

Benchmarking BRCA2 Missense Variant Pathogenicity Predictors Across Functional Domains and in a Subset of Variants Observed in Ashkenazi Jewish Individuals

Meirav Bocknek
Brooklyn Prospect Charter School
Brooklyn, NY, USA
meiravbocknek@gmail.com

Mariame Diabate, PhD
Stanford University
Stanford, CA, USA
diabatem@stanford.edu

Abstract

Pathogenic variants in *BRCA2* substantially increase hereditary cancer risk, yet many identified missense variants remain classified as variants of uncertain significance (VUS). This study evaluated SIFT, PolyPhen-2, and REVEL for *BRCA2* missense single-nucleotide variants using ClinVar classifications as the reference standard, with gene-wide, domain-specific, and Ashkenazi Jewish (AJ)-observed subset analyses. After missense-only and review-status filtering, 9,077 variants remained; 651 had definitive ClinVar classifications (96 Pathogenic/Likely pathogenic and 555 Benign/Likely benign). All three predictors returned scores for these 651 variants. Because the predictors can return intermediate calls, primary sensitivity and specificity were calculated among confident binary predictions and intermediate-call rates were reported separately. Gene-wide, REVEL had the highest accuracy (92.8%) and specificity (96.0%), whereas SIFT had the highest sensitivity (98.9%) but low specificity (48.7%); PolyPhen-2 showed 81.8% accuracy, 88.2% sensitivity, and 80.5% specificity. In the OB Folds, REVEL showed the best balanced binary-call performance (90.1% accuracy, 90.9% sensitivity, 89.7% specificity), although 93 of 194 variants received intermediate REVEL calls. The BRC Repeats contained no definitively pathogenic missense variants, so sensitivity was not estimable. Using an exact genomic merge and $AF_{ASJ} > 0$, 35 missense variants were observed in the AJ subset (27 Benign and 8 VUS/conflicting; no definitively Pathogenic variants). Consequently, AJ sensitivity was not estimable; specificity among confident benign calls was 100% for REVEL (95% CI 85.7%–100%), 100% for PolyPhen-2 (81.6%–100%), and 42.3% for SIFT (25.5%–61.1%). These results support context-aware use of in-silico predictions as supporting evidence, while emphasizing predictor abstention, domain-level uncertainty, and the limits of population-subset inference.

Keywords

BRCA2, in-silico predictors, variant classification, ClinVar, SIFT, PolyPhen-2, REVEL, Ashkenazi Jewish population, variants of uncertain significance

ACM Reference Format:

Meirav Bocknek and Mariame Diabate, PhD. 2026. Benchmarking BRCA2 Missense Variant Pathogenicity Predictors Across Functional Domains and

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

IJSCAR VOL. 3, ISSUE 4

© 2026 The Author(s). Published by IJSCAR.

in a Subset of Variants Observed in Ashkenazi Jewish Individuals. In *Proceedings of International Journal of Secondary Computing and Applications Research (IJSCAR VOL. 3, ISSUE 4)*. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.67149/yhjs2024.5/jbuwvpv61>

1 Introduction

BRCA2 is a tumor-suppressor gene that encodes a large protein forming connections with other repair factors and contributing to the repair of double-strand DNA breaks. Pathogenic variants in this gene substantially increase the risk of hereditary breast, ovarian, pancreatic, and prostate cancers; for example, female *BRCA2* carriers have an estimated lifetime breast cancer risk of 69% by age 80 [7]. Genetic testing is becoming more widespread because of the increasing availability of tests [4], and laboratories are identifying a large number of *BRCA2* variants, many of which are classified in ClinVar as variants of uncertain significance (VUS). This leaves many patients without clear guidance for treatment or protection planning. To help bridge this gap, *in-silico* predictors use computational approaches to predict whether an amino acid substitution is likely to disrupt protein function. While these predictions do not constitute standalone evidence, they contribute useful supporting information in clinical guidelines [9]. Accurate interpretation of *BRCA2* missense variants remains important because many detected variants cannot be classified confidently using currently available clinical evidence. Computational predictions can provide supporting evidence during variant interpretation, but their usefulness depends on how reliably they distinguish benign from pathogenic variation within the specific gene and functional context being evaluated.

Although this study examines variants observed in individuals of Ashkenazi Jewish ancestry, the predictors themselves are not population-conditioned: SIFT, PolyPhen-2, and REVEL score a given amino-acid substitution the same way regardless of who carries it. The AJ analysis therefore evaluates performance on the subset of *BRCA2* missense variants with a positive recorded AJ allele frequency, rather than ancestry-specific predictor behavior.

