

Predicting Pathogenicity: Benchmarking SIFT, REVEL and AlphaMissense on ClinVar Missense Variants

Zhang Enshuo

Dulwich College, UK

Sajjaad Hassan Kassim

Nanjing Medical University, China

[Doi:10.19044/esj.2026.v22n30p52](https://doi.org/10.19044/esj.2026.v22n30p52)

Submitted: 16 September 2026

Accepted: 13 October 2026

Published: 31 October 2026

Copyright 2026 Author(s)

Under Creative Commons CC-BY 4.0

OPEN ACCESS

Cite As:

Enshuo, Z., & Kassim, S.H. (2026). *Predicting Pathogenicity: Benchmarking SIFT, REVEL and AlphaMissense on ClinVar Missense Variants*. European Scientific Journal, ESJ, 22 (30), 52. <https://doi.org/10.19044/esj.2026.v22n30p52>

Abstract

This study compared SIFT, REVEL, and AlphaMissense using 5,000 ClinVar missense variants: an original sample of 1,000 variants and 4,000 additional variants. All variants were evaluated using the same fixed transcript-selection rule. The original sample yielded 752 complete cases, and the additional sample yielded 3,018. In the additional cases, ROC-AUC was 0.9035 for SIFT, 0.9794 for REVEL, and 0.9662 for AlphaMissense. The REVEL-AlphaMissense difference was small but clear at 0.0132 (95% paired bootstrap interval 0.0081-0.0187). Among 983 additional cases absent from a December 2023 ClinVar snapshot, the difference narrowed to 0.0045 (-0.0024 to 0.0115). Among selected-transcript missense variants in the expanded sample (n = 4,591), coverage was 97.84% for SIFT, 84.49% for REVEL, and 98.24% for AlphaMissense. A transcript audit found another missense transcript with a REVEL score for 109 of 155 original missing-score records. AlphaMissense placed 42 original complete cases in its ambiguous range. Both newer tools discriminated better than SIFT. REVEL showed a slight advantage in the additional sample, although this advantage was not evident in the historical snapshot-absent subset. These findings support the use of computational scores for variant prioritization, while emphasizing that clinical interpretation still requires calibrated evidence thresholds and other patient-level evidence.

Keywords: Computational Biology, AlphaMissense, ClinVar, precision medicine, computational genomics, patient-level evidence, SIFT, REVEL, Missense variant pathogenicity prediction

1. Introduction

A missense change may disrupt a protein without causing the disease seen in a patient. That gap between molecular effect and clinical explanation is why the ACMG/AMP framework combines computational predictions with population, functional, segregation, and other evidence (Richards et al., 2015). Prediction scores fill a practical need, especially when laboratory or family evidence is sparse. However, their limits are just as practical: a score can rank variants well and still be poorly suited to a clinical decision.

ClinVar provides a large set of submitted classifications (Landrum et al., 2025). However, it is also an imperfect test bed. Clearly classified pathogenic and benign records represent a selected part of human variation, while the uncertain variants that often prompt clinical testing are excluded from a binary benchmark.

The three tools studied here approach the problem differently. SIFT estimates whether sequence conservation permits an amino-acid substitution (Ng & Henikoff, 2003). REVEL combines 13 prediction signals, one of which is SIFT (Ioannidis et al., 2016). AlphaMissense learns from evolutionary, structural, and population information (Cheng et al., 2023). Their development data differ as well. REVEL used disease and rare neutral variants, whereas AlphaMissense was not trained directly on ClinVar labels; however, its released pathogenicity scores were calibrated using a ClinVar evaluation set and its categorical thresholds were selected using ClinVar variants. Earlier work has shown how this type of overlap can favor a predictor in retrospective benchmarks (Grimm et al., 2015). Other widely used approaches include integrative scores such as CADD (Kircher et al., 2014) and generative protein-sequence models such as EVE (Frazer et al., 2021). The present benchmark was prespecified to compare SIFT, REVEL, and AlphaMissense.

