Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Concordance of Automated ACMG Variant Classification with Expert-Curated Assertions: A Systematic Evaluation Using the ClinGen Evidence Repository

Muhammad Abiodun SULAIMAN
University of Ilorin https://orcid.org/0000-0001-9161-2608
Bolaji Fatai Oyeyemi
bolajioyeyemi@gmail.com
University of Ilorin https://orcid.org/0000-0001-5564-6165

Research Article

Keywords: variant classification, ACMG/AMP, concordance, ClinGen, evidence repository, pathogenicity prediction, REVEL, computational genomics

Posted Date: August 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10615801/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Background

Automated ACMG/AMP variant classification tools are increasingly deployed in clinical genomics, yet systematic evaluations against expert-curated gold standards remain sparse. The ClinGen Evidence Repository provides Expert Panel-adjudicated classifications at the highest evidentiary tier, serving as an ideal concordance benchmark.

Results

We evaluated VarTriage, a streaming ACMG classifier implementing 10 evidence criteria with ClinGen SVI-calibrated thresholds, against 15,334 Expert Panel-curated variants from the ClinGen Evidence Repository. Pathogenic sensitivity reached 68.9% (95% CI: 67.8–70.1%; PPV 80.5%), with binary concordance (pathogenic vs benign, VUS excluded) yielding Cohen's kappa of 0.98. REVEL prediction achieved AUC 0.920 on the scored subset (n = 8,094), with the ClinGen-calibrated moderate threshold (0.773) providing 85.2% sensitivity at 88.9% specificity. Ablation analysis identified REVEL as the dominant sensitivity contributor: removing computational predictions dropped sensitivity by 36.9 percentage points. Functional consequence strongly predicted expert classification (χ^2 = 9,117, $p < 10^{-300}$), with nonsense variants showing 160-fold pathogenic enrichment. In multi-tool comparison, VarTriage (68.9%) placed between InterVar (64.3%) and BIAS-2015 (74.0%) while employing fewer criteria (10 vs 18–19).

Conclusions

A minimal 10-criterion automated classifier achieves near-perfect binary discrimination (kappa = 0.98) and meaningful pathogenic sensitivity, with REVEL scores and the ClinGen SVI Bayesian combining rule as principal enablers. The accuracy gap concentrates in benign sensitivity (16.9% vs 80.2% for BIAS-2015), attributable to evidence types requiring non-computational data (functional assays, segregation, reputable sources). Targeted expansion of benign evidence criteria represents the highest-yield improvement path.

1 Background

Clinical whole-genome sequencing (WGS) generates approximately 4–5 million variant calls per individual (Telenti et al. 2016), and has become the preferred first-line diagnostic method for rare diseases in multiple healthcare systems (Bagger et al. 2024; Brlek et al. 2024). The diagnostic yield of WGS reaches 34–49% for pediatric rare disease (Nurchis et al. 2023; Kim et al. 2025), with trio WGS adding incremental yield for developmental delay (Uehara et al. 2024), but the clinical challenge lies not

in variant detection but interpretation: identifying pathogenic variants among millions of benign polymorphisms under reporting constraints that preclude manual review of every call (Bryen et al. 2023).

The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) established consensus guidelines in 2015 for sequence variant interpretation (Richards et al. 2015). These guidelines define 28 evidence criteria organized into strength tiers that combine through rule-based logic to produce five-tier classifications: Pathogenic, Likely Pathogenic, VUS, Likely Benign, and Benign. Despite this standardization, inter-laboratory concordance remains limited. Amendola et al. found that only 34% of variants received identical five-tier classifications across nine CSER laboratories in a 2016 pilot study (Amendola et al. 2016), improving to 71% after consensus discussion. A subsequent 2020 evaluation across nine eMERGE/CSER sites reported 54% complete five-category concordance (Amendola et al. 2020), with discordance driven primarily by differences in evidence access and criterion-strength interpretation. Similar concordance challenges have been documented in carrier screening (Kaseniit et al. 2019), somatic variant classification (Li et al. 2017; Maxwell et al. 2020), and ClinVar multi-submitter comparisons (Harrison et al. 2023).

The ClinGen Sequence Variant Interpretation (SVI) Working Group has progressively refined the ACMG/AMP framework through quantitative calibration. Tavtigian et al. demonstrated that the guidelines are fundamentally compatible with Bayesian reasoning (Tavtigian et al. 2018) and proposed a naturally scaled point system where evidence strength levels correspond to exponentially scaled odds of pathogenicity (Tavtigian et al. 2020). This quantitative recalibration has direct implications for automated classifiers: two moderate pathogenic criteria (combined odds approximately 18.5:1) exceed the posterior probability threshold for Likely Pathogenic, a combining rule not present in the original 2015 Table 5 but formally justified by the Bayesian framework.

Computational pathogenicity prediction has advanced substantially since 2015. REVEL, an ensemble of 13 individual tools, provides calibrated missense variant predictions (Ioannidis et al. 2016) that the ClinGen SVI Working Group validated for PP3/BP4 evidence at multiple strength levels (Pejaver et al. 2022). Early evaluations of in silico algorithms identified wide variability in individual tool concordance (Ghosh et al. 2017), motivating the development of ensemble metapredictors. The 2022 calibration established that REVEL $>$ 0.773 provides moderate pathogenic evidence and REVEL $<$ 0.183 provides moderate benign evidence (Pejaver et al. 2022), transforming PP3/BP4 from supporting-only to up to strong evidence strength. A 2025 expansion calibrated additional tools including AlphaMissense, ESM1b, and VARITY at comparable evidence levels (Bergquist et al. 2025). AlphaMissense, leveraging AlphaFold2's structural predictions, classified 89% of 71 million possible human missense variants (Cheng et al. 2023), while protein language models such as ESM1b enable genome-wide variant effect prediction from sequence alone (Brandes et al. 2023; Livesey and Marsh 2023). CADD v1.7 integrated ESM-1v protein language model features and regulatory CNN scores (Schubach et al. 2024), with CADD-Splice extending predictions to non-coding splice-altering variants (Rentzsch et al. 2021). SpliceAI uses deep residual networks to predict cryptic splicing from primary sequence (Jaganathan et al. 2019). The

ClinGen SVI Splicing Subgroup published standardized recommendations for applying splicing-related ACMG codes, including calibrated SpliceAI thresholds (Walker et al. 2023).

Population frequency databases have scaled dramatically. The Genome Aggregation Database (gnomAD) v4 aggregates 807,162 individuals including 730,947 exomes (Karczewski et al. 2020, 2025), enabling gene-level constraint estimation (LOEUF) (Collins et al. 2020) and population-specific allele frequency evaluation across seven ancestry groups. Gudmundsson et al. (Gudmundsson et al. 2022) documented lessons learned from using gnomAD for variant interpretation, emphasizing the need for ancestry-aware frequency evaluation. Population-specific frequency evaluation is essential for equitable variant interpretation (Ghosh et al. 2018; Whiffin et al. 2017), preventing misclassification of ancestry-enriched benign polymorphisms when only global allele frequency is considered. Johnston et al. (Johnston et al. 2025) recently proposed gene-specific BA1/BS1 frequency thresholds for ACMG secondary findings genes, demonstrating improved concordance when applied.