Prior benchmarking studies have established the general performance of in-silico predictors across large variant databases, but *BRCA2* performance can vary with local sequence constraint and functional context [6]. This study therefore asks two related but distinct questions: whether gene-wide predictor performance is preserved within defined *BRCA2* functional domains, and how the same predictors behave on missense variants documented in AJ individuals. The latter analysis is a variant-subset analysis, not a test of population-conditioned algorithms. These distinctions are clinically relevant because computational predictions are used only

as supporting evidence within broader variant-classification frameworks [9].

We hypothesized that gene-wide performance would not necessarily be preserved within all BRCA2 functional domains because the consequences of missense variation differ across regions with distinct structural and evolutionary constraints. We also evaluated whether gene-wide specificity was retained among variants observed in AJ individuals, while recognizing that population membership does not alter the predictor score itself. The primary goal was to compare SIFT, PolyPhen-2, and REVEL on definitively classified BRCA2 missense variants and to identify contexts in which predictor performance or predictor abstention limits interpretation.

2 Method

BRCA2 variant data were obtained from ClinVar and downloaded on October 14, 2025. The source file contained 10,019 BRCA2 records. The analytical cohort was restricted to germline single-nucleotide variants whose molecular consequence included missense variation and whose ClinVar germline review status corresponded to at least one review star. Nonsense and stop-gain variants were excluded before predictor benchmarking because the three evaluated tools are missense predictors. This produced 9,077 missense SNVs. ClinVar classifications were grouped as Pathogenic (Pathogenic and Likely pathogenic), Benign (Benign and Likely benign), or VUS (Uncertain significance and Conflicting classifications of pathogenicity). The final missense cohort contained 96 Pathogenic, 555 Benign, and 8,426 VUS/conflicting variants; the 651 Pathogenic/Benign variants formed the reference set for performance evaluation.

Canonical BRCA2 amino-acid positions were parsed from the NM_000059.4 transcript annotation in the ClinVar Name field. This avoided ambiguity in the comma-separated Protein change field, which can contain multiple transcript-specific protein consequences. All reported domain assignments and amino-acid identifiers therefore use the canonical BRCA2 transcript.

A binary truth label was assigned for definitive ClinVar classifications (Benign = 0; Pathogenic = 1). REVEL and SIFT scores were retrieved through the MyVariant.info API using GRCh38 genomic HGVS notation and dbNSFP annotations. PolyPhen-2 HumVar scores were retrieved separately through the Ensembl Variant Effect Predictor (VEP) REST API. For PolyPhen-2, the BRCA2 Ensembl transcript ENST00000380152 was explicitly selected; all 9,077 missense variants received PolyPhen-2 scores for this transcript. REVEL and SIFT scores were also available for all 9,077 variants in the corrected analysis.

Each predictor score was classified into three tiers (Benign, intermediate/VUS, Pathogenic) using prespecified literature-supported thresholds. PolyPhen-2 HumVar scores ≤ 0.30 were classified as Benign, ≥ 0.70 as Pathogenic, and intermediate values as VUS [1]. SIFT scores ≤ 0.05 were classified as Pathogenic, > 0.10 as Benign, and intermediate values as VUS [8]; lower SIFT scores indicate greater predicted damage. REVEL scores < 0.40 were classified as Benign, ≥ 0.75 as Pathogenic, and intermediate values as VUS [5].

2.1 Ashkenazi Jewish Population Filtering

AJ-observed variants were identified by merging ClinVar and BRCA Exchange [3] using the exact GRCh38 genomic SNV (chromosome, position, reference allele, and alternate allele), rather than protein-change text. A variant was included in the AJ-observed subset only when the recorded Ashkenazi Jewish allele frequency was strictly greater than zero ($AF_{AJ} > 0$). This criterion identified 35 missense variants: 27 Benign and 8 VUS/conflicting by ClinVar, with no definitively Pathogenic missense variants. Because there were no Pathogenic reference variants in this subset, sensitivity could not be estimated.

Because predictor scores are variant-specific rather than population conditioned, the AJ analysis should be interpreted as performance on variants observed in AJ individuals, not as evidence that predictor behavior changes by ancestry.