Therefore, the main question was straightforward: when all three tools score the same ClinVar variant, which one separates pathogenic from benign classifications most effectively? Two related questions emerged during the analysis. How much coverage does each annotation pipeline provide, and what happens to the comparison when a tool returns no score or an ambiguous call? We kept these questions together because a high AUC is less useful when many variants drop out before evaluation.

2. Methods

2.1 Data Source and Eligibility

The analysis used the ClinVar GRCh38 VCF dated 5 September 2026. It was downloaded on 13 September 2026 from the official NCBI ClinVar source, and the MD5 checksum matched the value supplied for that release. This confirmed that the analyzed file was the intended, uncorrupted release; it does not by itself establish the biological truth of each individual ClinVar classification. Eligible records had a missense_variant consequence and one of six aggregate germline classifications: Pathogenic, Likely pathogenic, Pathogenic/Likely pathogenic, Benign, Likely benign, or Benign/Likely benign. We grouped the first three as pathogenic or likely pathogenic (P/LP) and the last three as benign or likely benign (B/LB).

To reduce label uncertainty, each record was required to have at least one ClinVar review star and no aggregate conflict. Accepted review states were criteria provided by one submitter, criteria provided by multiple submitters without conflict, expert-panel review, or a practice guideline (National Center for Biotechnology Information, 2026). Records with CLNSIGCONF were removed. Non-SNV alleles, nonstandard chromosomes, and duplicate chromosome-position-reference-alternate keys were also excluded. These filters improve label consistency, but ClinVar classifications remain curated clinical assertions rather than experimental ground truth.

2.2 Sampling Decisions

A balanced 200-variant pilot served as a technical check. It showed that the annotation route worked, but the observed REST processing time implied roughly 20 hours for all 206,578 eligible variants. That estimate triggered the prespecified smaller-sample option in the original analysis plan. The original 1,000-input sample was therefore a resource-constrained design choice rather than the result of a conventional power calculation.

Sampling used a SHA-256 ordering of the fixed seed 20260913 and the genomic key. After removing the pilot, the first 500 variants in each class formed the original 1,000-input sample. A later-dated revision protocol extended the same fixed ordering to 2,500 variants per class, adding 4,000 inputs without changing the original sample. Statistical adequacy was assessed through the precision of the bootstrap confidence intervals and the stability of the main effect estimates across the original, additional, combined, and snapshot-absent analyses. The additional cohort yielded 3,018 complete cases, and the REVEL-AlphaMissense AUC difference remained similar in the additional and combined analyses (0.0132 and 0.0123, respectively), while confidence intervals narrowed with the larger sample.

2.3 Annotation and Transcript Choice

Variants were submitted to Ensembl VEP REST in batches of 100. The requests used GRCh38, VEP release 116, and REST release 15.12 and asked for SIFT, REVEL, AlphaMissense, MANE, canonical, protein, and PICK annotations (Ensembl, 2026b; McLaren et al., 2016). Every response was archived by variant with its request settings and timestamp.

The analysis selected one transcript before considering score availability. MANE Select was prioritized, followed by the VEP PICK transcript, using the order `mane_select`, `canonical`, `appris`, `tsl`, `biotype`, `ccds`, `rank`, and `length`. MANE provides a shared transcript reference for clinical genomics (Morales et al., 2022). The selected consequence was required to remain missense; otherwise, the variant was excluded from the main analysis.

For ROC analysis, SIFT was transformed to 1-SIFT so that larger values indicated a more pathogenic prediction for all three tools. The binary rules were fixed in advance: SIFT ≤ 0.05 (SIFT Developers, 2026) and REVEL ≥ 0.5 (Ioannidis et al., 2016). AlphaMissense classifies scores below 0.34 as benign-like and scores above 0.564 as pathogenic-like; the inclusive middle range is ambiguous (Ensembl, 2026a). Scores in this middle range remained available for ROC analysis but were excluded from AlphaMissense binary metrics.

