frontiers | Frontiers in Genetics

TYPE Brief Research Report
PUBLISHED 30 September 2022
DOI 10.3389/fgene.2022.982930

Check for updates

OPEN ACCESS

EDITED BY
Valentina Silvestri,
Sapienza University of Rome, Italy

REVIEWED BY
Nguyen Quoc Khanh Le,
Taipei Medical University, Taiwan
Corinna Ernst,
University of Cologne, Germany

*CORRESPONDENCE
Borbala Mifsud,
bmifsud@hbku.edu.qa

SPECIALTY SECTION
This article was submitted to Cancer Genetics and Oncogenomics, a section of the journal Frontiers in Genetics

RECEIVED 30 June 2022
ACCEPTED 12 September 2022
PUBLISHED 30 September 2022

CITATION
Khandakji MN and Mifsud B (2022), Gene-specific machine learning model to predict the pathogenicity of BRCA2 variants.
Front. Genet. 13:982930.
doi: 10.3389/fgene.2022.982930

COPYRIGHT
© 2022 Khandakji and Mifsud. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Gene-specific machine learning model to predict the pathogenicity of *BRCA2* variants

Mohannad N. Khandakji[1,2] and Borbala Mifsud[1,3]*

[1]College of Health and Life Sciences, Hamad Bin Khalifa University, Ar-Rayyan, Qatar, [2]Hamad Medical Corporation, Doha, Qatar, [3]William Harvey Research Institute, Queen Mary University of London, London, United Kingdom

Background: Existing *BRCA2*-specific variant pathogenicity prediction algorithms focus on the prediction of the functional impact of a subtype of variants alone. General variant effect predictors are applicable to all subtypes, but are trained on putative benign and pathogenic variants and do not account for gene-specific information, such as hotspots of pathogenic variants. Local, gene-specific information have been shown to aid variant pathogenicity prediction; therefore, our aim was to develop a *BRCA2*-specific machine learning model to predict pathogenicity of all types of *BRCA2* variants.

Methods: We developed an XGBoost-based machine learning model to predict pathogenicity of *BRCA2* variants. The model utilizes general variant information such as position, frequency, and consequence for the canonical *BRCA2* transcript, as well as deleteriousness prediction scores from several tools. We trained the model on 80% of the expert reviewed variants by the Evidence-Based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) consortium and tested its performance on the remaining 20%, as well as on an independent set of variants of uncertain significance with experimentally determined functional scores.

Results: The novel gene-specific model predicted the pathogenicity of ENIGMA *BRCA2* variants with an accuracy of 99.9%. The model also performed excellently on predicting the functional consequence of the independent set of variants (accuracy was up to 91.3%).

Conclusion: This new, gene-specific model is an accurate method for interpreting the pathogenicity of variants in the *BRCA2* gene. It is a valuable addition for variant classification and can prioritize unreviewed variants for functional analysis or expert review.

KEYWORDS

breast cancer, variant pathogenicity, in-silico predictions, variant prioritization, VUS

Introduction

Breast cancer is the most common cancer in women, impacting more than two million each year (Bray et al., 2020; Sung et al., 2021). The disease affects one in seven women worldwide and causes the greatest number of cancer-related deaths among them (McGuire, Brown, Malone, McLaughlin, & Kerin, 2015; Sung et al., 2021). In 2020, it resulted in 684,996 deaths: equal to 15.5% of all cancer deaths among women (Ferlay et al., 2020; Sung et al., 2021). Early breast cancer detection with suitable treatment could reduce breast cancer death rates significantly in the long-term. If the cancer is located only in the breast, the 5-year survival rate of women with breast cancer is 99%, however, if the cancer has spread to a distant part of the body, the 5-year survival rate decreases to 27% (Noone et al., 2018). Therefore, to improve breast cancer outcomes and survival, early detection is crucial. Early detection involves two strategies: screening and early diagnosis. Nevertheless, the balance of potential benefits over risks for mammographic breast cancer screening of the general population is controversial (Canelo-Aybar et al., 2021). A Cochrane review published in 2013 found that it is unclear if mammographic screening does more good or harm (Gøtzsche & KJ, 2013). Recent studies suggest that mammographic screening could be most effective if offered based on the personal risk of the patient calculated from family history, breast density, reproductive factors, demographic, clinical, imaging-related and genetic data (Clift et al., 2021). This highlights the great importance of genetic testing in identifying high risk individuals for screening and early detection. Mutations in several genes were associated with increased risks of breast cancer, according to the Breast Cancer Association Consortium, this includes: *BRCA1, BRCA2, PALB2, CHEK2, ATM, BARD1, MSH6, RAD51C, RAD51D, NF1, TP53* and *PTEN* (Dorling et al., 2021).

BRCA1 and BRCA2 are tumor suppressors that aid in repairing damaged DNA or destroy cells if DNA cannot be repaired (Yoshida & Miki, 2004). These genes are the two major breast and ovarian cancer predisposition genes. Mutations in *BRCA1* and *BRCA2* account for up to 90% of familial breast and ovarian cancer cases (Ford et al., 1998; Mahdavi et al., 2019). The prevalence of mutation in one of those genes was previously estimated to be approximately 1 in every 400 women, nonetheless, recent studies found an overall prevalence of up to 1 in 139 individuals of the general population (Group, 2000; McClain, Palomaki, Nathanson, & Haddow, 2005; Manickam et al., 2018; Abul-Husn et al., 2020). It was estimated that the cumulative breast cancer risk for a 70-year-old woman is up to 87% for *BRCA1* and 84% for *BRCA2* mutation carriers with corresponding ovarian cancer risks up to 68% and 30%, respectively. The prevalence of breast cancer in those females was estimated to be 10–30 times more than in those with no inherited gene mutation (Antoniou et al., 2003; Begg et al., 2008; Brohet et al., 2014; S. Chen et al., 2006; Evans et al., 2008; Ford et al., 1998; Gabai-Kapara et al., 2014; Hopper et al., 1999; Kuchenbaecker et al., 2017; Milne et al., 2008; Park et al., 2021).

Researchers have identified thousands of mutations in *BRCA* genes, some of which were determined to be harmful, while others have no proven impact. The risk associated with any given variant varies significantly and depends on the exact type and location of the variant (Dorling et al., 2022; H. Li et al., 2022; López-Urrutia et al., 2019; Morris & Gordon, 2010). High risk variants typically disrupt the gene function; however, the functional impact of many variants cannot be deduced from their sequence information alone. Such variants are defined as variants of uncertain significance (VUS) and they represent a major challenge for the management of families, in which they are identified (Eccles et al., 2015; López-Urrutia et al., 2019). Worldwide genetic testing has uncovered thousands of VUS in the *BRCA* genes, including missense substitutions, in-frame insertions and deletions, silent alterations that may influence splicing or translation and intronic changes of unknown influence on gene splicing or expression (Spurdle et al., 2012; López-Urrutia et al., 2019; NCBI ClinVar database, 2021).