2.2 Domain-Specific Filtering

BRCA2 functional domains were defined from established structural and functional annotations [2] using the following amino-acid boundaries: PALB2-binding region (10–40), BRC Repeats (1002–2113), Helical Domain (2481–2667), OB Folds (2670–3190), Tower Domain (3196–3232), and C-terminal BRCT-like region (3265–3418). Variants outside these intervals were categorized as non-domain. Domain truth counts were calculated before predictor-specific filtering. Comparative predictor interpretation focused on the BRC Repeats, Helical Domain, and OB Folds because the PALB2-binding, Tower, and C-terminal regions contained very small definitive reference sets.

2.3 Performance Metrics

Predictor performance was evaluated using confusion-matrix counts and standard classification metrics. Accuracy was $(TP + TN)/N$, sensitivity was $TP/(TP + FN)$, and specificity was $TN/(TN + FP)$. Sensitivity was reported as not estimable when a subset contained no Pathogenic reference variants. Because each tool can return an intermediate/VUS prediction, every predictor was evaluated with three denominators: the number of definitive ClinVar variants with a score (N scored), the number receiving a confident binary Benign/Pathogenic call (N confident), and the number receiving an intermediate call. Primary accuracy, sensitivity, and specificity were calculated among confident binary calls, and Wilson 95% confidence intervals were reported. As a sensitivity analysis addressing predictor abstention, intermediate calls were counted as unresolved/incorrect when calculating all-case accuracy. Pathogenic prevalence was calculated as $P/(P + B)$, whereas prior odds were calculated as P/B ; these quantities were not treated as interchangeable.

3 Results

3.1 Gene-wide Predictor Performance

The corrected definitive missense reference set contained 96 Pathogenic and 555 Benign variants ($n = 651$). The pathogenic prevalence was therefore $96/651 = 14.75\%$, and the corresponding prior odds were $96/555 = 0.173$. These quantities describe the composition of

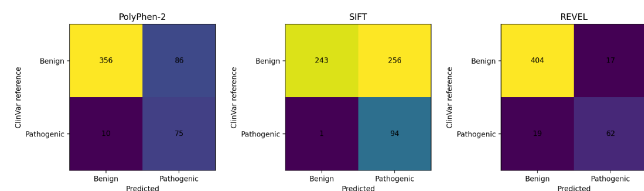


Figure 1: Gene-wide confusion matrices for confident binary predictions. Counts are shown separately for PolyPhen-2 (n = 527), SIFT (n = 594), and REVEL (n = 502). Intermediate predictor calls are not included in these matrices and are reported separately.

the definitive ClinVar missense set; they do not impose a maximum possible predictor odds ratio and should not be assumed to represent the prior probability among VUS.

All three predictors produced scores for all 651 definitive variants, but their intermediate-call rates differed. PolyPhen-2 made 527 confident binary calls and 124 intermediate calls; SIFT made 594 confident calls and 57 intermediate calls; REVEL made 502 confident calls and 149 intermediate calls. Accordingly, the confusion matrices and primary binary-call metrics use predictor-specific confident denominators rather than a common n.

Among confident binary calls, REVEL achieved the highest gene-wide accuracy at 92.8% (95% CI 90.2%–94.8%) and specificity at 96.0% (93.6%–97.5%), with sensitivity of 76.5% (66.2%–84.4%). SIFT achieved the highest sensitivity at 98.9% (94.3%–99.8%) but had substantially lower specificity at 48.7% (44.3%–53.1%) and accuracy at 56.7% (52.7%–60.7%). PolyPhen-2 showed intermediate performance, with 81.8% accuracy (78.3%–84.8%), 88.2% sensitivity (79.7%–93.5%), and 80.5% specificity (76.6%–84.0%).

The corrected results preserve the main sensitivity–specificity tradeoff, but the frequency of intermediate/VUS predictions differs across predictors and should be considered when interpreting their performance. SIFT made the fewest intermediate calls (57/651) and identified 94 of 96 Pathogenic variants as Pathogenic, whereas REVEL made 149 intermediate calls and classified 62 of 96 Pathogenic variants as Pathogenic. When intermediate calls were counted as unresolved/incorrect, all-case accuracy was 66.2% for PolyPhen-2, 51.8% for SIFT, and 71.6% for REVEL. Thus, REVEL’s high binary-call accuracy and specificity should be interpreted together with its lower sensitivity and higher intermediate-call rate.