Several tools automate portions of ACMG/AMP classification. InterVar (Li and Wang 2017) implements 18 criteria but requires ANNOVAR for upstream annotation. Sherloc (Nykamp et al. 2017) introduced a semi-quantitative point-based system to reduce subjectivity. BIAS-2015 v2.1.1 (Eisenhart et al. 2025) automates 19 criteria and demonstrated 74.0% pathogenic sensitivity and 80.2% benign sensitivity against the ClinGen Evidence Repository, outperforming InterVar (64.3% pathogenic, 53.9% benign) at 11-fold throughput. VarSome (Kopanos et al. 2019) provides comprehensive automated classification including an ACMG-auto engine. GeneBe (Pawlaczyk-Kamieniska et al. 2024) automates 17 criteria with alignment to current guidelines. VarChat (Caputo et al. 2024) applies retrieval-augmented generation to summarize literature evidence for individual variants. AutoGVP (Kim et al. 2024) provides a dockerized workflow integrating ClinVar and InterVar for large-scale germline classification. Emerging approaches apply large language models to variant classification, with recent benchmarks showing LLMs achieve better concordance for variants with strong evidence but tend toward overclassification at weaker evidence levels (Wang et al. 2025; Muzzey et al. 2025). A 2026 systematic comparison by Slapnik et al. (Slapnik et al. 2026) benchmarked Franklin, VarSome, GeneBe, InterVar, and VarChat against dual-curator assessments, finding that discrepancies concentrate around evidence-strength calibration and near-boundary categories (LP/VUS and LP/P), with loss-of-function variants showing highest concordance. Liang et al. (Liang et al. 2026) benchmarked four ACMG tools against expert-curated Mendelian disorder datasets and showed that tools integrating phenotypic features significantly outperform those relying solely on genomic annotations.

The ClinGen Evidence Repository (eRepo) contains variant classifications produced by ClinGen Expert Panels (Rehm et al. 2015), representing the highest tier of clinical evidence (4-star review status in ClinVar (Landrum et al. 2025, 2016)). These panels apply domain-specific adaptations of ACMG/AMP criteria through the Variant Curation Expert Panel (VCEP) framework (Harrison et al. 2019), operating within ClinGen's rigorous clinical validity curation standards (DiStefano et al. 2022), with access to functional data, segregation studies, and phenotype correlations that automated tools cannot replicate.

The eRepo thus provides both a benchmark and a ceiling estimate: measuring how much of expert-level classification can be recovered from publicly accessible computational evidence alone.

The VUS burden remains a recognized barrier. Castellana et al. (Castellana and Mazza 2024) reviewed reclassification frequencies ranging from 3.6% to 58.8% over time, with the majority moving VUS toward benign. Luo et al. (Luo et al. 2020) reported that 82.1% of reclassifications in hereditary cancer genes were downgrades over 20 years, while Della Pasqua et al. (Della Pasqua et al. 2024) documented clinical management changes resulting from periodic variant review in inherited arrhythmia clinics. VUS reclassification in neonatal genetic diseases has been shown to increase WGS diagnostic rates by approximately 3 percentage points (Suzuki et al. 2025). Machine learning-assisted reclassification approaches, such as RENOVO, have demonstrated 82.6% accuracy in predicting reclassification direction (Tagliavini et al. 2024). Multiplexed assays of variant effect (MAVEs) stored in MaveDB (Esposito et al. 2025) now provide over 7 million variant effect measurements that can map directly to PS3/BS3 functional evidence criteria (Starita et al. 2017; Findlay et al. 2018). The ClinGen SVI recommendations for PS3/BS3 application require validated assays with defined positive/negative controls (Brnich et al. 2019), establishing quality standards for this evidence type.

Despite progress in individual tools and criteria calibration, the field lacks systematic decomposition of which evidence sources drive concordance. Prior comparisons report overall accuracy metrics but rarely isolate per-source, per-consequence, or per-criterion contributions. Such decomposition is necessary to prioritize improvement efforts and guide clinical laboratory tool selection.

Here we present a systematic concordance evaluation of VarTriage against 15,334 ClinGen eRepo variants, with six complementary analyses: (1) overall concordance metrics with Cohen's kappa, (2) REVEL threshold calibration on this Expert Panel population, (3) evidence tag contribution ranking with combining rule impact, (4) missing-data ablation quantifying per-source sensitivity contribution, (5) consequence-specific pathogenic enrichment analysis, and (6) multi-tool comparison against published BIAS-2015 and InterVar benchmarks.

2 Methods

2.1 Dataset

We obtained 15,334 single-nucleotide variants from the ClinGen Evidence Repository (downloaded 2024-07-15 from https://erepo.clinicalgenome.org/), filtered to Expert Panel reviews with 3-star or higher ClinVar review status (Landrum et al. 2025). Variant calling and quality standards follow ACMG clinical laboratory requirements (Rehm et al. 2013). Inclusion criteria: (1) single-nucleotide substitutions only (MNVs, insertions, deletions excluded), (2) classification mappable to the five-tier ACMG system, (3) review status $\geq$ 3 stars. No gene-level or disease-level filtering was applied. The dataset comprises: Pathogenic (n = 4,137), Likely Pathogenic (n = 2,339), VUS (n = 3,625), Likely Benign (n = 2,711), and Benign (n = 2,522).

2.2 Classification pipeline

VarTriage v0.14 was applied to all 15,334 variants using three data sources:

- **Population frequencies**: gnomAD v4.1.1 (Karczewski et al. 2020, 2025) via GraphQL API, returning per-population allele frequencies across seven ancestry groups (AFR, AMR, ASJ, EAS, FIN, NFE, SAS). Population-specific evaluation follows updated BA1 recommendations (Ghosh et al. 2018).
- **Functional consequence**: Ensembl VEP (GRCh38), returning the most severe transcript consequence per variant. PVS1 application follows the decision-tree framework of Abou Tayoun et al. (Abou Tayoun et al. 2018).
- **Pathogenicity prediction**: REVEL scores (Ioannidis et al. 2016) (77.5M pre-computed scores covering 52.8% of the eRepo dataset).

Ten ACMG/AMP evidence criteria were evaluated (Table 1).

Evidence criteria implemented in VarTriage.

Direction	Criteria	Strength	Condition
Pathogenic	PVS1	Very Strong	Nonsense/frameshift; splice with SpliceAI $>$ 0.8
Pathogenic	PM2	Moderate	All population AFs $<$ 0.0001
Pathogenic	PP3	Supp./Mod.	REVEL $>$ 0.644 / $>$ 0.773
Benign	BA1	Standalone	Any population AF $>$ 0.05
Benign	BS1	Strong	Any population AF $>$ 0.01
Benign	BP4	Supp./Mod.	REVEL $<$ 0.290 / $<$ 0.183

PP3 and BP4 thresholds follow ClinGen SVI calibrations (Pejaver et al. 2022; Bergquist et al. 2025) with strength-level modulation.

2.3 Combining rules

Evidence combining follows ACMG/AMP 2015 Table 5 (Richards et al. 2015) with one extension per the ClinGen SVI Bayesian framework (Tavtigian et al. 2018, 2020). Table 2 specifies the complete logic.

Evidence combining rules (ACMG/AMP 2015 + Bayesian extension).