A consistent variant classification system is essential to the use of genomics in patient care. The 2015 joint recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP 2015 guidelines) classifies sequence variants into five categories: pathogenic, likely pathogenic, uncertain significance, likely benign, and benign (Richards et al., 2015). For best classification of cancer gene variants, the probability of pathogenicity is based on *in silico* analysis of the sequence alteration in combination with the available genetic, epidemiological and clinical data, such as: segregation analysis, personal and family history, tumor histopathology, and co-occurrence (Lindor et al., 2012; Richards et al., 2015). While several assumptions are made in these calculations, this approach has been widely used to classify variants as pathogenic or benign. It is the currently accepted method for classifying *BRCA* variants by the Evidence-Based Network for the Interpretation of Germline Mutant Alleles Consortium (ENIGMA) that specializes in clinical classification of *BRCA* variants (Spurdle et al., 2012), the ClinVar database of variants, and the Global Alliance for Genomic Health organization in their BRCA Exchange database (Cline et al., 2018). It is noteworthy that there are still >29,000 *BRCA2* variants in the BRCA Exchange database that have not been reviewed for classification.

The number of identified germline variants in *BRCA2* outpace the clinical annotation due to the limited availability of genetic, epidemiological, and clinical data, which highlights the importance and the practicality of computational methods for risk assessment, as well as the need to prioritize *BRCA2* variants for functional testing or classification. In fact, there was a recent call to action to complement the use of the ClinVar database with computational predictors to enhance the actionability of rare breast cancer-gene variants (Saad et al.,

TABLE 1 The receiver operating characteristic (ROC) curve analysis for the different in silico predictions. AUC: Area under the curve, Obs: Number of observations, Std.Err: Standard error, CI: Confidence interval.

In silico prediction method	Observation	AUC	Std.Err	95% CI Low	95% CI High
XGBoost	820	1	0	0.99996	1
Consequence	4102	0.9978	0.0009	0.9961	0.99945
IMPACT	4102	0.9986	0.0005	0.99765	0.99957
SIFT_Score	142	0.1166	0.0237	0.07008	0.16318
PolyPhen_Scor	142	0.8811	0.0493	0.78446	0.97781
BayesDel_addAF_rankscore	717	0.9896	0.005	0.97974	0.99946
BayesDel_noAF_rankscore	717	0.9657	0.0105	0.94517	0.98617
CADD_raw_rankscore	717	0.9916	0.0041	0.9836	0.99968
ClinPred_rankscore	142	0.9922	0.0054	0.98165	1
DANN_rankscore	717	0.6899	0.0341	0.6231	0.75676
Eigen-PC-raw_coding_rankscore	717	0.8116	0.0255	0.76169	0.86157
Eigen-raw_coding_rankscore	717	0.8641	0.0225	0.81998	0.90832
FATHMM_converted_rankscore	142	0.9238	0.0437	0.8381	1
GERP++_RS_rankscore	717	0.658	0.0271	0.60495	0.71103
GM12878_fitCons_rankscore	717	0.4681	0.0266	0.41598	0.52016
GenoCanyon_rankscore	717	0.5597	0.0264	0.50789	0.61149
H1-hESC_fitCons_rankscore	717	0.5535	0.0285	0.49759	0.6095
HUVEC_fitCons_rankscore	717	0.5132	0.0257	0.46293	0.56354
LRT_converted_rankscore	717	0.5833	0.0273	0.52976	0.63693
M-CAP_rankscore	120	0.9181	0.0341	0.85119	0.985
MPC_rankscore	141	0.9294	0.0295	0.87164	0.98714
MVP_rankscore	135	0.9563	0.0184	0.92017	0.99247
MetaLR_rankscore	142	0.9414	0.0373	0.86837	1
MetaRNN_rankscore	142	0.995	0.0038	0.98745	1
MetaSVM_rankscore	142	0.9096	0.0688	0.77467	1
MutPred_rankscore	46	0.9821	0.0147	0.95336	1
MutationTaster_rankscore	717	0.984	0.0077	0.96888	0.99921
PROVEAN_converted_rankscore	142	0.596	0.068	0.46263	0.72934
PrimateAI_rankscore	141	0.9153	0.0666	0.78482	1
REVEL_rankscore	142	0.9531	0.0217	0.91052	0.99573
SiPhy_29way_logOdds_rankscore	717	0.6476	0.0266	0.59544	0.69976
VEST4_rankscore	717	0.9948	0.0023	0.99039	0.99925
bStatistic_converted_rankscore	717	0.525	0.0271	0.47193	0.57798
fathmm-MKL_coding_rankscore	717	0.6825	0.0281	0.62749	0.73749
fathmm-XF_coding_rankscore	717	0.4945	0.031	0.43373	0.55523
integrated_fitCons_rankscore	717	0.5015	0.0262	0.45006	0.55286
phastCons17way_primate_rankscore	717	0.5545	0.0287	0.49815	0.61085
phyloP17way_primate_rankscore	717	0.4938	0.0341	0.42692	0.56064
MaxEntScan_alt	64	0.1129	0.0416	0.03141	0.19444
MaxEntScan_diff	64	0.8333	0.0496	0.73613	0.93054
MaxEntScan_ref	64	0.3891	0.0843	0.22388	0.55435
SpliceAI_pred_DS_AG	3655	0.5148	0.0048	0.50537	0.52418
SpliceAI_pred_DS_AL	3655	0.5149	0.0029	0.50917	0.52065
SpliceAI_pred_DS_DG	3655	0.5016	0.0032	0.49532	0.5078
SpliceAI_pred_DS_DL	3655	0.5202	0.0032	0.51396	0.52638

2022). Moreover, the existing *BRCA2* pathogenicity prediction algorithms focus on the prediction of the functional impact, as measured by functional assays, of missense variants only. Therefore, our aim is to develop a gene-specific machine learning model to predict pathogenicity according to the comprehensive ACMG guidelines and for all types of *BRCA2* variants, utilizing novel features. We will use this new model to predict the pathogenicity of all *BRCA2* variants that have not been classified yet and prioritize them according to their predicted level of pathogenicity.

Materials and methods

BRCA2 set of variants

We downloaded *BRCA2* variants from the BRCA Exchange database, which contains information drawn from multiple databases that provide a comprehensive list of *BRCA1* and *BRCA2* variants with their annotations (https://brcaexchange.org/variants; accessed on 14 March 2022). It contains variants curated and classified by an international consortium of investigators (ENIGMA consortium) to assess variant pathogenicity. At the time of this study, there were 33,550 *BRCA2* variants, of which 4,102 variants were reviewed by the ENIGMA expert panel and had known effect of being pathogenic (2,672), likely benign (738) or benign (692).