3.2 Domain-Specific Predictor Performance

To evaluate domain-specific performance, the 651 definitive missense variants were mapped to annotated BRCA2 functional regions before predictor filtering. Of these, 462 variants fell within the six predefined domains and 189 fell outside the annotated intervals. The corrected domain composition is shown in Table 1. Pathogenic missense prevalence was highly uneven across regions, including no definitively Pathogenic missense variants in the BRC Repeats, Tower Domain, or C-terminal region.

The BRC Repeats contained 159 definitive missense variants, all Benign. The Helical Domain contained 33 Pathogenic and 38 Benign variants (n = 71), and the OB Folds contained 45 Pathogenic

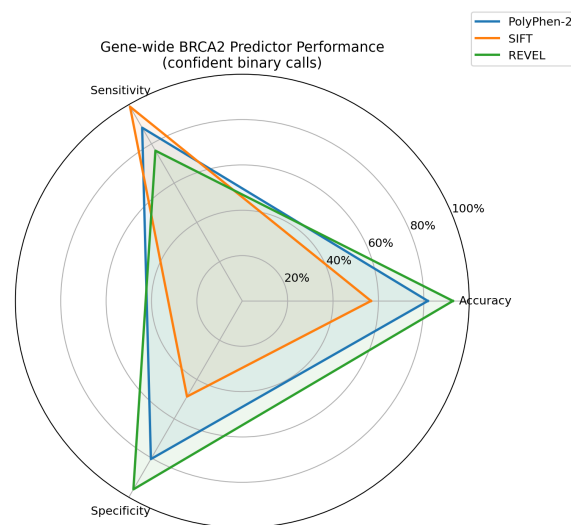


Figure 2: Radar plot of gene-wide accuracy, sensitivity, and specificity among confident binary calls for PolyPhen-2, SIFT, and REVEL.

and 149 Benign variants (n = 194). These pre-predictor counts are distinct from the smaller predictor-specific confident denominators used for performance metrics.

Comparative predictor interpretation focused on the BRC Repeats, Helical Domain, and OB Folds. The PALB2-binding region (n = 6), Tower Domain (n = 8), and C-terminal region (n = 24) were retained in the descriptive domain table but were considered too small for stable comparative performance claims.

In the BRC Repeats, all 159 definitive missense variants were Benign; therefore sensitivity was not estimable for any predictor. Among confident calls, specificity was 96.6% for REVEL (95% CI 92.3%–98.5%), 90.4% for PolyPhen-2 (84.3%–94.3%), and 62.1% for SIFT (53.9%–69.8%). The absence of definitively Pathogenic missense variants is a robust descriptive finding, but it does not by itself establish that pathogenic disruption of this region is predominantly truncating; unresolved missense variants may remain in the VUS category.

In the Helical Domain, REVEL had 90.5% accuracy, 100% sensitivity, and 60.0% specificity among 42 confident calls, but 29 of 71 variants received intermediate REVEL calls. PolyPhen-2 had 71.0% accuracy, 100% sensitivity, and 37.9% specificity among 62 confident calls, while SIFT had 61.8% accuracy, 100% sensitivity, and 25.7% specificity among 68 confident calls. In the OB Folds, REVEL showed the best balanced binary-call performance: 90.1% accuracy (82.7%–94.5%), 90.9% sensitivity (76.4%–96.9%), and 89.7% specificity (80.2%–94.9%) among 101 confident calls. PolyPhen-2 had 69.8% accuracy, 95.0% sensitivity, and 60.6% specificity among 149 confident calls; SIFT had 44.0% accuracy, 100% sensitivity, and 25.5% specificity among 182 confident calls. REVEL’s OB Fold performance should nevertheless be interpreted alongside its 93 intermediate calls in this domain.

Overall, domain-level results reinforce the need to report both discrimination and abstention. REVEL generally had the highest

Table 1: BRCA2 functional domains, amino acid boundaries, and ClinVar classification counts.

Domain	AA Range	Function	Pathogenic	Benign	Total	Pathogenic Prevalence
PALB2-binding	10–40	Protein interaction, N-terminus	3	3	6	0.5000
BRC Repeats	1002–2113	RAD51 binding, DNA repair	0	159	159	0.0000
Helical Domain	2481–2667	Structural / DNA-binding domain	33	38	71	0.4648
OB Folds	2670–3190	ssDNA binding, DNA-binding domain	45	149	194	0.2320
Tower Domain	3196–3232	DNA binding	0	8	8	0.0000
C-terminal	3265–3418	C-terminus	0	24	24	0.0000

Table 2: Domain-specific predictor performance. N truth is the number of definitive ClinVar variants in the domain; N confident is the number receiving a binary Benign/Pathogenic prediction. The difference between these values is the number of intermediate predictor calls. Metrics are reported with Wilson 95% confidence intervals.