2.4 Statistical Analysis

A complete case was a selected-transcript missense variant with all three numerical scores. The primary comparison used 2,000 paired, class-stratified bootstrap resamples with seed 20260913. The same resampled variant indices were used for all predictors within each paired comparison, preserving the paired AUC differences. Bootstrap resampling was used to quantify sampling uncertainty; it did not alter the underlying observations or point estimates. Reported 95% confidence intervals were percentile bootstrap intervals defined by the 2.5th and 97.5th percentiles of the relevant bootstrap distribution (Efron & Tibshirani, 1993). We interpreted paired differences by whether their 95% confidence intervals included zero rather than deriving p-values from the bootstrap distribution. Threshold results included sensitivity, specificity, and F1. We repeated the binary comparison on the AlphaMissense-confident subset so that all three tools were evaluated on the same records.

The original protocol also specified a sensitivity analysis limited to records with at least two review stars. Analyses added during revision used 5,000 paired, class-stratified bootstrap resamples for F1 differences with seed 20260915 and 2,000 paired, class-stratified bootstrap resamples for the expanded cohorts with seed 20260916. The date-coded seeds identified separate reproducible analysis stages and were not selected on the basis of the results. Where p-values are reported for paired binary correctness in the

supplementary analyses, they were calculated separately using a two-sided exact test of the discordant pairs. These revision analyses were distinct from the initial analysis plan.

The 4,000 additional inputs were evaluated separately because they had not contributed to the original results. We also report the combined 5,000-input sample for precision. A historical sensitivity analysis then removed any additional variant whose genomic key or ClinVar Variation ID appeared in the 26 December 2023 GRCh38 VCF (National Center for Biotechnology Information, 2023). The 26 December 2023 release was selected as a stable, reproducible historical reference close to the 2023 publication of AlphaMissense. Because ClinVar variants were used in AlphaMissense evaluation/calibration and in defining its categorical thresholds, this analysis was intended to reduce potential historical overlap and to test whether the main comparison persisted among variants absent from that archived dataset. We refer to the remaining records as snapshot-absent. The combined and restricted datasets are nested analyses within this study.

2.5 Audits Prompted by the First Results

The lower REVEL coverage than the other tools prompted a return to the original VEP responses. For each missing score, we checked whether another annotated missense transcript carried a value. The VEP plugin matching rules were then reviewed, and all records on chromosome 11 were compared with the authors' REVEL v1.3 archive (Ensembl, 2026c; Rothstein & Sieh, 2021). Chromosome 11 was chosen as a manageable diagnostic slice before inspecting the author scores; performance did not influence this choice. AlphaMissense posed a different problem. Because it makes no binary call within its ambiguous range, both error and coverage are affected. Therefore, the abstention interval was widened symmetrically around 0.452, the midpoint of the default AlphaMissense ambiguous interval (0.340–0.564), with the midpoint retained.

3. Results

3.1 Sample Construction

The source VCF contained 4,469,028 records. The missense consequence filter retained 2,537,042. The accepted P/LP or B/LB labels retained 225,306. Review-status and conflict filters reduced the set to 207,842. Removing 1,255 non-SNV records and nine records on nonstandard chromosomes left 206,578 eligible variants.

In the original sample, the selected transcript remained missense for 926 of 1,000 variants, evenly divided between P/LP and B/LB. MANE Select supplied 925 of those transcripts; VEP PICK supplied one. Seventy-four variants therefore left the analysis at the transcript check. Another 174 lacked

at least one score, giving 752 complete cases: 380 P/LP and 372 B/LB variants from 560 genes.