Variant annotation

The Ensembl Variant Effect Predictor determines the effect of any variant on genes, transcripts, and protein sequence, as well as on regulatory regions. It is a tool for the analysis and annotation of genomic variants. It provides information on the affected transcript, protein, non-coding region, on the frequency and the phenotypes associated with the variant. Additionally, it provides access to numerous *in silico* pathogenicity prediction scores that are present in the dbNSFP database (Liu, Jian, & Boerwinkle, 2011). The *in silico* predictions we included in the model were BayesDel_addAF, BayesDel_noAF, bStatistic, CADD, ClinPred, DANN, Eigen, EigenPC, FATHMM-XF coding, FATHMM-MKL coding, GenoCanyon, GERP++RS, GM12878fitCons, H1hESCfitCons, HUVECfitCons, integratedfitCons, LRT, MaxEntScan, MCAP, MetaLR, MetaSVM, MutationAssessor, MutationTaster2, MutPred, MPC, MVP, phastCons, PhyloP, Polyphen, PrimateAI, PROVEAN, REVEL, SIFT, SiPhy, SpliceAI, and VEST4. For the *in silico* predictions, we used the rank scores whenever they were provided. The detailed list is presented in Table1.

Other variables that were collected or derived from VEP included the position of the variant, variant length (number of bases involved based on reference and alternative alleles), presence in protein domain, variant association with phenotype, presence as a somatic mutation, variant impact, and variant consequences. The variant consequence variable included 18 different effects of the variant position (Supplementary Table S1). We ranked them based on the assumed pathogenicity of the effect with downstream variants having the least effect and stop gained variants having the highest effect.

Allele frequencies

We obtained population frequency of the variants from both the BRCA Exchange database and from VEP, which included population frequency data from: Exome Aggregation Consortium, NHLBI exome sequencing, 1000 Genomes Project, gnomAD, UK10K cohort data, and the NHLBI Exome Sequencing Project ESP6500 data. We used the highest frequency reported for any given variant as a variable called maximum allele frequency in the model.

XGBoost

XGBoost (Extreme Gradient Boosting) is an open-source software, which provides a regularizing gradient boosting framework (T. Chen & Guestrin, 2016). It implements a highly flexible, optimized distributed gradient boosting machine learning algorithm under the Gradient Boosting framework through parallel processing to speed up calculations, regularization to avoid overfitting, tree-pruning and handling of missing values.

We chose XGBoost, because it is widely used in bioinformatics; some of those applications were for analyzing protein translocation between cellular organelles (Mendik et al., 2019); predicting gene expression values (W. Li, Yin, Quan, & Zhang, 2019); predicting early-stage prostate cancer (Danciu et al., 2020); identifying the origin of DNA replication (Do & Le, 2020); and predicting Kruppel-like factors (Le, Do, & Le, 2021). Additionally, XGBoosted Machine learning performed better than other predictive models, including Linear models, Gradient Boosting Machines, Neural Networks, Random Forests, and Extremely Randomized Forests, in predicting the functional impact of *BRCA2* missense variants (Hart, Polley, Shimelis, Yadav, & Couch, 2020).

Model building

We used the XGBoost R package (version 1.4.1.1) with default parameters (booster = "gbtree", objective = "binary: logistic", eta = 0.3, gamma = 0, max_depth = 6,

min_child_weight = 1, subsample = 1, colsample_bytree = 1, nrounds = 100) to train a classifier model on the variant annotations for predicting pathogenicity. Pathogenicity was based on the ENIGMA expert panel's review. Therefore, only the variants that have been reviewed were included in building the model (4,012 variants), and they were split into 80% training set and 20% test set. The original variant pathogenicity groups were recategorized as pathogenic ("pathogenic" and "likely pathogenic") and benign ("benign" and "likely benign") for the binary classification. The model was trained to predict the expert classification of either pathogenic or benign variants and we performed 5-fold cross-validation of the model. We used the xgb.plot.importance function to show which are the top 10 most important features of the model (gain was used as the measure of importance). The Shapely values were also examined to find the most predictive characteristics and prediction scores (xgb.plot.shap function and SHAPforxgboost package). Finally, we predicted the pathogenicity of the 29,448 unreviewed variants.

Testing the model on independent VUS with functional data

Richardson et al., in 2021 assessed the functional effect of 252 *BRCA2* VUS by a *BRCA2* homology-directed DNA repair (HDR) assay. Utilizing the Variant Recoder tool in Ensembl, the 252 BRCA2 amino acid changes corresponded to 276 missense sequence variants. Out of the 276 variants, 251 were not reviewed by the BRCA Exchange database and 4 of them were both missense and splice region variants. Accordingly, 247 variants were used for independent assessment of the model on missense VUS. Those VUS had functional data on their ability to complement DR-GFP BRCA2 deficient V-C8 cells in a BRCA2 homology-directed DNA repair (HDR) assay. Known pathogenic variants with functional defects had HDR scores <1.66, and known benign variants that were nonfunctional had HDR scores >2.44 (Richardson et al., 2021). More extreme HDR scores of <1.0 and >3.0 have also been utilized in the literature for pathogenic and benign variants, respectively (Guidugli et al., 2018), therefore we tested the model's performance with both cut-offs.

Results

Variant datasets

At the time of data collection, the BRCA Exchange database had 33,550 *BRCA2* variants. The largest proportion of those variants were intronic (36%), and of those found in the coding region, the majority were missense (53%: Supplementary Figures S1A,B). Only 4,102 variants were reviewed by the expert panel and had a known effect: 2,672 were pathogenic, and 1,430 were benign. The distribution of the variants in the expert reviewed portion markedly differed from the distribution of all variants. Out of the reviewed 4,102 variants, 29% were frame shift, compared to the 5% of all variants, and intronic variants were only 9%, compared to the 36% of all variants (Supplementary Figures S1A,C). There was an even more pronounced distribution difference for coding sequence variants, with the proportion of frameshift, synonymous and stop gained variants being much higher for the reviewed variants, while only 4% of them were missense variants (Supplementary Figures S1B,D). This highlights that a large proportion of missense variants could not be unambiguously assigned to either pathogenic or benign categories.