Domain	Predictor	N truth	N conf.	TP	FP	FN	TN	Acc. (95% CI)	Sens. (95% CI)	Spec. (95% CI)
BRC Repeats	PolyPhen-2	159	136	0	13	0	123	90.4% (84.3–94.3)	Not est.	90.4% (84.3–94.3)
BRC Repeats	REVEL	159	147	0	5	0	142	96.6% (92.3–98.5)	Not est.	96.6% (92.3–98.5)
BRC Repeats	SIFT	159	140	0	53	0	87	62.1% (53.9–69.8)	Not est.	62.1% (53.9–69.8)
Helical Domain	PolyPhen-2	71	62	33	18	0	11	71.0% (58.7–80.8)	100% (89.6–100)	37.9% (22.7–56.0)
Helical Domain	REVEL	71	42	32	4	0	6	90.5% (77.9–96.2)	100% (89.3–100)	60.0% (31.3–83.2)
Helical Domain	SIFT	71	68	33	26	0	9	61.8% (49.9–72.4)	100% (89.6–100)	25.7% (14.2–42.1)
OB Folds	PolyPhen-2	194	149	38	43	2	66	69.8% (62.0–76.6)	95.0% (83.5–98.6)	60.6% (51.2–69.2)
OB Folds	REVEL	194	101	30	7	3	61	90.1% (82.7–94.5)	90.9% (76.4–96.9)	89.7% (80.2–94.9)
OB Folds	SIFT	194	182	45	102	0	35	44.0% (36.9–51.2)	100% (92.1–100)	25.5% (19.0–33.4)

specificity and balanced binary-call accuracy in the BRC Repeats, Helical Domain, and OB Folds, whereas SIFT prioritized sensitivity at the cost of specificity. However, REVEL also generated substantially more intermediate calls in the Helical Domain and OB Folds, limiting direct comparison based on binary-call metrics alone.

3.3 Ashkenazi Jewish Population Subset

Using the corrected definition of AJ-observed variation (exact genomic match and $AF_{ASJ} > 0$), 35 BRCA2 missense variants were identified. Of these, 27 were definitively Benign and 8 were VUS/conflicting; none were definitively Pathogenic. Consequently, this subset can inform specificity and false-positive behavior among known Benign variants, but it cannot estimate sensitivity for Pathogenic missense variants.

Among confident AJ-observed Benign calls, PolyPhen-2 correctly classified all 17/17 variants as Benign (specificity 100%, 95% CI 81.6%–100%) but returned 10 intermediate calls. REVEL correctly classified all 23/23 confident variants as Benign (specificity 100%, 85.7%–100%) and returned 4 intermediate calls. SIFT classified 11/26 confident variants as Benign and 15 as Pathogenic, yielding specificity of 42.3% (25.5%–61.1%); one additional variant received an intermediate SIFT call. Because the AJ reference set contains only Benign definitive variants, accuracy is numerically identical to specificity and sensitivity is not estimable. The wide confidence intervals emphasize the small sample size.

Statistical Note: Sensitivity and specificity are conditional on true variant status, whereas PPV and NPV depend on the prior probability of pathogenicity. The pathogenic prevalence of the definitive ClinVar missense set (14.75%) should not automatically be used as the prior probability for ClinVar VUS, because definitively

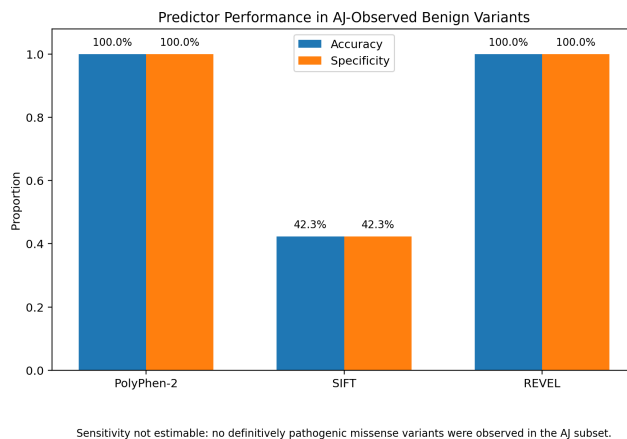


Figure 3: Corrected predictor performance among AJ-observed definitively Benign missense variants. Accuracy and specificity are shown; sensitivity is not estimable because no definitively Pathogenic missense variants were present.

classified variants and VUS are not exchangeable populations. Accordingly, this study does not assign a single posterior probability of pathogenicity or benignity to VUS.