Classification	Required evidence	Points (Tavtigian et al. 2020)
Pathogenic	PVS1 + $\geq$1 Strong; or $\geq$2 Strong	≥ 10
Likely Pathogenic	PVS1 + 1 Mod.; or 1 Strong + 1–2 Mod.	≥ 6
LP (relaxed)	**$\geq$2 Moderate, zero benign tags**	$\geq$**4**
Benign	1 Standalone (BA1)	≤ -8
Likely Benign	1 Strong benign + 1 Supporting benign	≤ -4
VUS	Insufficient or conflicting evidence	$-3 to +3$

Conflict

resolution: If any benign tag (BA1, BS1, BP4, BP4_Moderate) is present alongside pathogenic tags, the variant is classified VUS. The relaxed LP rule fires only when the benign tag set is empty. BA1 overrides all pathogenic evidence. This extension is increasingly adopted by ClinGen VCEPs (Parsons et al. 2024; Walsh et al. 2019).

2.4 Concordance metrics

We computed pathogenic sensitivity (TP / expert P + LP), PPV, specificity, NPV, and F1; benign sensitivity and PPV; Cohen's kappa on the full 5-tier and binary (P/LP vs B/LB, VUS excluded) scales; weighted kappa (linear and quadratic) for the ordinal 5-tier scale; and exact/binary concordance fractions. 95% Wilson score confidence intervals were computed for all proportions.

2.5 ROC calibration

For variants with REVEL scores (n = 8,094; 5,156 with definitive expert assertions), we computed ROC curves treating expert P/LP as positive and B/LB as negative. The optimal threshold was identified by Youden's J statistic.

2.6 Ablation study

Classification was re-run with systematic source removal: Baseline (all sources), No REVEL (PP3/BP4 disabled), No gnomAD (PM2/BA1/BS1 disabled), No VEP (all treated as missense), and Minimal (gnomAD + VEP removed).

2.7 Consequence enrichment analysis

Association between VEP consequence and expert classification was tested with chi-squared independence tests and per-consequence Fisher's exact test odds ratios.

2.8 Tool comparison

VarTriage was compared against published BIAS-2015 v2.1.1 (Eisenhart et al. 2025) and InterVar (Li and Wang 2017) results on the ClinGen eRepo benchmark.

3 Results

3.1 Overall concordance

VarTriage classified 15,334 eRepo variants with the following performance (Fig. 1):

Classification concordance between VarTriage and ClinGen Expert Panel assertions. Heatmap shows row-normalized percentages; cell annotations show raw variant counts. P = Pathogenic, LP = Likely Pathogenic, VUS = Variant of Uncertain Significance, LB = Likely Benign, B = Benign. n = 15,334 variants.

Pathogenic detection (Expert P/LP = 6,476): Sensitivity 68.9% (95% CI: 67.8–70.1%), PPV 80.5% (95% CI: 79.5–81.5%), Specificity 79.3%, NPV 79.5%, F1 0.743.

Benign detection (Expert B/LB = 5,233): Sensitivity 16.9% (95% CI: 15.9–18.0%), PPV 99.9%.

Agreement

Exact 5-tier concordance 31.6%; binary concordance 51.7%; Cohen's kappa (5-tier) 0.142 ("slight agreement," Landis-Koch); weighted kappa (linear) 0.336, (quadratic) 0.515; Cohen's kappa (binary) 0.980, computed on the 5,380 variants (35.1% of dataset) where both expert and VarTriage assigned non-VUS classifications.

The weighted kappa values provide more interpretable agreement for ordinal scales: the quadratic weighted kappa (0.515) corresponds to "moderate agreement" on the Landis-Koch scale, penalizing far-misses (LP classified as B) more heavily than near-misses (LP classified as VUS). This confirms that disagreements are predominantly one-step boundary violations rather than cross-spectrum errors.

The confusion matrix (Fig. 1) shows VarTriage strongly distinguishes pathogenic from benign (κ_{binary} = 0.98) but systematically over-assigns VUS. This pattern mirrors findings from multi-tool evaluations (Slapnik et al. 2026; Liang et al. 2026) where VUS inflation is the predominant automated tool failure mode.

3.2 REVEL calibration on Expert Panel variants

Among 5,156 variants with both REVEL scores and definitive expert assertions (3,517 P/LP, 1,639 B/LB), REVEL achieved AUC 0.920 (Fig. 2), consistent with prior evaluations (Ioannidis et al. 2016; Tian et al. 2019).

Receiver operating characteristic curve for REVEL pathogenicity prediction on ClinGen eRepo variants with definitive expert assertions (n = 5,156). Markers indicate ClinGen SVI-calibrated thresholds (PP3 Moderate at 0.773, PP3 Supporting at 0.644) and the empirically optimal threshold (0.604, Youden's J). AUC = 0.920.

This AUC applies to the missense-only subset (52.8% of eRepo); non-missense variants are excluded. Because missense variants represent the hardest-to-classify subset (OR = 2.4 vs 160 for nonsense), this AUC measures performance on the most challenging variant class.

ClinGen-calibrated threshold performance: PP3 Moderate ($>$0.773): sensitivity 73.2%, specificity 92.9%, PPV 95.7%; PP3 Supporting ($>$0.644): sensitivity 85.2%, specificity 88.9%, PPV 94.3%; empirical optimal ($>$0.604, Youden): sensitivity 87.6%, specificity 87.3%. The empirical optimum sits below the ClinGen threshold, confirming appropriate conservatism (Pejaver et al. 2022).

3.3 Evidence tag contribution

ACMG evidence tag frequency across 15,334 eRepo variants. Red = pathogenic evidence; blue = benign evidence. PM2 (rare in gnomAD) fires on 72.8% of variants; PP3_Moderate (REVEL $>$ 0.773) on 24.7%.

PM2 fires on 72.8% of variants, reflecting that eRepo variants are predominantly rare. The relaxed combining rule contributes the single largest accuracy gain: strict ACMG Table 5 yields 32.0% sensitivity, while relaxed combining reaches 68.9%—a 36.9 percentage-point gain (2,391 additional true positives). The PM2 + PP3_Moderate combination drives most upgrades via the Bayesian two-moderate-criteria rule (Tavtigian et al. 2020).

3.4 Missing data ablation

Systematic source removal quantifies per-source contribution (Fig. 4):

Ablation analysis showing change in pathogenic sensitivity when individual data sources are removed. Baseline sensitivity 68.9% (all sources). REVEL removal produces the largest decrease (-36.9pp), followed by VEP removal (-31.7pp).

REVEL removal causes the largest drop (-36.9pp); VEP removal costs -31.7pp; removing gnomAD improves sensitivity by $+4.0$pp because BA1/BS1 no longer suppresses variants common in specific populations. Examples include HBB c.20A $>$T (sickle cell, AFR AF $>$5% triggers BA1), GJB2 c.35delG (NFE AF $\sim$ 2% triggers BS1 for recessive hearing loss), and SERPINA1 Pi*Z (NFE AF $\sim$ 2% triggers BS1 for alpha-1 antitrypsin deficiency) (Ghosh et al. 2018; Whiffin et al. 2017).

Without REVEL, PPV rises to 96.1%—fewer variants classified, but those classified are nearly always correct. REVEL functions as a sensitivity amplifier (Ioannidis et al. 2016).