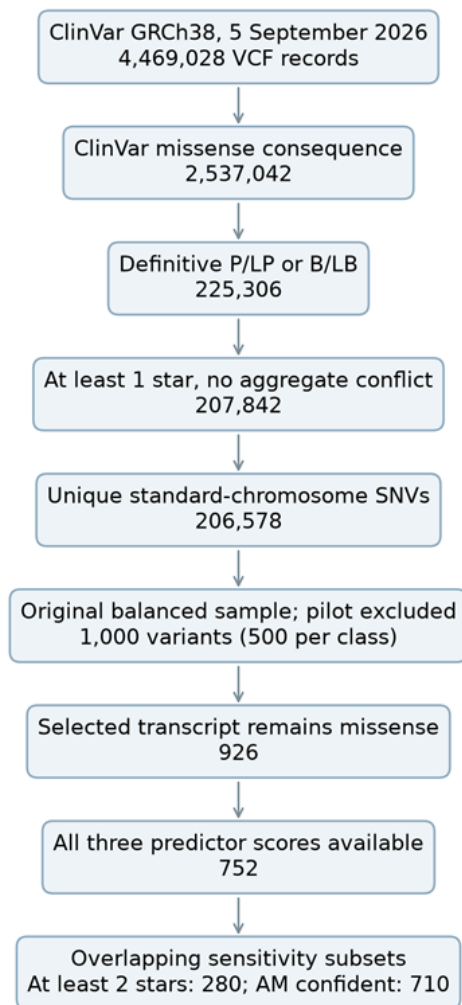


Figure 1

Table 1: Sampling and analysis populations

Stage	Original	Expanded including original
Eligible variants in release	206,578	Same pool
Technical pilot excluded	200	Same excluded pilot
Fixed balanced inputs	1,000	5,000
Selected-transcript missense	926	4,591
Three scores available	752	3,770
P/LP in complete cases	380	1,941
B/LB in complete cases	372	1,829

3.2 Score Coverage

SIFT scored 904 of the 926 original selected-missense variants (97.62%). REVEL scored 771 (83.26%), and AlphaMissense scored 912 (98.49%). In the expanded sample, 4,591 inputs remained missense on the selected transcript. SIFT scored 4,492 of these variants (97.84%), REVEL scored 3,879 (84.49%), and AlphaMissense scored 4,510 (98.24%). Relative to all 5,000 variants, the coverage would be 89.84%, 77.58%, and 90.20%, respectively.

AlphaMissense returned a number for most records, but a number was not always a binary decision. 54 of its 912 original scores were ambiguous. 42 of those occurred among the 752 complete cases, leaving 710 records for its default binary metrics. The two denominators answer different questions: 54/912 describes ambiguity among scored variants, whereas 42/752 describes its effect on the common comparison set.

Table 2: Numerical score coverage

Predictor and sample	Scored / missense	Missense %	All inputs %
SIFT	904/926	97.62	90.40
REVEL	771/926	83.26	77.10
AlphaMissense	912/926	98.49	91.20
SIFT (expanded)	4492/4591	97.84	89.84
REVEL (expanded)	3879/4591	84.49	77.58
AlphaMissense (expanded)	4510/4591	98.24	90.20

3.3 Discrimination

The first 752 complete cases gave AUCs of 0.8903 for SIFT, 0.9731 for REVEL, and 0.9639 for AlphaMissense. Both newer tools outperformed SIFT. REVEL was 0.0092 AUC higher than AlphaMissense. However, its 95% interval crossed zero (−0.0011 to 0.0207).

Table 3: Original complete-case discrimination

Predictor	AUC	95% paired-bootstrap CI
SIFT	0.8903	0.8667 to 0.9129
REVEL	0.9731	0.9623 to 0.9821
AlphaMissense	0.9639	0.9502 to 0.9759

Table 4: Complete case composition

Cohort	n	P/LP n (%)	B/LB n (%)
Additional	3018	1561 (51.72%)	1457 (48.28%)
Combined	3770	1941 (51.49%)	1829 (48.51%)
Snapshot absent	983	395 (40.18%)	588 (59.82%)