3.4 Application to Variants of Uncertain Significance

A total of 8,426 missense variants in the corrected cohort were classified as VUS/conflicting by ClinVar. The purpose of applying the three predictors to these variants was descriptive: to characterize

Table 3: Predictor performance among the 27 definitively Benign AJ-observed missense variants. Sensitivity is not estimable because the subset contains no definitively Pathogenic variants.

Predictor	N truth	N conf.	Interm.	TN	FP	Specificity (95% CI)
PolyPhen-2	27	17	10	17	0	100% (81.6–100)
SIFT	27	26	1	11	15	42.3% (25.5–61.1)
REVEL	27	23	4	23	0	100% (85.7–100)

how often each tool produced Pathogenic, Benign, or intermediate calls under the prespecified thresholds. These call proportions are not measures of clinical accuracy because the true status of the VUS is unknown.

Note: The proportion called Pathogenic is a predictor call proportion, not a posterior probability of pathogenicity.

REVEL classified 289 of 8,426 VUS/conflicting variants as Pathogenic (3.43%), 6,303 as Benign, and 1,834 as intermediate. SIFT classified 4,336 as Pathogenic (51.46%), 3,158 as Benign, and 932 as intermediate. PolyPhen-2 classified 1,456 as Pathogenic (17.28%), 5,025 as Benign, and 1,945 as intermediate. These distributions are consistent with the gene-wide error profiles: SIFT makes Pathogenic calls much more frequently, while REVEL more often returns Benign or intermediate calls.

A Bayesian sensitivity analysis was performed across assumed pathogenicity priors from 1% to 50% using the observed gene-wide sensitivity and specificity. The resulting PPV and NPV changed substantially with the assumed prior, demonstrating why a single posterior probability should not be assigned to VUS without a justified VUS-specific prior. For example, at an illustrative 10% prior, estimated PPV/NPV were 67.8%/97.4% for REVEL, 33.5%/98.4% for PolyPhen-2, and 17.6%/99.8% for SIFT. These values are scenario estimates, not validated probabilities for individual VUS.

Therefore, predictor outputs should be used as supporting computational evidence rather than as stand-alone reclassification of a VUS. The corrected analysis does not support statements such as “94.8% confidence that an AJ VUS is benign” based on specificity alone.

4 Discussion

This study compared SIFT, PolyPhen-2, and REVEL on a corrected missense-only BRCA2 cohort and examined both domain-level performance and a subset of variants observed in AJ individuals. The revised analysis changes the interpretation in two important ways: predictor-specific intermediate calls must be considered alongside binary-call performance, and the AJ-observed subset contains no definitively Pathogenic missense variants, preventing sensitivity estimation in that subset.

Gene-wide, REVEL provided the strongest balance of accuracy and specificity among confident binary calls (92.8% accuracy; 96.0% specificity), but its sensitivity was 76.5% and it returned the largest number of intermediate calls (149/651). SIFT had the highest sensitivity (98.9%) and the fewest intermediate calls, but specificity was only 48.7%, reflecting frequent false Pathogenic calls among Benign variants. PolyPhen-2 occupied an intermediate position, with 81.8% accuracy, 88.2% sensitivity, and 80.5% specificity. These results argue against describing any single metric as overall predictor reliability

without also reporting abstention and the sensitivity-specificity tradeoff.

4.1 Clinical Implications for High-Risk Populations

Ashkenazi Jewish individuals have an elevated burden of BRCA1/2 founder variants [10], but the common founder alleles are predominantly truncating and fall outside the missense-only predictors evaluated here. The present AJ analysis therefore addresses only missense variants with a positive recorded AJ allele frequency and should not be interpreted as a comprehensive assessment of BRCA2 risk prediction in this population.

The corrected AJ-observed subset contained 27 definitively Benign missense variants and no definitively Pathogenic missense variants. REVEL and PolyPhen-2 made no false Pathogenic calls among their confident predictions, whereas SIFT misclassified 15 of 26 confident Benign variants as Pathogenic. However, the sample is small, confidence intervals are wide, and sensitivity cannot be estimated. These results support caution about SIFT false-positive behavior in this subset but do not establish AJ-specific superiority or clinical actionability for any predictor.