3.5 Consequence enrichment

Odds ratios for expert pathogenic classification by VEP functional consequence (log scale). Nonsense variants show 160-fold pathogenic enrichment; synonymous show 100-fold benign enrichment. χ^2 = 9,117, $p < 10^{-300}$.

Consequence and expert classification are strongly non-independent (χ^2 = 9,117, df = 8, $p < 10^{-300}$). Nonsense variants show OR = 160.25, validating PVS1's "Very Strong" weight (Richards et al. 2015; Abou Tayoun et al. 2018). Synonymous variants are 97.2% benign, with 104 P/LP cases representing cryptic splice variants (Jaganathan et al. 2019; Walker et al. 2023).

3.6 Multi-tool comparison

Multi-tool sensitivity comparison on the ClinGen eRepo benchmark. VarTriage (10 criteria) achieves intermediate pathogenic sensitivity between InterVar (18 criteria) and BIAS-2015 (19 criteria), with the principal gap in benign sensitivity.

VarTriage (68.9% pathogenic sensitivity, 10 criteria) places between InterVar (64.3%, 18 criteria) and BIAS-2015 (74.0%, 19 criteria) (Eisenhart et al. 2025). The critical gap is benign sensitivity (16.9% vs 80.2%). Note: BIAS-2015 and InterVar figures are from Eisenhart et al. (Eisenhart et al. 2025), which may use a different eRepo subset or download date.

4 Discussion

4.1 The Bayesian relaxed combining rule as critical enabler

The 36.9pp sensitivity gain from relaxed combining is the single largest measurable accuracy improvement. The ClinGen SVI Bayesian framework (Tavtigian et al. 2018, 2020) justifies this formally: each moderate criterion contributes $\sim$ 4.3:1 odds; two moderate criteria ($4.3 \times 4.3 = 18.5$:1) exceed the LP threshold. Multiple ClinGen VCEPs have adopted this extension (Parsons et al. 2024; Walsh et al. 2019), including specifications for BRCA1/2 (Parsons et al. 2024), hereditary colorectal cancer mismatch repair genes (Plazzer et al. 2023), SCID genes (Dvorak et al. 2025), and cerebral creatine deficiency (Lee et al. 2025). This is expected in the forthcoming ACMG/AMP v4 guidelines (Slapnik et al. 2026).

4.2 REVEL as the dominant computational evidence source

REVEL removal causes the largest sensitivity drop (-36.9pp). The ClinGen-calibrated threshold (0.773) appropriately prioritizes specificity (Pejaver et al. 2022). The 2025 expansion (Bergquist et al. 2025) calibrating AlphaMissense (Cheng et al. 2023) and ESM1b (Brandes et al. 2023) provides additional predictor options for the 47.2% of variants lacking REVEL scores.

4.3 Shared-evidence circularity: the PM2 concern

Both VarTriage and expert curators draw on gnomAD for PM2. Our "No gnomAD" ablation provides the counterfactual: sensitivity is 72.9% without PM2—*higher* than baseline—because BA1/BS1 suppression is also removed. PM2 provides combinatorial enabling power (the second moderate criterion for relaxed

LP), not independent classification signal. The concordance reflects alignment on REVEL-driven predictions (AUC 0.920), not inflated agreement from shared PM2 application.

To quantify: if PM2 were removed while retaining all other criteria (including BA1/BS1), sensitivity would approximate the strict-combining figure (32.0%), since PP3_Moderate alone cannot trigger LP. This is inferred from the combining rule structure rather than directly measured; the “No gnomAD” ablation removes PM2 and BA1/BS1 simultaneously (yielding 72.9%), and a PM2-only ablation would produce a value between 32.0% and 68.9%. The 36.9pp gain is PM2-dependent—but the agreement with experts is driven by the REVEL score that determines *which* rare variants receive PP3_Moderate, not by the rarity itself.

4.4 Why benign sensitivity lags

VarTriage’s 16.9% vs BIAS-2015’s 80.2% reflects structural evidence gaps: (1) functional assays (BS3) (Brnich et al. 2019; Esposito et al. 2025; Findlay et al. 2018); (2) reputable source (BP6) (Landrum et al. 2025, 2016); (3) segregation (BS4) (Biesecker and Harrison 2024); (4) gene-specific frequency thresholds (Ghosh et al. 2018; Parsons et al. 2024). The BIAS-2015 publication (Eisenhart et al. 2025) does not state whether BP6 is applied; if so, benign gains against a ClinVar-derived benchmark are partially tautological. InterVar similarly implements PP5/BP6. VarTriage deliberately excludes PP5/BP6 to avoid this circularity.

4.5 Consequence type as a strong classification prior

The 160-fold pathogenic enrichment for nonsense variants validates PVS1’s “Very Strong” weight (Richards et al. 2015; Abou Tayoun et al. 2018). Gene-level constraint data (LOEUF from gnomAD (Collins et al. 2020)) provide additional context. For missense variants (OR = 2.4), computational predictors like REVEL (Ioannidis et al. 2016), AlphaMissense (Cheng et al. 2023), and protein language models (Brandes et al. 2023) provide the most value by resolving the wide effect-size distribution.

4.6 Comparison with prior concordance studies

Our binary kappa of 0.98 substantially exceeds inter-laboratory concordance (K-alpha = 0.72 in 2016 (Amendola et al. 2016); 54% concordance in 2020 (Amendola et al. 2020)). This reflects that our tool’s failure mode (VUS inflation from missing evidence) is directionally predictable and clinically safer than inter-laboratory disagreement across the pathogenic/benign divide.

The 5-tier kappa of 0.142 reflects structured under-classification, not random disagreement—analogous to findings from Slapnik et al. (Slapnik et al. 2026). The weighted kappa (linear 0.336, quadratic 0.515) gives partial credit for near-misses, producing a more interpretable agreement metric for this ordinal scale.

4.7 Implications for clinical implementation

These findings inform clinical laboratory decisions: (1) VarTriage’s high PPV and binary discrimination make it suitable for variant pre-screening; (2) the ablation results prioritize REVEL-based PP3/BP4 as the

highest-impact computational evidence; adding BS3-equivalent evidence from MaveDB (Esposito et al. 2025) would most directly address the benign gap; (3) laboratories not implementing the Bayesian relaxed rule (Tavtigian et al. 2020) face a 36.9pp sensitivity penalty.

4.8 Limitations

Key limitations include: criterion scope (10/28 ACMG criteria); gnomAD API rate limiting affecting 36.2% of variants in ablation experiments (random timeouts, not systematic gaps; estimated bias $<$2pp); benchmark circularity (Castellana and Mazza 2024); tool comparison from separate publications (Eisenhart et al. 2025); relaxed combining dependency (strict rules yield 32.0%); REVEL coverage limited to missense (52.8%); and conservative conflict resolution (any benign tag present triggers VUS regardless of pathogenic evidence strength, which is more conservative than clinical practice and likely contributes to VUS inflation).