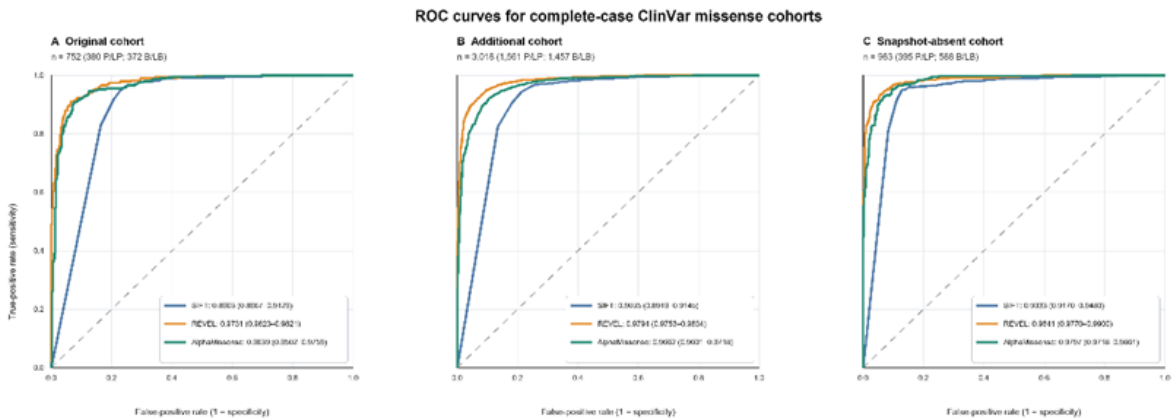


Figure 2: ROC curves for the original, additional, and snapshot-absent complete-case cohorts

Results were more precise in the 3,018 additional complete cases. This sample contained 1,561 P/LP and 1,457 B/LB variants. AUCs were 0.9035 for SIFT, 0.9794 for REVEL, and 0.9662 for AlphaMissense. REVEL exceeded AlphaMissense by 0.0132 (0.0081-0.0187). Combining the original and additional cases produced a similar positive difference of 0.0123 (0.0077-0.0171).

The snapshot-absent analysis contained 983 additional complete cases: 395 P/LP and 588 B/LB. Here, AUC was 0.9333 for SIFT, 0.9841 for REVEL, and 0.9797 for AlphaMissense. Their difference narrowed to 0.0045 (-0.0024 to 0.0115). Records with at least two review stars produced the same overall ordering. The snapshot-absent cohort was less balanced than the original and additional cohorts, with 59.82% B/LB variants. ROC-AUC is not determined by class prevalence alone; however, the historical filtering that produced this subset may also have altered the underlying case mix and class-conditional score distributions. Threshold-dependent measures such as F1 and overall accuracy are more directly affected by class composition.

Table 5: Discrimination with 95% intervals

Cohort	SIFT	REVEL	AlphaMissense
Additional	0.9035 (0.8919 to 0.9145)	0.9794 (0.9753 to 0.9834)	0.9662 (0.9601 to 0.9718)
Combined	0.9008 (0.8906 to 0.9100)	0.9781 (0.9740 to 0.9819)	0.9658 (0.9605 to 0.9709)
Snapshot absent	0.9333 (0.9170 to 0.9480)	0.9841 (0.9770 to 0.9900)	0.9797 (0.9718 to 0.9861)

3.4 Threshold Results

At the fixed thresholds, SIFT showed higher sensitivity. It detected 369 of 380 original P/LP variants (97.11%) but classified only 267 of 372 B/LB

variants correctly (71.77% specificity). REVEL traded some sensitivity for specificity: 91.05% and 93.01%, respectively. AlphaMissense reached 90.65% sensitivity and 94.68% specificity among the 710 cases outside its ambiguous range.

Their corresponding F1 scores were 0.8642, 0.9202, and 0.9249. The last value uses fewer cases, so we also compared the tools on the same 710 records. REVEL and AlphaMissense then had F1 scores of 0.9224 and 0.9249. In the 2,846 additional confident cases, the ordering reversed: 0.9310 for REVEL and 0.9167 for AlphaMissense. The paired difference was 0.0144 (0.0022–0.0262).