The absence of definitively Pathogenic missense variants in the corrected AJ-observed subset is consistent with the fact that major AJ BRCA2 founder variants, including 6174delT, are truncating rather than missense. This database composition limits the ability of the present study to evaluate pathogenic-missense detection in AJ individuals and motivates larger ancestry-linked datasets with variant-level phenotypic and functional evidence.

For an AJ individual with a BRCA2 missense VUS, these computational predictions may contribute supporting evidence, but they should not be translated directly into patient-level probabilities or management recommendations. Clinical interpretation requires integration with segregation, functional, population-frequency, and other evidence under established variant-classification frameworks.

All reported performance metrics in this study apply to missense single-nucleotide variants only; truncating founder variants and other variant classes were not benchmarked.

4.2 Domain-Specific Considerations

The corrected domain analysis confirms that pathogenic missense variation is unevenly distributed across BRCA2. No definitively Pathogenic missense variants were present in the BRC Repeats, Tower Domain, or C-terminal region, whereas the Helical Domain contained 33 and the OB Folds contained 45. The BRC-repeat finding is biologically interesting, but the absence of definitive pathogenic missense classifications should not be equated with proof that disruptive variation in this region is exclusively truncating.

Table 4: Predictor call distribution among 8,426 ClinVar VUS/conflicting missense variants.

Predictor	Pathogenic call	Benign call	Intermediate call	Proportion Pathogenic
REVEL	289	6303	1834	0.0343
SIFT	4336	3158	932	0.5146
PolyPhen-2	1456	5025	1945	0.1728

In the Helical Domain and OB Folds, REVEL showed the strongest balanced binary-call performance, particularly in the OB Folds, where accuracy was 90.1%, sensitivity 90.9%, and specificity 89.7%. One possible explanation is that REVEL integrates multiple sources of evidence, including conservation metrics, protein-level functional features, and the outputs of other prediction methods, whereas SIFT relies primarily on evolutionary conservation. Structural and functional regions such as the Helical Domain and OB Folds may therefore benefit from the broader feature set incorporated by REVEL. SIFT retained very high sensitivity but poor specificity, while PolyPhen-2 showed intermediate performance. At the same time, REVEL returned 29 intermediate calls in the Helical Domain and 93 in the OB Folds, so its strong binary-call metrics apply to a more selectively classified subset. This abstention behavior is an important part of domain-specific performance.

These domain-level findings support a context-aware interpretation rather than a simple ranking of predictors. For regions with no Pathogenic reference variants, sensitivity is not estimable. For regions with both classes represented, specificity, sensitivity, confidence intervals, and intermediate-call rates should all be considered before using a computational prediction as supporting evidence.

4.3 Limitations

Several limitations should be considered. First, ClinVar classifications were used as the reference standard, but these classifications are not fully independent of the predictors being benchmarked. SIFT, PolyPhen-2, and REVEL may contribute PP3/BP4 computational evidence to ClinVar submissions, creating potential circularity. The present design cannot quantify how much of the observed agreement, including REVEL’s apparent advantage, reflects this dependence.

Second, the corrected AJ-observed subset contained only 35 missense variants, including 27 definitively Benign variants and no definitively Pathogenic variants. Sensitivity therefore cannot be estimated for this subset, and specificity estimates have wide confidence intervals. The AJ analysis should be viewed as descriptive rather than population-specific validation.

Third, classification thresholds affect both binary-call performance and the number of intermediate calls. The thresholds used here were prespecified from the literature, but different thresholds would alter sensitivity, specificity, and abstention. Gene-specific threshold optimization was beyond the scope of this study.

Fourth, the analysis was restricted to missense single-nucleotide variants. Frameshift, nonsense, splice-altering, and other variant classes were not evaluated; this is particularly important for AJ founder variation. Several domains also contained very small definitive reference sets, limiting stable performance estimation.

Fifth, the AJ criterion identifies variants with a positive recorded AJ allele frequency but does not establish enrichment, ancestry specificity, or causal association with AJ cancer risk.

Finally, performance on definitively classified variants may not generalize to VUS. Intermediate predictor calls were common, especially for REVEL and PolyPhen-2, and the true pathogenicity prevalence among ClinVar VUS is unknown. Accordingly, the Bayesian calculations are sensitivity analyses over assumed priors rather than calibrated posterior probabilities for individual VUS.