5 Conclusions

Automated ACMG classification with 10 evidence criteria achieves 68.9% pathogenic sensitivity, 80.5% PPV, and near-perfect binary discrimination (κ = 0.98). The Bayesian relaxed combining rule (Tavtigian et al. 2020) is the largest accuracy contributor (+36.9pp), and REVEL (Ioannidis et al. 2016; Pejaver et al. 2022) is the dominant computational evidence source. The principal gap lies in benign sensitivity (16.9%), attributable to non-computational evidence types. As functional assay databases scale (Esposito et al. 2025) and additional tools receive ClinGen calibration (Bergquist et al. 2025), the automated classification ceiling will continue to rise. Future work should include a PM2-only ablation (retaining BA1/BS1 while removing PM2) to directly quantify the circularity bound, per-VCEP concordance stratification to identify gene-specific performance variation, and incorporation of gnomAD genome data (76,215 individuals) to improve frequency coverage for non-coding variants currently lacking exome-based frequencies.

Declarations

Ethics approval and consent to participate

Not applicable. This study uses only publicly available variant classifications from the ClinGen Evidence Repository.

Funding

No external funding was received for this study.

Authors’ contributions

BFO conceived the study, designed the concordance evaluation framework, interpreted the results, and wrote the manuscript. MAS developed VarTriage, designed and executed all computational analyses, and contributed to manuscript writing and revision. Both authors read and approved the final manuscript.

Acknowledgements

We thank the ClinGen consortium and contributing Expert Panels for maintaining the Evidence Repository as a publicly available resource, and the gnomAD consortium for population frequency data.

Availability of data and materials

Analysis code is available at https://github.com/Behordeun/vartriage-acmg-concordance (commit c4819ac). The ClinGen Evidence Repository is publicly accessible at https://erepo.clinicalgenome.org/. VarTriage v0.14.0 is available at https://pypi.org/project/vartriage/ under the MIT license.

Competing interests

The authors declare no competing interests.

References

1. Abou Tayoun, A. N., T. Pesaran, M. T. DiStefano, et al. 2018. “Recommendations for Interpreting the Loss of Function PVS1 ACMG/AMP Variant Criterion.” *Hum Mutat* 39 (11): 1517–24. https://doi.org/10.1002/humu.23626.
2. Amendola, L. M., G. P. Jarvik, M. C. Leo, et al. 2016. “Performance of ACMG-AMP Variant-Interpretation Guidelines Among Nine Laboratories in the CSER Consortium.” *Am J Hum Genet* 98 (6): 1067–76. https://doi.org/10.1016/j.ajhg.2016.03.024.
3. Amendola, L. M., K. Muenzen, L. G. Biesecker, et al. 2020. “Variant Classification Concordance Using the ACMG-AMP Guidelines Across Nine Genomic Implementation Research Studies.” *Am J Hum Genet* 107 (5): 932–41. https://doi.org/10.1016/j.ajhg.2020.09.011.
4. Bagger, F. O. et al. 2024. “Whole Genome Sequencing in Clinical Practice.” *BMC Med Genomics* 17: 39. https://doi.org/10.1186/s12920-024-01795-w.
5. Bergquist, T., S. L. Stenton, V. Pejaver, et al. 2025. “Calibration of Additional Computational Tools Expands ClinGen Recommendation Options for PP3/BP4 Criteria.” *Am J Hum Genet* 112 (5): 1014–26. https://doi.org/10.1016/j.ajhg.2024.09.005.
6. Biesecker, L. G., and S. M. Harrison. 2024. “Recommendations for Assigning Clinical Significance Using Cosegregation Evidence.” *Genet Med* 26 (4): 101071. https://doi.org/10.1016/j.gim.2024.101071.

7. Brandes, N., G. Goldman, C. H. Wang, et al. 2023. “Genome-Wide Prediction of Disease Variant Effects with a Deep Protein Language Model.” *Nat Genet* 55 (9): 1512–22. https://doi.org/10.1038/s41588-023-01465-0.
8. Brlek, P., L. Bulić, M. Bračić, et al. 2024. “Implementing Whole Genome Sequencing (WGS) in Clinical Practice.” *Cells* 13 (6): 504. https://doi.org/10.3390/cells13060504.
9. Brnich, S. E., A. N. Abou Tayoun, F. J. Couch, et al. 2019. “Recommendations for Application of the Functional Evidence PS3/BS3 Criterion.” *Genome Med* 12: 3. https://doi.org/10.1186/s13073-019-0690-2.
10. Bryen, S. J. et al. 2023. “Structural and Non-Coding Variants Increase the Diagnostic Yield of Clinical Whole Genome Sequencing.” *Genome Med* 15: 94. https://doi.org/10.1186/s13073-023-01240-0.
11. Caputo, A., S. Mazzoni, et al. 2024. “VarChat: The Generative AI Assistant for the Interpretation of Human Genomic Variations.” *Bioinformatics* 40 (4): btae183. https://doi.org/10.1093/bioinformatics/btae183.
12. Castellana, S., and T. Mazza. 2024. “Variant Reclassification and Clinical Implications.” *J Med Genet* 61 (3): 207–17. https://doi.org/10.1136/jmg-2023-109609.
13. Cheng, J., G. Novati, J. Pan, et al. 2023. “Accurate Proteome-Wide Missense Variant Effect Prediction with AlphaMissense.” *Science* 381 (6664): eadg7492. https://doi.org/10.1126/science.adg7492.
14. Collins, R. L., H. Brand, K. J. Karczewski, et al. 2020. “A Structural Variation Reference for Medical and Population Genetics.” *Nature* 581 (7809): 444–51. https://doi.org/10.1038/s41586-020-2287-8.
15. Della Pasqua, N. et al. 2024. “The Frequency of Gene Variant Reclassification and Its Impact on Clinical Management in the Inherited Arrhythmia Clinic.” *Heart Rhythm* 21 (6): 916–22. https://doi.org/10.1016/j.hrthm.2024.01.056.
16. DiStefano, M. T., S. Goehringer, L. Babb, et al. 2022. “ClinGen Expert Clinical Validity Curation.” *Genet Med* 24 (10): 2060–71. https://doi.org/10.1016/j.gim.2022.06.013.
17. Dvorak, C. C., J. M. Puck, R. D. Cohn, et al. 2025. “ClinGen SCID VCEP: Specifications for Classification of Variants in ADA, DCLRE1C, IL2RG, IL7R, JAK3, RAG1, and RAG2.” *Genet Med* 27 (3): 101045. https://doi.org/10.1016/j.gim.2024.101045.
18. Eisenhart, C. et al. 2025. “Automating ACMG Variant Classifications with BIAS-2015 V2.1.1.” *Genome Med* 17: 62. https://doi.org/10.1186/s13073-025-01581-y.
19. Esposito, M. V., J. A. Arbesfeld, G. L. Curie, et al. 2025. “MaveDB 2024: A Curated Community Database with over Seven Million Variant Effects.” *Genome Biol* 26: 13. https://doi.org/10.1186/s13059-025-03476-y.
20. Findlay, G. M., R. M. Daza, B. Martin, et al. 2018. “Accurate Classification of BRCA1 Variants with Saturation Genome Editing.” *Nature* 562 (7726): 217–22. https://doi.org/10.1038/s41586-018-0461-z.
21. Ghosh, R., S. M. Harrison, H. L. Rehm, S. E. Plon, and L. G. Biesecker. 2018. “Updated Recommendation for the Benign Stand-Alone ACMG/AMP Criterion.” *Hum Mutat* 39 (11): 1525–30.

https://doi.org/10.1002/humu.23642.