One further check chose a single cutoff for each tool in the original cases to reach at least 95% sensitivity, then left those cutoffs untouched in the additional cases. Sensitivity remained close to the target at 95.90% for SIFT, 95.64% for REVEL, and 95.13% for AlphaMissense. Specificity was 75.91%, 88.06%, and 83.60%.

Table 6: Original default-threshold classification and denominator-matched comparison

Predictor	n	Sens %	Spec %	F1	TP	FN	TN	FP
SIFT	752	97.11	71.77	0.8642	369	11	267	105
REVEL	752	91.05	93.01	0.9202	346	34	346	26
AlphaMissense	710	90.65	94.68	0.9249	320	33	338	19
SIFT (matched)	710	97.17	73.95	0.8695	343	10	264	93
REVEL (matched)	710	90.93	93.84	0.9224	321	32	335	22
AlphaMissense (matched)	710	90.65	94.68	0.9249	320	33	338	19

Note: P/LP was the positive class. Default analysis: SIFT and REVEL n = 752; AlphaMissense n = 710 after excluding 42 ambiguous calls. Matched analysis: all three predictors n = 710 (353 P/LP, 357 B/LB).

3.5 Missing Scores and Ambiguous Calls

The transcript audit explained much of REVEL's missingness: among 155 original records lacking a selected-transcript REVEL score, 109 had a value on another missense transcript (18 of 22 for SIFT; 0 of 14 for AlphaMissense). The chromosome 11 check illustrated the matching issue: all 43 scored controls agreed with the author archive, and six unscored records matched a genomic coordinate and amino acid substitution but not the selected transcript ID. This paper did not use these alternative scores in the benchmark, because doing so would have changed the transcript rule after seeing the missing data.

In the expanded sample, 712 REVEL scores were missing across 347 genes. The five genes with the most missing values contributed 121 records (16.99% of all missing values), indicating that missingness was spread across many genes rather than driven by a single locus.

The 42 AlphaMissense-ambiguous complete cases were also notable: SIFT misclassified 13 (30.95%) and REVEL 6 (14.29%), compared with error

rates of 14.51% and 7.61% among the 710 confident cases. As the AlphaMissense abstention interval widened, retained error generally decreased while coverage fell (Figure 3).