5 Conclusion

This study evaluated SIFT, PolyPhen-2, and REVEL on 9,077 BRCA2 missense SNVs, including 651 definitively classified variants, with gene-wide, domain-specific, and AJ-observed subset analyses. Among confident gene-wide binary calls, REVEL had the highest accuracy and specificity, SIFT had the highest sensitivity but substantially lower specificity, and PolyPhen-2 showed intermediate performance. Domain analyses showed that performance and abstention vary by functional region, while sensitivity was not estimable in regions containing no definitively Pathogenic missense variants.

The corrected AJ-observed subset contained no definitively Pathogenic missense variants, so the study cannot support claims about AJ-specific sensitivity or patient-level posterior probabilities. Overall, the findings support using in-silico predictors as context dependent supporting evidence rather than stand-alone classifiers. Future work should use independent functional reference sets, larger ancestry-linked cohorts, and explicit evaluation of intermediate-call behavior to improve BRCA2 missense interpretation.

6 Reproducibility

The full analysis work, including all data-processing scripts, filtering steps, and code used to generate tables and figures, is available in a publicly accessible Google Colab notebook (https://colab.research.google.com/drive/1MPReq_riDLVYMWvcCOImU_kseKJhpy41?usp=sharing).

This notebook contains the complete reproducible pipeline and instructions for regenerating all results presented in the manuscript.

References

- [1] Ivan A. Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. 2010. A Method and Server for Predicting Damaging Missense Mutations. *Nature Methods* 7, 4 (2010), 248–249. <https://doi.org/10.1038/nmeth0410-248>
- [2] Kajal Biswas et al. 2012. Functional Evaluation of BRCA2 Variants Mapping to the PALB2-Binding and C-Terminal DNA-Binding Domains Using a Mouse ES Cell-Based Assay. *Human Molecular Genetics* 21, 18 (2012), 3993–4006. <https://doi.org/10.1093/hmg/dds222>

- [3] Melissa S. Cline et al. 2018. BRCA Challenge: BRCA Exchange as a Global Resource for Variants in BRCA1 and BRCA2. *PLoS Genetics* 14, 12 (2018), e1007752. <https://doi.org/10.1371/journal.pgen.1007752>
- [4] Alyssa L. Halbisen and Christine Y. Lu. 2023. Trends in Availability of Genetic Tests in the United States, 2012–2022. *Journal of Personalized Medicine* 13, 4 (2023), 638. <https://doi.org/10.3390/jpm13040638>
- [5] Nilah M. Ioannidis, Joseph H. Rothstein, Vikas Pejaver, Sumit Middha, Shannon K. McDonnell, Saurabh Baheti, et al. 2016. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics* 99, 4 (2016), 877–885. <https://doi.org/10.1016/j.ajhg.2016.08.016>
- [6] Mohannad N. Khandakji and Borbala Mifsud. 2022. Gene-Specific Machine Learning Model to Predict the Pathogenicity of BRCA2 Variants. *Frontiers in Genetics* 13 (2022), 982930. <https://doi.org/10.3389/fgene.2022.982930>
- [7] Karoline B. Kuchenbaecker, John L. Hopper, Daniel R. Barnes, Kelly-Anne Phillips, Thea M. Mooij, Marie-José Roos-Blom, Sarah Jarvis, Flora E. van Leeuwen, Roger L. Milne, Nadine Andrieu, David E. Goldgar, Mary Beth Terry, Matti A. Rookus, Douglas F. Easton, Antonis C. Antoniou, Timothy R. Rebbeck, et al. 2017. Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA* 317, 23 (2017), 2402–2416. <https://doi.org/10.1001/jama.2017.7112>
- [8] Pauline C. Ng and Steven Henikoff. 2003. SIFT: Predicting Amino Acid Changes That Affect Protein Function. *Nucleic Acids Research* 31, 13 (2003), 3812–3814. <https://doi.org/10.1093/nar/gkg509>
- [9] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, and Heidi L. Rehm. 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. *Genetics in Medicine* 17, 5 (2015), 405–424. <https://doi.org/10.1038/gim.2015.30>
- [10] Jaya M. Satagopan, Kenneth Offit, William Foulkes, Mark E. Robson, Sholom Wacholder, Christine M. Eng, Stephen E. Karp, and Colin B. Begg. 2001. The Lifetime Risks of Breast Cancer in Ashkenazi Jewish Carriers of BRCA1 and BRCA2 Mutations. *Cancer Epidemiology, Biomarkers & Prevention* 10, 5 (2001), 467–473. PMID: 11352856.