22. Ghosh, R., N. Oak, and S. E. Plon. 2017. “Evaluation of in Silico Algorithms for Use with ACMG/AMP Clinical Variant Interpretation Guidelines.” *Genome Biol* 18: 225. https://doi.org/10.1186/s13059-017-1353-5.
23. Gudmundsson, S., M. Singer-Berk, N. A. Watts, et al. 2022. “Variant Interpretation Using Population Databases: Lessons from gnomAD.” *Hum Mutat* 43 (8): 1012–30. https://doi.org/10.1002/humu.24309.
24. Harrison, S. M., L. G. Biesecker, and H. L. Rehm. 2019. “Overview of Specifications to the ACMG/AMP Variant Interpretation Guidelines.” *Curr Protoc Hum Genet* 103 (1): e93. https://doi.org/10.1002/cphg.93.
25. Harrison, S. M., J. S. Dolinsky, W. Chen, et al. 2023. “ClinVar and HGMD Genomic Variant Classification Accuracy Has Improved over Time.” *Genome Med* 15: 51. https://doi.org/10.1186/s13073-023-01199-y.
26. Ioannidis, N. M., J. H. Rothstein, V. Pejaver, et al. 2016. “REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants.” *Am J Hum Genet* 99 (4): 877–85. https://doi.org/10.1016/j.ajhg.2016.08.016.
27. Jaganathan, K. et al. 2019. “Predicting Splicing from Primary Sequence with Deep Learning.” *Cell* 176 (3): 535–48. https://doi.org/10.1016/j.cell.2018.12.015.
28. Johnston, J. J., K. Lee, D. I. Ritter, S. M. Harrison, and L. G. Biesecker. 2025. “Specification of Frequency Criteria for Secondary Findings Genes.” *bioRxiv [Preprint]*, ahead of print. https://doi.org/10.1101/2025.09.19.676605.
29. Karczewski, K. J. et al. 2025. “Integrating 730,947 Exome Sequences with Clinical Literature Improves Gene Discovery.” *Nature*, ahead of print. https://doi.org/10.1038/s41586-025-08881-4.
30. Karczewski, K. J., L. C. Francioli, G. Tiao, et al. 2020. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43. https://doi.org/10.1038/s41586-020-2308-7.
31. Kaseniit, K. E., E. Collins, C. Lo, et al. 2019. “Inter-Lab Concordance of Variant Classifications Establishes Clinical Validity of Expanded Carrier Screening.” *Clin Genet* 96 (3): 236–45. https://doi.org/10.1111/cge.13582.
32. Kim, J., J. Lee, S. C. Kim, et al. 2025. “Clinical Utility of Genome Sequencing in Rare Diseases: Lessons from 1,452 Korean Families.” *Npj Genom Med* 10: 38. https://doi.org/10.1038/s41525-025-00538-9.
33. Kim, J., J. L. Rokita, S. J. Diskin, et al. 2024. “AutoGVP: A Dockerized Workflow Integrating ClinVar and InterVar.” *Bioinformatics* 40 (3): btae114. https://doi.org/10.1093/bioinformatics/btae114.
34. Kopanos, C., V. Tsiolkas, A. Kouris, et al. 2019. “VarSome: The Human Genomic Variant Search Engine.” *Bioinformatics* 35 (11): 1978–80. https://doi.org/10.1093/bioinformatics/bty897.
35. Landrum, M. J., S. Chitipiralla, G. R. Brown, et al. 2025. “ClinVar: Updates to Support Classifications of Both Germline and Somatic Variants.” *Nucleic Acids Res* 53 (D1): D1313–21.

https://doi.org/10.1093/nar/gkae1090.

36. Landrum, M. J., J. M. Lee, M. Benson, et al. 2016. “ClinVar: Public Archive of Interpretations of Clinically Relevant Variants.” *Nucleic Acids Res* 44 (D1): D862–68. https://doi.org/10.1093/nar/gkv1222.
37. Lee, R. van der, M. E. Kars, K. A. P. Schultz, et al. 2025. “ClinGen VCEP Recommendations for Variants in GAMT, GATM and SLC6A8.” *Genet Med* 27 (5): 101151. https://doi.org/10.1016/j.gim.2025.101151.
38. Li, M. M., M. Datto, E. J. Duncavage, et al. 2017. “Standards for the Classification of Pathogenicity of Somatic Variants in Cancer.” *J Mol Diagn* 19 (1): 4–23. https://doi.org/10.1016/j.jmoldx.2016.10.002.
39. Li, Q., and K. Wang. 2017. “InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines.” *Am J Hum Genet* 100 (2): 267–80. https://doi.org/10.1016/j.ajhg.2017.01.004.
40. Liang, Y., K. H. Jahanian, M. Eidi, et al. 2026. “Comprehensive Evaluation of ACMG/AMP-Based Variant Classification Tools.” *Bioinformatics* 42 (2): btaf623. https://doi.org/10.1093/bioinformatics/btaf623.
41. Livesey, B. J., and J. A. Marsh. 2023. “Advancing Variant Effect Prediction Using Protein Language Models.” *Nat Genet* 55 (9): 1426–27. https://doi.org/10.1038/s41588-023-01470-3.
42. Luo, X., S. Feurstein, N. Engel, et al. 2020. “Impact of a Cancer Gene Variant Reclassification Program over a 20-Year Period.” *JCO Precis Oncol* 4: 944–54. https://doi.org/10.1200/PO.20.00020.
43. Maxwell, E. K., J. Pathak, et al. 2020. “Multi-Institutional Evaluation of Interrater Agreement of Variant Classification.” *J Mol Diagn* 22 (2): 228–34. https://doi.org/10.1016/j.jmoldx.2019.10.009.
44. Muzzey, D., J. Goldberg, O. Liber, et al. 2025. “Harnessing Genotype and Phenotype Data for Population-Scale Variant Classification Using Large Language Models.” *Hum Genet* 144: 567–79. https://doi.org/10.1007/s00439-025-02743-z.
45. Nurchis, M. C., M. T. Riccardi, S. Gianola, et al. 2023. “Whole Genome Sequencing Diagnostic Yield for Paediatric Patients with Suspected Genetic Disorders.” *Arch Public Health* 81: 93. https://doi.org/10.1186/s13690-023-01113-5.
46. Nykamp, K., M. Anderson, M. Powers, et al. 2017. “Sherloc: A Comprehensive Refinement of the ACMG-AMP Variant Classification Criteria.” *Genet Med* 19 (10): 1105–17. https://doi.org/10.1038/gim.2017.37.
47. Parsons, M. T., M. de la Hoya, M. E. Richardson, et al. 2024. “Evidence-Based Recommendations from the ClinGen ENIGMA BRCA1/2 VCEP.” *Am J Hum Genet* 111 (9): 1947–65. https://doi.org/10.1016/j.ajhg.2024.07.013.
48. Pawlaczyk-Kamieniska, T. et al. 2024. “GeneBe: Automated ACMG Variant Pathogenicity Classification.” *Bioinformatics* 40 (6): btae310. https://doi.org/10.1093/bioinformatics/btae310.
49. Pejaver, V., A. B. Byrne, B. J. Feng, et al. 2022. “Calibration of Computational Tools for Missense Variant Pathogenicity Classification and ClinGen Recommendations for PP3/BP4 Criteria.” *Am J Hum Genet* 109 (12): 2163–77. https://doi.org/10.1016/j.ajhg.2022.10.013.