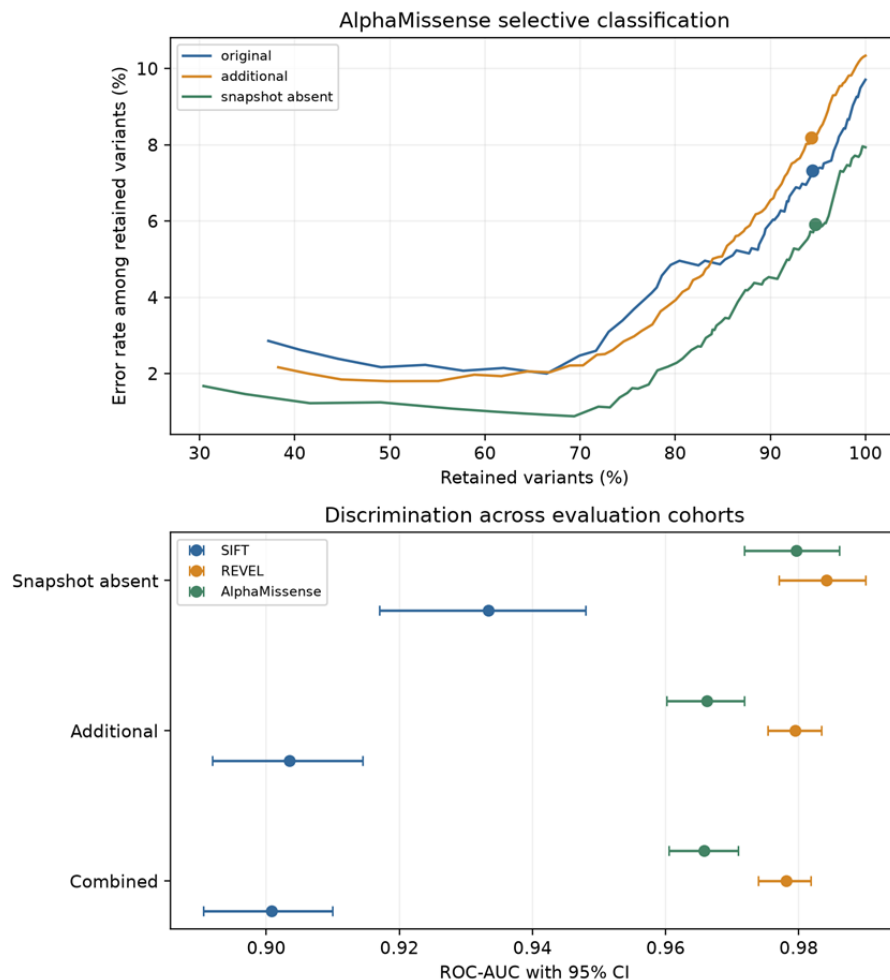


Figure 3

4. Discussion

REVEL and AlphaMissense both discriminated ClinVar pathogenic from benign missense variants better than SIFT. REVEL held a modest AUC advantage in the additional sample ($\Delta = 0.0132$), but the difference was small. In the snapshot-absent subset, the advantage shrank to 0.0045, and the confidence interval included zero. Historical data overlap and differences in case mix are both plausible explanations; distinguishing them will require an external dataset with documented development-set membership.

Coverage and selective classification were as important as discrimination. REVEL's lower coverage often reflected transcript matching

rather than absent predictor data: 109 of 155 original missing-score records had a REVEL value on another missense transcript, and chromosome 11 controls showed the same pattern. Using those alternative scores would have changed the transcript rule after data inspection, so the benchmark retained the original rule. AlphaMissense posed a different tradeoff: its ambiguous interval removed cases that were harder for SIFT and REVEL, which improved its binary metrics but reduced coverage. An unresolved variant still requires review, so coverage and accuracy should be reported together when comparing tools that abstain with those that always return a binary call.

AUC measures ranking, not clinical evidence strength. Clinicians need calibrated score ranges mapped to PP3/BP4 evidence levels, and ClinGen guidance recommends selecting the tool and threshold before inspecting results (Bergquist et al., 2025; Pejaver et al., 2022). We did not estimate those evidence strengths. REVEL and SIFT are also not independent, because SIFT contributes to the REVEL ensemble.

Several design features limit the scope of these findings. ClinVar provides curated labels rather than experimental truth. The balanced sample omits uncertain variants and does not reflect clinical prevalence. One selected transcript cannot represent all relevant isoforms. Missing scores and AlphaMissense abstention alter the evaluated set and the historical snapshot addresses only some sources of potential development-data overlap. Therefore, a stronger next study would begin with an external reference set whose labels and development-set membership are traceable, specify sample size, transcript rules, missing-score handling, and the primary comparison before scoring, and use functional assays where available, while recognizing that molecular readouts are not equivalent to clinical pathogenicity (Livesey & Marsh, 2025).

Conclusion

REVEL and AlphaMissense outperformed SIFT in this ClinVar missense benchmark. REVEL had a modest AUC advantage in the additional sample, whereas the historical snapshot-absent subset did not clearly separate the two newer tools. Performance should be interpreted jointly with score coverage and abstention: high discrimination among scored variants does not guarantee a result for every variant, and an ambiguous call may be preferable to a forced binary classification. Transcript matching can also influence practical score availability. These predictors are therefore most useful for variant prioritization and as calibrated components of a broader evidence framework, rather than as stand-alone clinical classifiers; patient-level interpretation still depends on additional clinical, population, functional, and segregation evidence.