Variant location

Out of the 33,550 *BRCA2* variants, 14,259 were present along the 27 *BRCA2* exons. Out of the 4,012 reviewed variants, the highest number of both benign and pathogenic variants were found in exon 11 (47.6%) followed by exon 10 (10.4%) (Figure 1A). Similarly, the highest number of specifically missense variants were present in Exons 10 and 11 (41.1% and 15.6% of all missense variants, respectively). However, exon 10 and 11 had only benign missense variants. Pathogenic missense variants were present in exons 13, 17, 18, 24 and 25 (Supplementary Figure S2). Only a small fraction of the reviewed variants were intronic, and most of those were determined to be benign (Figure 1B).

Predicting pathogenicity of the ENIGMA reviewed variants

To develop the prediction model, we used an extreme gradient boosting machine learning algorithm (XGBoost) and included the variants with known expert reviewed effect (4,102 variants). Variants were divided into a training (80%) and a test set (20%). The training model included 3,282 variants and it was trained to predict the expert classification of either pathogenic (2,118 variants) or benign (1,164 variants) variants. The test model included 554 pathogenic and 266 benign variants.

The model was used to predict the test group of 820 variants and yielded an accuracy of 0.999 with sensitivity of 99.6% and the specificity of 100% (Figure 2A). The most important variable was the variant consequence followed by a combination of different *in silico* prediction tools (Figure 2B, Supplementary Figure S3A). Removing the consequence variable from the model did not affect the accuracy (0.996) and the maximum allele frequency became the most important feature followed by the Combined Annotation Dependent Depletion (CADD) Phred score and the number of involved bases "variant length" (Figure 2C,

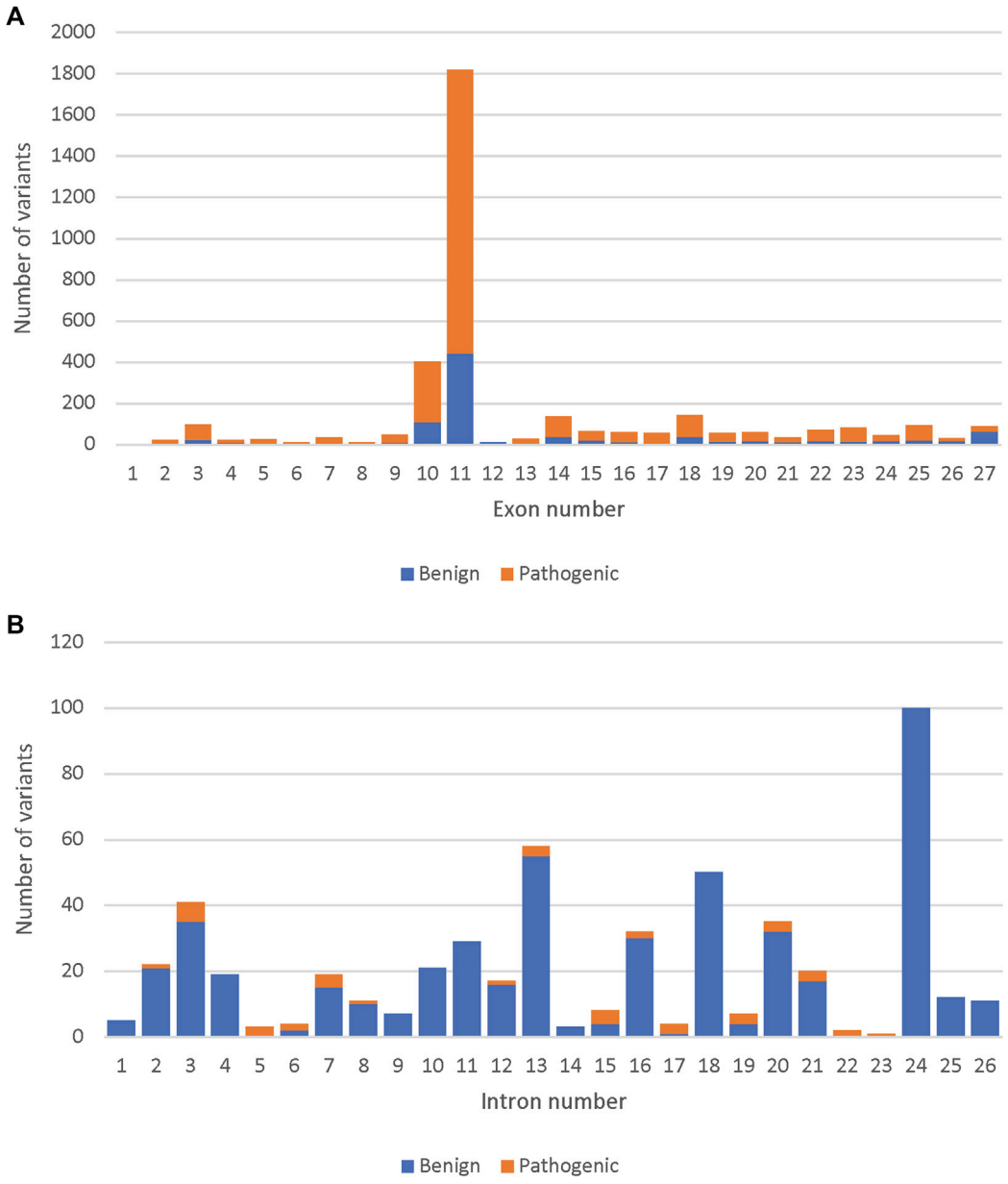


FIGURE 1
Comparison of the number of pathogenic and benign variants among the 4,102 reviewed, across the BRCA2 gene. **(A)** The number of pathogenic and benign variants per BRCA2 exons. **(B)** The number of pathogenic and benign variants per BRCA2 introns.

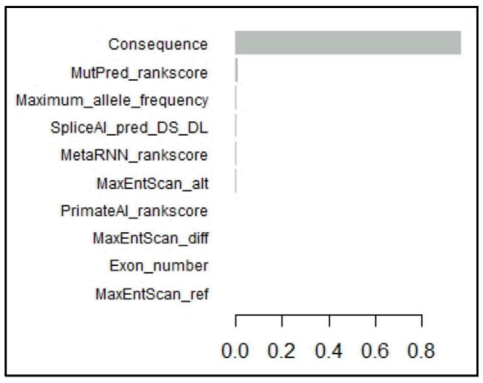
Supplementary Figure S3B). We performed cross validation of the *BRCA2* model with 5 different subsamples that included random training and test groups. All models demonstrated similar accuracies between 99.6% and 99.9% (Supplementary Table S2). Similar to the original model, the variant consequence was the most important variable across the 5 subsamples, and when it was removed, the maximum allele frequency became the most important feature.