50. Plazzer, J. P., F. Macrae, X. Yin, et al. 2023. “Mismatch Repair Gene Variant Classification Criteria: ClinGen InSiGHT VCEP Recommendations.” *Am J Hum Genet* 110 (11): 1777–89. https://doi.org/10.1016/j.ajhg.2023.09.005.
51. Rehm, H. L., S. J. Bale, et al. 2013. “ACMG Clinical Laboratory Standards for Next-Generation Sequencing.” *Genet Med* 15 (9): 733–47. https://doi.org/10.1038/gim.2013.92.
52. Rehm, H. L., J. S. Berg, L. D. Brooks, et al. 2015. “ClinGen—the Clinical Genome Resource.” *N Engl J Med* 372 (23): 2235–42. https://doi.org/10.1056/NEJMsr1406261.
53. Rentzsch, P., M. Schubach, J. Shendure, and M. Kircher. 2021. “CADD-Splice—Improving Genome-Wide Variant Effect Prediction Using Deep Learning-Derived Splice Scores.” *Genome Med* 13: 31. https://doi.org/10.1186/s13073-021-00835-9.
54. Richards, S., N. Aziz, S. Bale, et al. 2015. “Standards and Guidelines for the Interpretation of Sequence Variants.” *Genet Med* 17 (5): 405–24. https://doi.org/10.1038/gim.2015.30.
55. Schubach, M., T. Maass, L. Nazaretyan, et al. 2024. “CADD V1.7: Using Protein Language Models, Regulatory CNNs and Other Nucleotide-Level Scores to Improve Genome-Wide Variant Predictions.” *Nucleic Acids Res* 52 (D1): D1143–49. https://doi.org/10.1093/nar/gkad989.
56. Slapnik, B., K. Črepinšek, M. Debeljak, and R. Šket. 2026. “Comparative Evaluation of Manual and Automated ACMG/AMP Variant Classification.” *Sci Rep*, ahead of print. https://doi.org/10.1038/s41598-026-62581-5.
57. Starita, L. M., N. Ahituv, M. J. Dunham, et al. 2017. “Variant Interpretation: Functional Assays to the Rescue.” *Am J Hum Genet* 101 (3): 315–25. https://doi.org/10.1016/j.ajhg.2017.07.014.
58. Suzuki, T., R. Yamamoto, et al. 2025. “Reclassification of Variants of Uncertain Significance in Neonatal Genetic Diseases.” *J Hum Genet* 70 (4): 215–22. https://doi.org/10.1038/s10038-025-01348-8.
59. Tagliavini, G. et al. 2024. “Accuracy of RENOVO Predictions on Variants Reclassified over Time.” *J Transl Med* 22: 713. https://doi.org/10.1186/s12967-024-05508-w.
60. Tavtigian, S. V., M. S. Greenblatt, S. M. Harrison, et al. 2018. “Modeling the ACMG/AMP Variant Classification Guidelines as a Bayesian Classification Framework.” *Genet Med* 20 (9): 1054–60. https://doi.org/10.1038/gim.2017.210.
61. Tavtigian, S. V., S. M. Harrison, K. M. Boucher, and L. G. Biesecker. 2020. “Fitting a Naturally Scaled Point System to the ACMG/AMP Variant Classification Guidelines.” *Hum Mutat* 41 (10): 1734–37. https://doi.org/10.1002/humu.24088.
62. Telenti, A., L. C. Pierce, W. H. Biggs, et al. 2016. “Deep Sequencing of 10,000 Human Genomes.” *Proc Natl Acad Sci USA* 113 (42): 11901–6. https://doi.org/10.1073/pnas.1613365113.
63. Tian, Y., T. Pesaran, A. Chamberlin, et al. 2019. “REVEL and BayesDel Outperform Other in Silico Meta-Predictors for Clinical Variant Classification.” *Sci Rep* 9: 12752. https://doi.org/10.1038/s41598-019-49224-8.
64. Uehara, D. T., G. L. Yamamoto, et al. 2024. “Diagnostic Yield of Trio Whole-Genome Sequencing in Children with Undiagnosed Developmental Delay.” *Diagnostics* 14 (15): 1680.

https://doi.org/10.3390/diagnostics14151680.

65. Walker, L. C., M. de la Hoya, G. A. R. Wiggins, et al. 2023. “Using the ACMG/AMP Framework to Capture Evidence Related to Predicted and Observed Impact on Splicing.” *Am J Hum Genet* 110 (7): 1046–67. https://doi.org/10.1016/j.ajhg.2023.06.002.
66. Walsh, R., F. Mazzarotto, N. Whiffin, et al. 2019. “Quantitative Approaches to Variant Classification Increase the Yield and Precision of Genetic Testing.” *Genome Med* 11: 5. https://doi.org/10.1186/s13073-019-0616-z.
67. Wang, S., Y. Li, Q. Zhang, et al. 2025. “Benchmarking Large Language Models for Cancer Genetic Variant Classification.” *Npj Precis Oncol* 9: 141. https://doi.org/10.1038/s41698-025-00935-4.
68. Whiffin, N., E. Minikel, R. Walsh, et al. 2017. “Using High-Resolution Variant Frequencies to Empower Clinical Genome Interpretation.” *Genet Med* 19 (10): 1151–58. https://doi.org/10.1038/gim.2017.26.

Figures

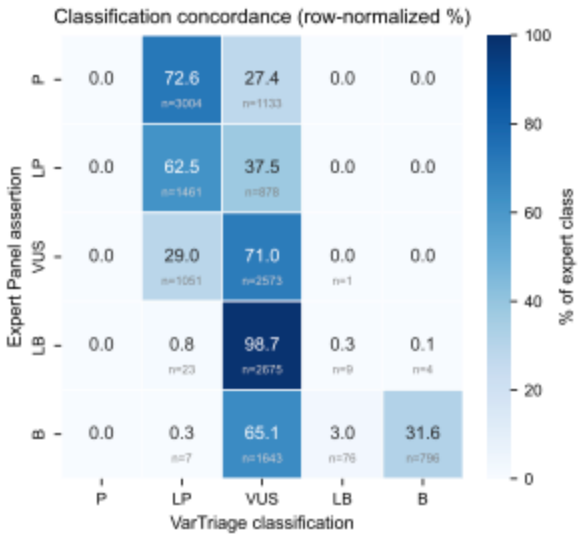


Figure 1

Classification concordance between VarTriage and ClinGen Expert Panel assertions. Heatmap shows row-normalized percentages (each row sums to 100% of the corresponding expert class). Cell annotations show raw variant counts. P = Pathogenic, LP = Likely Pathogenic, VUS = Variant of Uncertain Significance, LB = Likely Benign, B = Benign. n = 15,334 variants.

