or splicing mutation ²	2.7 (2.0 – 3.6)
	Intercept	0.9 (0.7 – 1.3)
	Age at onset of carriers ³	1.1 (0.7 – 1.6)
Second Stage Model	Family history ⁴	1.0 (0.8 – 1.2)
	BRCA1 vs. BRCA2	0.9 (0.6 – 1.5)
	GV = 0	2.1 (0.6 – 7.0)
	GD ≥ 65	1.4 (0.5 – 3.8)
	SIFT ≤ 0.05	1.5 (0.8 – 2.5)

¹ The first stage model also included individual covariates for all common polymorphisms (a total of 24 with MAF > 2.5%, see Table III).

² Presence of a truncation (frameshift, nonsense) or splice mutation (total of 96 controls and 107 cases).

³ Average age at diagnosis for all carriers of a specific mutation dichotomized at age 45.

⁴ Relative frequency of family history of breast cancer among first degree relatives of carriers of the variant, standardized to account for differing family sizes.

Table III**Relative Risks of Common SNPs and Other Polymorphic Variants**

Gene	SNP	Relative Risk (95% CI)	Controls	Cases
BRCA1	Q356R	0.9 (0.7, 1.3)	170	81 (32%)
BRCA1	D693N	1.5 (1.1, 2.2)	190	106 (36%)
BRCA1	S694S	0.8 (0.4, 1.4)	703	337 (32%)
BRCA1	L771L	1.2 (0.4, 3.3)	666	321 (33%)
BRCA1	P871L	0.9 (0.5, 1.4)	696	333 (32%)
BRCA1	E1038G	1.5 (0.6, 3.7)	672	324 (33%)
BRCA1	K1183R	1.4 (0.7, 2.7)	670	328 (33%)
BRCA1	S1436S	0.5 (0.2, 1.2)	682	323 (32%)
BRCA1	S1613G	1.0 (0.5, 1.9)	721	348 (33%)
BRCA2	N289H	1.3 (0.2, 10.7)	97	51 (34%)
BRCA2	N372H	1.1 (0.9, 1.4)	662	360 (35%)
BRCA2	S455S	1.8 (0.3, 9.1)	95	51 (35%)
BRCA2	H743H	0.5 (0.1, 3.4)	95	49 (34%)
BRCA2	N991D	0.8 (0.2, 3.0)	97	51 (34%)
BRCA2	K1132K	1.3 (0.9, 1.7)	608	318 (34%)
BRCA2	V1269V	1.0 (0.8, 1.2)	439	212 (33%)
BRCA2	T1915M	0.8 (0.5, 1.2)	78	30 (28%)
BRCA2	S2414S	0.9 (0.7, 1.2)	498	245 (33%)
BRCA1	IVS7+36del14	0.9 (0.6, 1.6)	777	377 (33%)
BRCA1	IVS7-34C>T	0.9 (0.7, 1.2)	525	254 (33%)
BRCA1	IVS8-58delT	1.0 (0.5, 1.9)	702	342 (33%)
BRCA1	IVS17-53C>T	0.5 (0.3, 0.9)	83	31 (27%)
BRCA2	IVS16-14T>C	1.1 (0.9, 1.4)	731	391 (35%)
BRCA2	203G/A	1.0 (0.7, 1.3)	528	269 (34%)
All subjects			1398	705 (34%)

Table IV
Variants Exhibiting Relative Risk Estimates Greater Than 2.0

Gene	Variant	Relative Risk (95% CI)		Frequencies		FH ²	Age ³	Bioinformatic Predictors		
		Logistic Regression	Hierarchical Model	Controls	Cases			GV=0	GD ≥ 65	SIFT≤0.05
BRCA1	C61G [/]	12.4 (2.5, 62.1)	4.9 (1.8, 13.0)	2	8	4/35	38	✓	✓	✓
BRCA2	G1529R		4.2 (1.2, 14.4)	0	2	0/5	52	✓	✓	✓
BRCA2	W2626C		4.2 (1.2, 15.0)	0	1	1/2	32	✓	✓	✓
BRCA1	C44S [/]		4.1 (1.3, 13.6)	0	2	2/9	42	✓	✓	✓
BRCA1	A1708E [/]		3.9 (1.1, 13.2)	0	1	0/2	31	✓	✓	✓
BRCA2	D2665G	2.8 (0.2, 48.8)	3.6 (1.0, 12.6)	1	1	0/14	49	✓	✓	✓
BRCA1	R1699W [/]		3.4 (1.0, 11.7)	1	0	2/4	40	✓	✓	✓
BRCA2	R3052W		3.4 (1.0, 11.9)	1	0	0/4	50	✓	✓	✓
BRCA2	E2663V		3.3 (0.9, 11.7)	1	0	0/1	51	✓	✓	✓
BRCA1	G1738E [/]		3.2 (0.9, 11.1)	1	0	0/5	39	✓	✓	✓
BRCA1	R866C		3.1 (0.9, 10.8)	1	0	0/3	51	✓	✓	✓
BRCA2	T3349A		2.5 (0.5, 11.4)	1	0	1/7	54	✓		✓
BRCA2	K2950N		2.3 (0.7, 7.4)	0	3	1/10	53		✓	✓
BRCA1	I1391T		2.0 (0.6, 6.7)	0	1	1/2	38		✓	✓

[/] Denotes variants that were classified as deleterious by standard BIC criteria. A sixth missense deleterious variant, BRCA1 M1775R, was not included in this table since its relative risk estimate was less than 2. Its estimate was 1.6 (95% CI 0.5 – 5.5), and it was observed in a single control patient.

² Among 1st degree female relatives of all carriers of the variant the column FH represents the number of relatives who have experienced breast cancer (numerator) over the number of female relatives (denominator). Note that we used a standardized version of this proportion in the model to account for varying denominators (see footnote to Table II).

³ Average age at diagnosis among carriers of a specific variant.