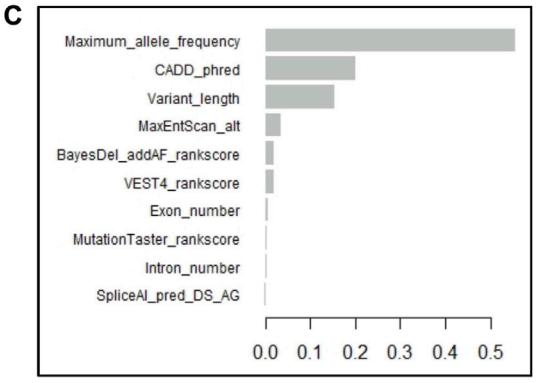
Comparison of the novel model to previous *in silico* prediction algorithms

We compared the area under the curve for our novel XGBoost model and the input *in silico* prediction tools on their own. The XGBoost model had the highest AUC of 1.00, followed by VEST4, ClinPred, and CADD rank scores (Table 1). It should be noted that, the receiver

A

Model	Accuracy (95%)	Sensitivity	Specificity	P-value
Full model	0.999 (0.993, 1)	0.996	1	<2.00E-16
Without consequence	0.995 (0.988, 0.999)	0.993	0.996	<2.00E-16





D

Model	Accuracy (95%)	Sensitivity	Specificity	P-value
HDR cutoff <= 1.66 =>2.44	0.834 (0.782, 0.878)	0.918	0.682	2.835E-11
HDR cutoff <= 1.0 >=3.0	0.913 (0.863, 0.952)	0.913	0.909	0.03782

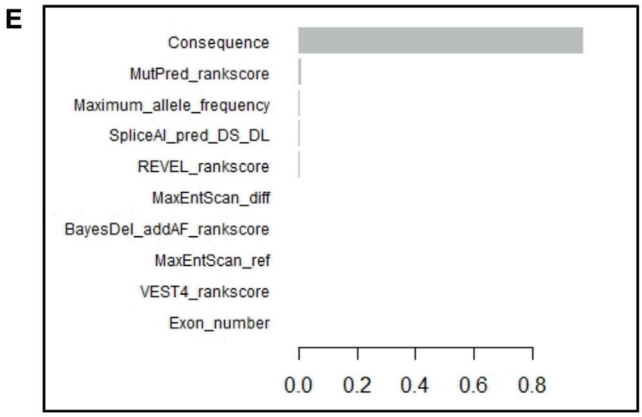


FIGURE 2
The BRCA2 XGBoost models. **(A)** The models characteristics (AUC: Area Under the Curve). **(B)** Feature importance of the XGBoost model. **(C)** Feature importance of the XGBoost model without consequence. The BRCA2 XGBoost model trained on the whole reviewed dataset (4,102 variants) was used to predict VUS pathogenicity based on HDR functional assay scores. **(D)** The performance of the model on a set of pathogenic and benign variants according to HDR cutoffs <= 1.66 and =>2.44 (247 variants) and cutoffs <= 1.0 and >=3.0 (160 variants). **(E)** Feature importance of the XGBoost model trained on the whole reviewed dataset.

operating characteristic (ROC) analysis of the different *in silico* predictions was performed using the full sample of 4,102 variants and the AUCs were calculated for only those variants that had prediction scores for the given tool.

We also performed ROC analysis for the association of the consequence, which demonstrated excellent diagnostic abilities with the area under the curve (AUC) equal to 99.8% (Supplementary Figure S4). The cutoff point of more than 12 (missense variant) had the best balance between sensitivity (99.4%) and specificity (99.9%). Thus, we can expect that the model's accuracy in identifying only VUS will decrease, because VUS are usually missense variants.

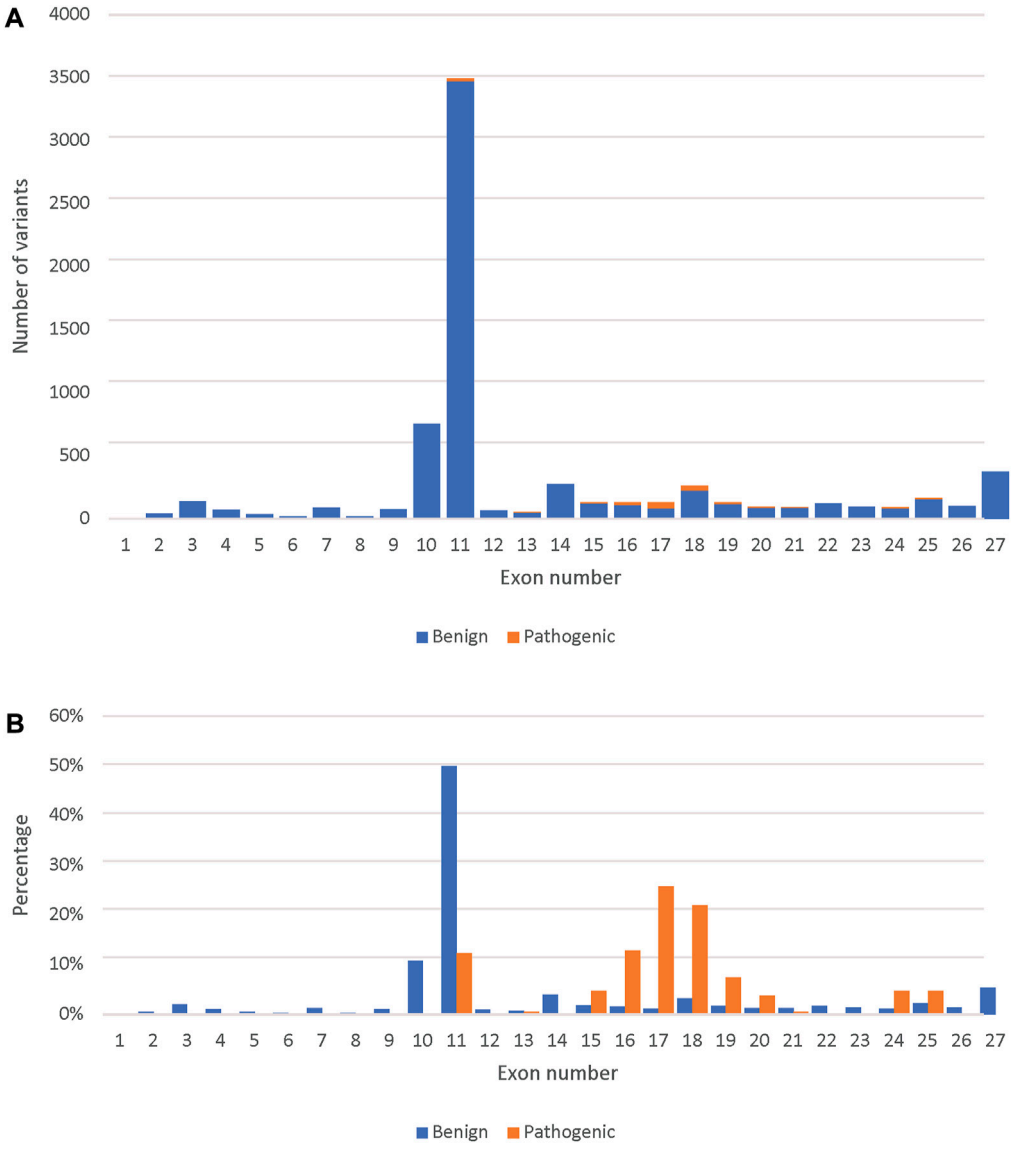


FIGURE 3
Comparison between the predicted pathogenic and benign variants across the BRCA2 exons. **(A)** The number of predicted pathogenic and benign missense variants (7,131) per BRCA2 exons. **(B)** Percent distribution of predicted pathogenic and benign missense variants across the BRCA2 exons.