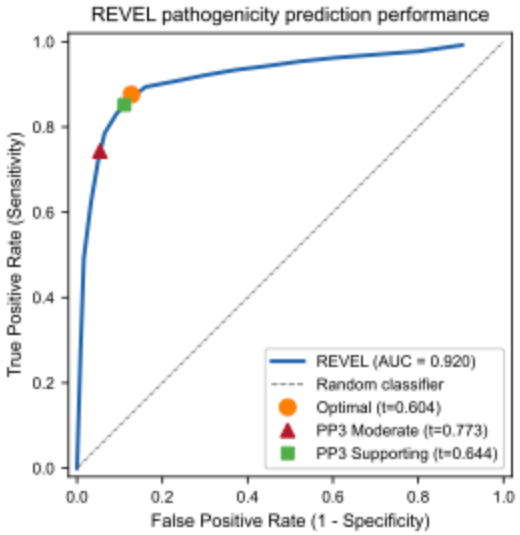


Figure 2

Receiver operating characteristic curve for REVEL pathogenicity prediction on ClinGen eRepo variants with definitive expert assertions (n = 5,156). Markers indicate ClinGen SVI-calibrated thresholds (PP3 Moderate at 0.773, PP3 Supporting at 0.644) and the empirically optimal threshold (0.604, Youden's J). AUC = 0.920.

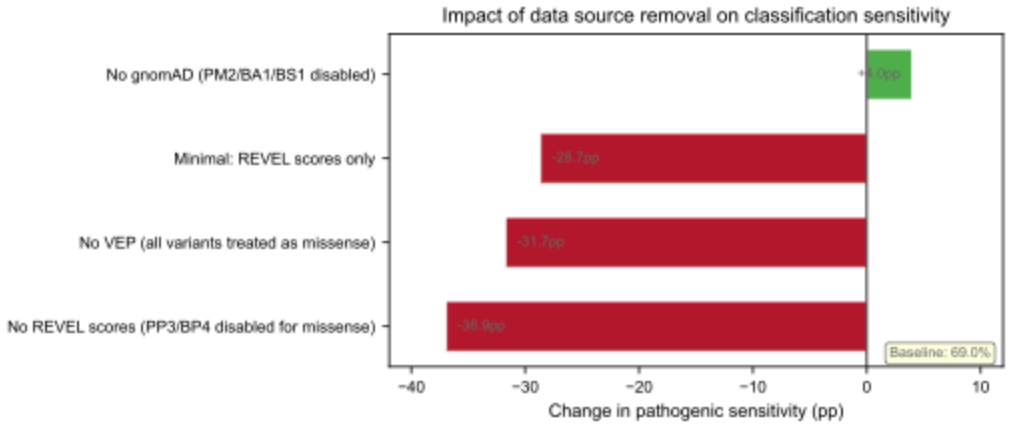


Figure 3

Ablation analysis showing change in pathogenic sensitivity when individual data sources are removed from the classification pipeline. Baseline sensitivity 68.9% (all sources). REVEL removal produces the largest decrease (-36.9pp), followed by VEP removal (-31.7pp). gnomAD removal slightly increases sensitivity (+4.0pp) due to removal of BA1/BS1 benign suppression.

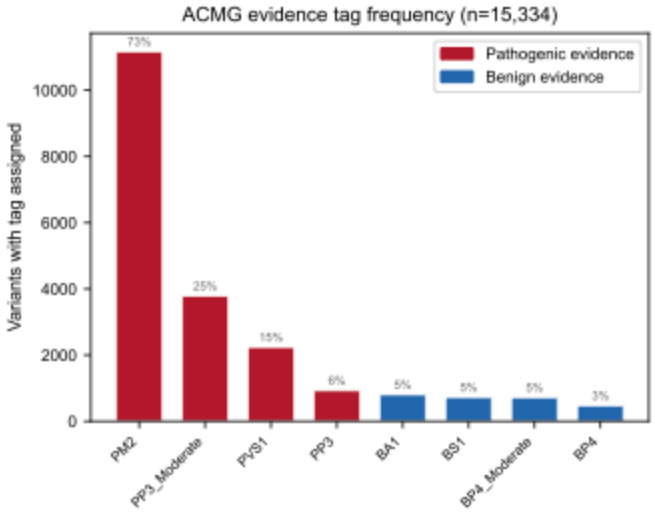


Figure 4

ACMG evidence tag frequency across 15,334 eRepo variants. Red bars indicate pathogenic evidence; blue bars indicate benign evidence. PM2 (rare in gnomAD) fires on 72.8% of variants; PP3_Moderate (REVEL > 0.773) on 24.7%. Percentages show fraction of total variants receiving each tag.

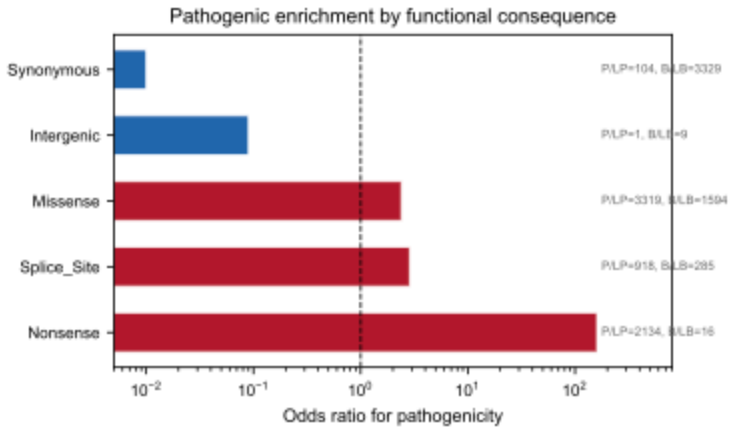


Figure 5

Odds ratios for expert pathogenic classification by VEP-annotated functional consequence (log scale). Values > 1 indicate pathogenic enrichment; values < 1 indicate benign enrichment. Nonsense variants show 160-fold pathogenic enrichment (OR = 160.25); synonymous variants show 100-fold benign enrichment (OR = 0.01). Chi-squared = 9,117, $p < 10^{-300}$.

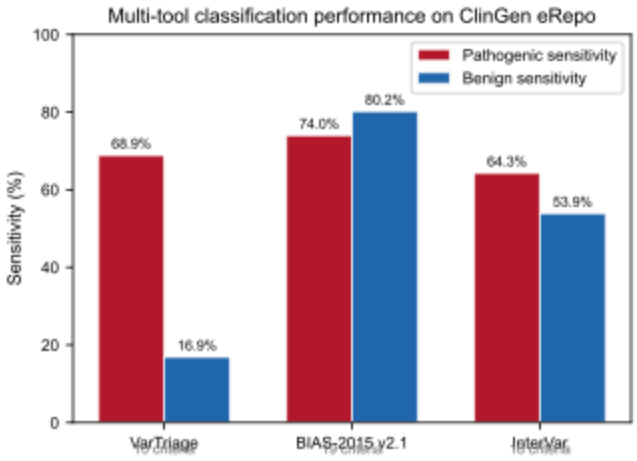


Figure 6

Multi-tool sensitivity comparison on the ClinGen eRepo benchmark. VarTriage (10 criteria) achieves intermediate pathogenic sensitivity between InterVar (18 criteria) and BIAS-2015 (19 criteria), with the principal performance gap in benign sensitivity (16.9% vs 80.2%). BIAS-2015 and InterVar figures from Eisenhart et al. (2025).