Conflict of Interest: The authors reported no conflict of interest.

Data Availability: All data are included in the content of the paper.

Funding Statement: The authors did not obtain any funding for this research.

References:

1. Bergquist, T., Stenton, S. L., Nadeau, E. A. W., et al. (2025). Calibration of additional computational tools expands ClinGen recommendation options for variant classification with PP3/BP4 criteria. *Genetics in Medicine*, 27, 101402. <https://doi.org/10.1016/j.gim.2025.101402>
2. Cheng, J., Novati, G., Pan, J., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381, eadg7492. <https://doi.org/10.1126/science.adg7492>
3. Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.
4. Ensembl. (2026a). AlphaMissense VEP plugin release 116 [Plugin documentation and source code]. Retrieved September 13, 2026, from https://github.com/Ensembl/VEP_plugins/blob/release/116/AlphaMissense.pm
5. Ensembl. (2026b). POST vep species region [REST API documentation]. Retrieved September 13, 2026, from https://rest.ensembl.org/documentation/info/vep_region_post
6. Ensembl. (2026c). REVEL VEP plugin release 116 [Source code]. Retrieved September 13, 2026, from https://github.com/Ensembl/VEP_plugins/blob/release/116/REVEL.pm
7. Frazer, J., Notin, P., Dias, M., et al. (2021). Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599, 91–95. <https://doi.org/10.1038/s41586-021-04043-8>
8. Grimm, D. G., Azencott, C. A., Aicheler, F., et al. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation*, 36, 513–523. <https://doi.org/10.1002/humu.22768>
9. Ioannidis, N. M., Rothstein, J. H., Pejaver, V., et al. (2016). REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *American Journal of Human Genetics*, 99, 877–885. <https://doi.org/10.1016/j.ajhg.2016.08.016>
10. Kircher, M., Witten, D. M., Jain, P., et al. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46, 310–315. <https://doi.org/10.1038/ng.2892>

11. Landrum, M. J., Chitipiralla, S., Kaur, K., et al. (2025). ClinVar: Updates to support classifications of both germline and somatic variants. *Nucleic Acids Research*, 53, D1313–D1321. <https://doi.org/10.1093/nar/gkae1090>
12. Livesey, B. J., & Marsh, J. A. (2025). Variant effect predictor correlation with functional assays is reflective of clinical classification performance. *Genome Biology*, 26, 104. <https://doi.org/10.1186/s13059-025-03575-w>
13. McLaren, W., Gil, L., Hunt, S. E., et al. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17, 122. <https://doi.org/10.1186/s13059-016-0974-4>
14. Morales, J., Pujar, S., Loveland, J. E., et al. (2022). A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, 604, 310–315. <https://doi.org/10.1038/s41586-022-04558-8>
15. National Center for Biotechnology Information. (2023). ClinVar GRCh38 archived VCF dated 26 December 2023 [Data archive]. Retrieved September 13, 2026, from https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/archive_2.0/2023/clinvar_20231226.vcf.gz
16. National Center for Biotechnology Information. (2026). Review status in ClinVar. Retrieved September 13, 2026, from https://www.ncbi.nlm.nih.gov/clinvar/docs/review_status/
17. Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31, 3812–3814. <https://doi.org/10.1093/nar/gkg509>
18. Pejaver, V., Byrne, A. B., Feng, B. J., et al. (2022). Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *American Journal of Human Genetics*, 109, 2163–2177. <https://doi.org/10.1016/j.ajhg.2022.10.013>
19. Richards, S., Aziz, N., Bale, S., 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. *Genetics in Medicine*, 17, 405–424. <https://doi.org/10.1038/gim.2015.30>
20. Rothstein, J., & Sieh, W. (2021). REVEL rare exome variant ensemble learner scores (Version 1.3) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.7072866>
21. SIFT Developers. (2026). SIFT help. Retrieved September 13, 2026, from https://sift.bii.a-star.edu.sg/www/SIFT_help.html