Model validation in predicting VUS

We obtained missense VUS with functional data from a recent study by Richardson et al. After removing variants that were already included in building the model, 247 variants were used to assess the model's performance in predicting variants of uncertain significance. Out of 247 VUS, 88 demonstrated functional defects with HDR scores <1.66, and 159 variants were considered benign with HDR >2.44.

The *BRCA2* model, trained on the full set of ENIGMA *BRCA2* variants (4,102), was tasked to predict the VUS that demonstrated functional defects. The model had high accuracy of 0.834 with sensitivity of 91.8% and specificity of 68.2% (Figure 2D). The most important variable was the variant consequence (Figure 2E, Supplementary Figure S5). The diagnostic performance of the model significantly improved with the more extreme HDR cutoff points of <= 1 for pathogenic and >=3 for benign variants (160 variants). The accuracy of the model increased to

0.913 with a sensitivity of 91.3% and specificity of 90.9% (Figure 2D).

Model pathogenicity predictions for the not reviewed variants

Finally, we used the novel gene-specific *BRCA2* model to predict the remaining 29,448 variants present in the BRCA Exchange database that are not yet reviewed by the expert panel. We predicted 2,092 variants to be pathogenic and prioritized them according to the total SHAP values of the different predictors (Supplementary Table S3). We predicted 186 pathogenic missense variants (Figure 3A). The majority of those are in the DNA-binding domain (exons 12–26), however 23 were in exon 11, which is outside of it (Figure 3B).

Discussion

The highest number of both benign and pathogenic variants were found in exon 11 followed by exon 10, which was expected as those two exons represent around 65% of the *BRCA2* coding sequence. However, looking only at expert reviewed missense variants, exons 10 and 11 had only benign missense variants. This is in agreement with previous studies that referred to exon 10 and 11 as "coldspots" which were described as "regions of a gene that are tolerant of variation, where pathogenic missense variants are unlikely" (Dines et al., 2020). However, we predicted 23 pathogenic missense variants in exon 11, which fell into the BRCA2 BRC repeats that binds to RAD51 and DSS1 resulting in the RAD51–BRCA2–DSS1 complex (Shailani, Kaur, & Munshi, 2018), indicating that these missense variants are likely to effect the complex's stability. This suggests that only exon 10 is a "coldspot".

We have demonstrated that the gene-specific *BRCA2* model is an extremely accurate method for predicting variant pathogenicity in the *BRCA2* gene according to the classification by the ENGIMA group. Moreover, the model demonstrated excellent abilities in predicting damaging missense variants of uncertain significance. The gene-specific model demonstrated better diagnostic probabilities than other *in silico* prediction tools. In contrast to other gene-specific models or *in silico* predictions, our model was built to predict the ENGIMA final classification (Hart et al., 2019; Hart et al., 2020). Therefore, it encompasses not only missense variants that are tested in functional studies but all possible variant types. The previously published models or predictions are built specifically for missense variants and to predict their functional impact as tested by functional assays. In fact, the *BRCA2* model developed by Hart *et al.* was limited to missense mutations in the DNA-binding domain of the BRCA2 protein known to be associated with impaired function (Hart et al., 2020). Moreover, the existing *BRCA2* model was trained and tested with only 202 *BRCA2* variants. It is based on small numbers of known damaging mutations, which limits both the model's ability to capture the variability of variant data and the direct comparison between the two gene-specific models. Nevertheless, to compare our model to the previous one, we examined the performance of our model to predict only the missense variants, which are present in the testing group and calculated the Matthews Correlation Coefficient (MCC). Our model had an MCC of 0.849 which is better than the MCC of 0.73 reported for the previous *BRCA2* model (Hart et al., 2020).

Despite the increasing number of variants that have been functionally tested, there are still 29,448 *BRCA2* variants that have not been classified by the BRCA Exchange expert panel (ENIGMA). Variant classification is based on the probability of pathogenicity that includes *in silico* analysis of the sequence alteration in combination with the available genetic, epidemiological, and clinical data, as well as functional studies (Lindor et al., 2012). All these underscore the importance and current need of computational methods to predict and prioritize variants for classification or functional testing. Our prioritized list of so far unreviewed variants could guide future efforts in studying damaging mutations and aid genetic counselors and researchers for interpreting the pathogenicity of different *BRCA2* variants.

There are still limitations for the *BRCA2* model. The fact that variants at the lower and upper extremes of HDR scores had better predictions emphasize that variants with intermediate HDR scores are more challenging for the model. These variants are also likely to represent a set of variants with variable functional effect that might depend on other variants or external factors. Moreover, we did not optimize the model parameters, therefore the model might perform better with optimal settings. We also did not systematically test whether leaving out certain *in silico* prediction variables would improve the model's performance. While a model with an optimal set of variables might exist, the XGBoost algorithm is resistant to redundant information, and therefore we do not foresee a significant improvement over including all available *in silico* predictions.

The *BRCA2* gene-specific model is an accurate method for interpreting the pathogenicity of all types of variants in the *BRCA2* gene as they were classified according to the ACMG criteria. It is a valuable addition for variant classification and can prioritize unreviewed variants for functional analysis or expert review. Finally, our approach could be utilized for other high-risk cancer genes that have a large number of variants with high-confidence pathogenicity annotation.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

MK and BM conceptualized the project. MK performed the analyses under BM's supervision. MK and BM wrote the manuscript.

Acknowledgments

We are grateful to William Villiers, who critically read the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.982930/full#supplementary-material

References

Abul-Husn, N. S., Soper, E. R., Odgis, J. A., Cullina, S., Bobo, D., Moscati, A., et al. (2020). Exome sequencing reveals a high prevalence of BRCA1 and BRCA2 founder variants in a diverse population-based biobank. *Genome Med.* 12 (1), 2–12. doi:10.1186/s13073-019-0691-1

Antoniou, A., Pharoah, P. D., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., et al. (2003). Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies. *Am. J. Hum. Genet.* 72 (5), 1117–1130. doi:10.1086/375033

Begg, C. B., Haile, R. W., Borg, Å., Malone, K. E., Concannon, P., Thomas, D. C., et al. (2008). Variation of breast cancer risk among BRCA1/2 carriers. *Jama* 299 (2), 194–201. doi:10.1001/jama.2007.55-a

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R., Torre, L., and Jemal, A. (2020). Erratum: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca. Cancer J. Clin.* 70 (4), 313. doi:10.3322/caac.21609

Brohet, R. M., Velthuizen, M. E., Hogervorst, F. B., Meijers-Heijboer, H. E., Seynaeve, C., Collée, M. J., et al. (2014). Breast and ovarian cancer risks in a large series of clinically ascertained families with a high proportion of BRCA1 and BRCA2 Dutch founder mutations. *J. Med. Genet.* 51 (2), 98–107. doi:10.1136/jmedgenet-2013-101974

Canelo-Aybar, C., Ferreira, D. S., Ballesteros, M., Posso, M., Montero, N., Solà, I., et al. (2021). Benefits and harms of breast cancer mammography screening for women at average risk of breast cancer: A systematic review for the European commission initiative on breast cancer. *J. Med. Screen.* 28, 389–404. doi:10.1177/0969141321993866

Chen, S., Iversen, E. S., Friebel, T., Finkelstein, D., Weber, B. L., Eisen, A., et al. (2006). Characterization of BRCA1 and BRCA2 mutations in a large United States sample. *J. Clin. Oncol.* 24 (6), 863–871. doi:10.1200/JCO.2005.03.6772

Chen, T., and Guestrin, C. (2016). "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. Paper presented at the.

Clift, A. K., Dodwell, D., Lord, S., Petrou, S., Brady, S. M., Collins, G. S., et al. (2021). The current status of risk-stratified breast screening. *Br. J. Cancer* 126, 533–550. doi:10.1038/s41416-021-01550-3

Cline, M. S., Liao, R. G., Parsons, M. T., Paten, B., Alquaddoomi, F., Antoniou, A., et al. (2018). BRCA challenge: BRCA exchange as a global resource for variants in BRCA1 and BRCA2. *PLoS Genet.* 14 (12), e1007752. doi:10.1371/journal.pgen.1007752

Danciu, I., Erwin, S., Agasthya, G., Janet, T., McMahon, B., Tourassi, G., et al. (2020). "Using longitudinal PSA values and machine learning for predicting progression of early stage prostate cancer in veterans," in *American society of clinical oncology*.

Dines, J. N., Shirts, B. H., Slavin, T. P., Walsh, T., King, M.-C., Fowler, D. M., et al. (2020). Systematic misclassification of missense variants in BRCA1 and BRCA2 "coldspots". *Genet. Med.* 22 (5), 825–830. doi:10.1038/s41436-019-0740-6

Do, D. T., and Le, N. Q. K. (2020). Using extreme gradient boosting to identify origin of replication in *Saccharomyces cerevisiae* via hybrid features. *Genomics* 112 (3), 2445–2451. doi:10.1016/j.ygeno.2020.01.017

Dorling, L., Carvalho, S., Allen, J., Gonzalez-Neira, A., Luccarini, C., Wahlström, C, ., et al. (2021). Breast cancer risk genes-association analysis in more than 113, 000 women. *N. Engl. J. Med.* 384 (5), 428–439. doi:10.1056/NEJMoa1913948

Dorling, L., Carvalho, S., Allen, J., Parsons, M. T., Fortuno, C., González-Neira, A., et al. (2022). Breast cancer risks associated with missense variants in breast cancer susceptibility genes. *Genome Med.* 14 (1), 51–17. doi:10.1186/s13073-022-01052-8

Eccles, D., Mitchell, G., Monteiro, A., Schmutzler, R., Couch, F., Spurdle, A., et al. (2015). BRCA1 and BRCA2 genetic testing—Pitfalls and recommendations for managing variants of uncertain clinical significance. *Ann. Oncol.* 26 (10), 2057–2065. doi:10.1093/annonc/mdv278

Evans, D. G., Shenton, A., Woodward, E., Lalloo, F., Howell, A., and Maher, E. R. (2008). Penetrance estimates for BRCA1 and BRCA2based on genetic testing in a Clinical Cancer Genetics service setting: Risks of breast/ovarian cancer quoted should reflect the cancer burden in the family. *BMC cancer* 8 (1), 155–159. doi:10.1186/1471-2407-8-155

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., et al. (2020). *Global cancer observatory: Cancer today*. Lyon, France: International Agency for Research on Cancer. Available from: https://gco.iarc.fr/today (accessed November 20, 2021).

Ford, D., Easton, D., Stratton, M., Narod, S., Goldgar, D., Devilee, P., et al. (1998). Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am. J. Hum. Genet.* 62 (3), 676–689. doi:10.1086/301749

Gabai-Kapara, E., Lahad, A., Kaufman, B., Friedman, E., Segev, S., Renbaum, P., et al. (2014). Population-based screening for breast and ovarian cancer risk due to BRCA1 and BRCA2. *Proc. Natl. Acad. Sci. U. S. A.* 111 (39), 14205–14210. doi:10.1073/pnas.1415979111

Gøtzsche, P., and Kj, J. (2013). Cochrane breast cancer group. *Screen. breast cancer Mammogr. Cochrane Database Syst Rev* 156 (101002), 14651858.

Group, A. B. C. S. (2000). Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. *Br. J. Cancer* 83 (10), 1301.

Guidugli, L., Shimelis, H., Masica, D. L., Pankratz, V. S., Lipton, G. B., Singh, N., et al. (2018). Assessment of the clinical relevance of BRCA2 missense variants by functional and computational approaches. *Am. J. Hum. Genet.* 102 (2), 233–248. doi:10.1016/j.ajhg.2017.12.013

Hart, S. N., Hoskin, T., Shimelis, H., Moore, R. M., Feng, B., Thomas, A., et al. (2019). Comprehensive annotation of BRCA1 and BRCA2 missense variants by functionally validated sequence-based computational prediction models. *Genet. Med.* 21 (1), 71–80. doi:10.1038/s41436-018-0018-4

Hart, S. N., Polley, E. C., Shimelis, H., Yadav, S., and Couch, F. J. (2020). Prediction of the functional impact of missense variants in BRCA1 and BRCA2 with BRCA-ML. *NPJ breast cancer* 6 (1), 13–14. doi:10.1038/s41523-020-0159-x

Hopper, J. L., Southey, M. C., Dite, G. S., Jolley, D. J., Giles, G. G., and McCredie, M. R. (1999). Population-based estimate of the average age-specific cumulative risk of breast cancer for a defined set of protein-truncating mutations in BRCA1 and BRCA2. Australian Breast Cancer Family Study. *Cancer Epidemiol. Biomarkers Prev.* 8 (9), 741–747.

Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R., Phillips, K.-A., Mooij, T. M., Roos-Blom, M.-J., et al. (2017). Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *Jama* 317 (23), 2402–2416. doi:10.1001/jama.2017.7112

Le, N. Q. K., Do, D. T., and Le, Q. A. (2021). A sequence-based prediction of Kruppel-like factors proteins using XGBoost and optimized features. *Gene* 787, 145643. doi:10.1016/j.gene.2021.145643

Li, H., Engel, C., de la Hoya, M., Peterlongo, P., Yannoukakos, D., Livraghi, L., et al. (2022). Risks of breast and ovarian cancer for women harboring pathogenic missense variants in BRCA1 and BRCA2 compared with those harboring protein truncating variants. *Genet. Med.* 24 (1), 119–129. doi:10.1016/j.gim.2021.08.016

Li, W., Yin, Y., Quan, X., and Zhang, H. (2019). Gene expression value prediction based on XGBoost algorithm. *Front. Genet.* 10, 1077. doi:10.3389/fgene.2019.01077

Lindor, N. M., Guidugli, L., Wang, X., Vallée, M. P., Monteiro, A. N., Tavtigian, S., et al. (2012). A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). *Hum. Mutat.* 33 (1), 8–21. doi:10.1002/humu.21627

Liu, X., Jian, X., and Boerwinkle, E. (2011). dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* 32 (8), 894–899. doi:10.1002/humu.21517

López-Urrutia, E., Salazar-Rojas, V., Brito-Elías, L., Coca-González, M., Silva-García, J., Sánchez-Marín, D., et al. (2019). BRCA mutations: Is everything said? *Breast Cancer Res. Treat.* 173 (1), 49–54. doi:10.1007/s10549-018-4986-5

Mahdavi, M., Nassiri, M., Kooshyar, M. M., Vakili-Azghandi, M., Avan, A., Sandry, R., et al. (2019). Hereditary breast cancer; Genetic penetrance and current status with BRCA. *J. Cell. Physiol.* 234 (5), 5741–5750. doi:10.1002/jcp.27464

Manickam, K., Buchanan, A. H., Schwartz, M. L., Hallquist, M. L., Williams, J. L., Rahm, A. K., et al. (2018). Exome sequencing–based screening for BRCA1/2 expected pathogenic variants among adult biobank participants. *JAMA Netw. Open* 1 (5), e182140. doi:10.1001/jamanetworkopen.2018.2140

McClain, M. R., Palomaki, G. E., Nathanson, K. L., and Haddow, J. E. (2005). Adjusting the estimated proportion of breast cancer cases associated with BRCA1 and BRCA2 mutations: Public health implications. *Genet. Med.* 7 (1), 28–33. doi:10.1097/01.gim.0000151155.36470.ff

McGuire, A., Brown, J. A., Malone, C., McLaughlin, R., and Kerin, M. J. (2015). Effects of age on the detection and management of breast cancer. *Cancers* 7 (2), 908–929. doi:10.3390/cancers7020815

Mendik, P., Dobronyi, L., Hári, F., Kerepesi, C., Maia-Moco, L., Buszlai, D., et al. (2019). Translocatome: A novel resource for the analysis of protein translocation between cellular organelles. *Nucleic Acids Res.* 47 (D1), D495–D505. doi:10.1093/nar/gky1044

Milne, R. L., Osorio, A., Cajal, T. R. n. y., Vega, A., Llort, G., De La Hoya, M., et al. (2008). The average cumulative risks of breast and ovarian cancer for carriers of mutations in BRCA1 and BRCA2 attending genetic counseling units in Spain. *Clin. Cancer Res.* 14 (9), 2861–2869. doi:10.1158/1078-0432.CCR-07-4436

Morris, J. L., and Gordon, O. K. (2010). *Positive results: Making the best decisions when you're at high risk for breast or ovarian cancer*. Amherst, NY: Prometheus Books.

NCBI ClinVar database (2021). BRCA1 and BRCA2 database. Retrieved from https://www.ncbi.nlm.nih.gov/clinvar/.

Noone, A., Howlader, N., Krapcho, M., Miller, D., Brest, A., Yu, M., et al. (2018). *SEER cancer statistics review*. Bethesda, MD: National Cancer Institute, 1975–2015.

Park, K.-S., Lee, W., Seong, M.-W., Kong, S.-Y., Lee, K.-A., Ha, J.-S., et al. (2021). A population-based analysis of brca1/2 genes and associated breast and ovarian cancer risk in Korean patients: A multicenter cohort study. *Cancers* 13 (9), 2192. doi:10.3390/cancers13092192

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of medical genetics and genomics and the association for molecular Pathology. *Genet. Med.* 17 (5), 405–424. doi:10.1038/gim.2015.30

Richardson, M. E., Hu, C., Lee, K. Y., LaDuca, H., Fulk, K., Durda, K. M., et al. (2021). Strong functional data for pathogenicity or neutrality classify BRCA2 DNA-binding-domain variants of uncertain significance. *Am. J. Hum. Genet.* 108 (3), 458–468. doi:10.1016/j.ajhg.2021.02.005

Saad, M., Mokrab, Y., Halabi, N., Shan, J., Razali, R., Kunji, K., et al. (2022). Genetic predisposition to cancer across people of different ancestries in Qatar: A population-based, cohort study. *Lancet. Oncol.* 23 (3), 341–352. doi:10.1016/S1470-2045(21)00752-X

Shailani, A., Kaur, R. P., and Munshi, A. (2018). A comprehensive analysis of BRCA2 gene: Focus on mechanistic aspects of its functions, spectrum of deleterious mutations, and therapeutic strategies targeting BRCA2-deficient tumors. *Med. Oncol.* 35 (3), 18–10. doi:10.1007/s12032-018-1085-8

Spurdle, A. B., Healey, S., Devereau, A., Hogervorst, F. B., Monteiro, A. N., Nathanson, K. L., et al. (2012). ENIGMA—Evidence-based network for the interpretation of germline mutant alleles: An international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Hum. Mutat.* 33 (1), 2–7. doi:10.1002/humu.21628

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca. Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Yoshida, K., and Miki, Y. (2004). Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. *Cancer Sci.* 95 (11), 866–871. doi:10.1111/j.1349-7006.2004.tb02195.x