Genetics in Medicine (2025) 27, 101402

ELSEVIER

Genetics in Medicine
An Official Journal of the ACMG

www.journals.elsevier.com/genetics-in-medicine

BRIEF REPORT

Calibration of additional computational tools expands ClinGen recommendation options for variant classification with PP3/BP4 criteria

Timothy Bergquist[1], Sarah L. Stenton[2,3], Emily A.W. Nadeau[4], Alicia B. Byrne[2], Marc S. Greenblatt[4], Steven M. Harrison[2,5], Sean V. Tavtigian[6], Anne O'Donnell-Luria[2,3], Leslie G. Biesecker[7], Predrag Radivojac[8], Steven E. Brenner[9], Vikas Pejaver[1,10,*], ClinGen Sequence Variant Interpretation Working Group

ARTICLE INFO

Article history:
Received 26 September 2024
Received in revised form
26 February 2025
Accepted 4 March 2025
Available online 10 March 2025

Keywords:
ACMG/AMP classification
AlphaMissense
Calibration
ESM1b
VARITY

ABSTRACT

Purpose: We previously developed an approach to calibrate computational tools for clinical variant classification, updating recommendations for the reliable use of variant impact predictors to provide evidence strength up to *Strong*. A new generation of tools using distinctive approaches has since been released, and these methods must be independently calibrated for clinical application.

Methods: Using our local posterior probability-based calibration and our established data set of ClinVar pathogenic and benign variants, we determined the strength of evidence provided by 3 new tools (AlphaMissense, ESM1b, and VARITY) and calibrated scores meeting each evidence strength.

Results: All 3 tools reached the *Strong* level of evidence for variant pathogenicity and *Moderate* for benignity, although sometimes for few variants. Compared with previously recommended tools, these yielded at best only modest improvements in the trade-offs between evidence strength and false-positive predictions.

Conclusion: At calibrated thresholds, 3 new computational predictors provided evidence for variant pathogenicity at similar strength to the 4 previously recommended predictors (and comparable with functional assays for some variants). This calibration broadens the scope of computational tools for application in clinical variant classification. Their new approaches offer promise for future advancement of the field.

The Article Publishing Charge (APC) for this article was paid by Vikas Pejaver.

The names of the ClinGen Sequence Variant Interpretation Working Group members will appear at the end of the article.

*Correspondence and requests for materials should be addressed to Vikas Pejaver, One Gustave L. Levy Pl, Box 1041, New York, NY 10029. *Email address:* vikas.pejaver@mssm.edu

Affiliations are at the end of the document.

doi: https://doi.org/10.1016/j.gim.2025.101402

Introduction

The classification of variants as pathogenic or benign by clinical genetic testing laboratories is a key component of modern genomic medicine. The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) have made recommendations to standardize the practice of clinical variant classification.[1] These recommendations identified distinct sources of evidence regarding the pathogenicity or benignity of a variant (eg, genetic, functional, computational, case observation, and population data), assigned strengths to them, and specified rules to combine evidence to classify a variant into 1 of 5 classes: pathogenic, likely pathogenic, uncertain significance, likely benign, or benign. Rules for multiple evidence types subsequently have been refined to improve classification.[2-4] In the original ACMG/AMP recommendations, the PP3 and BP4 criteria evidence from computational tools (eg, rule-based, statistical, and/or machine-learning-based) was limited to the weakest strength, ie, supporting-level evidence. However, powerful, new variant impact predictors (VIPs) have rapidly emerged, with over 400 now developed.[5]

Recently, we undertook a rigorous quantitative calibration of computational tools, demonstrating that some tools could reliably provide higher levels of evidence strength.[6] Our approach maps scores from a computational tool to local posterior probabilities, which, in turn, correspond to levels of evidential strength in the ACMG/AMP recommendations and points in a Bayesian framework-based adaptation of these recommendations: *Indeterminate* or 0 points, *Supporting* or ±1 point, *Moderate* or ±2 points, *Strong* or ±4 points, and *Very Strong* or ±8 points.[7,8] By applying this approach to 13 tools that predict the impact of missense variation, we demonstrated that at certain score thresholds, 4 tools can provide *Strong* evidence for pathogenicity and *Moderate* evidence for benignity: BayesDel,[9] MutPred2,[10] REVEL,[11] and VEST4.[12] Based on our findings, ClinGen[13] recommended modifications to the PP3 and BP4 criteria that stipulated consistent use of a single tool defined in advance (per laboratory or per gene) with score thresholds calibrated to specific evidential strength levels up to *Moderate* benign (BP4_Moderate; −2 points) and *Strong* pathogenic (PP3_Strong; +4 points). Additional context about these clinical recommendations is provided in Stenton et al,[14] along with practical guidance on their intended use and their implications for variant curation in disease-associated genes.

Since then, advances in protein structure prediction, protein language models, and experiments, such as deep mutational scanning and massively parallel reporter assays, among others, have led to the emergence of new VIPs, with claimed improvements in predictive performance when compared with existing tools.[15-19] However, it is unclear if these improvements in performance translate to the clinical context, in which computational tools serve as one line of evidence for variant pathogenicity/benignity among many. Furthermore, the objectives of these tools may vary, often focusing on the discovery of novel variants in research studies rather than the assertion of clinical pathogenicity and predicting different notions of variant impact, eg, distinguishing unobserved from observed ones. Thus, default score thresholds for these tools do not necessarily correspond to those for strengths of evidence defined by the ACMG/AMP recommendations. Here, we estimate thresholds corresponding to ACMG/AMP evidential strength for newer computational tools, using the same rigorous data sets and approaches. We also estimate additional thresholds for the above 4 previously calibrated tools corresponding to the ACMG/AMP point-based system for variant classification.[8] We then compare and contrast these clinically performant methods with 3 recently published ones. Finally, we discuss our findings in light of the development and use of computational tools in the clinical classification of variants, reiterating the important role that we expect such tools to play in the future.

Materials and Methods

Data sets, calibration procedures, and post hoc analyses

We applied the methods and data sets developed in Pejaver et al.[6] Specifically, we used the *ClinVar 2019* data set for calibration and the *ClinVar 2020* set for post hoc assessments of tools and their thresholds. We used the *gnomAD* data set (v2.1.1) for both calibration and post hoc assessments.[20] We calibrated each tool using our local posterior probability-based approach and estimated score thresholds through bootstrapping with the same parameters and local likelihood ratio cutoffs as before. We adopted the same post hoc assessment pipelines as in the study by Pejaver et al.[6]

Selection of computational tools and data preparation

We selected tools for this study using a purposive sampling strategy. Based on recency of publication (within the past 4 years), the use of modern machine learning approaches (such as protein language models), their performance in the *Annotate All Missense* challenge[21] in the Critical Assessment of Genome Interpretation (CAGI),[22] anecdotal feedback on interest in adoption by the clinical genetics community, and the minimal need for access to original training data, we chose 4 tools for calibration: AlphaMissense,[18] ESM1b,[17] EVE,[15] and VARITY[16] (specifically, VARITY_R, the model trained on only rare variants). Important for this effort and also for utility within the clinical genetics community, these tools make precomputed scores for all possible single-nucleotide or amino acid variants freely and publicly available, albeit in slightly different formats and with gene/protein identifiers from different

databases. To address these differences, we developed customized mapping protocols for each tool to maximize the number of variants in our data sets with scores, leaving 3 tools with sufficient data for calibration (Supplemental Materials and Methods). Except for VARITY, none of these tools were explicitly trained on variants from ClinVar.[23] However, for VARITY, the precomputed score for each variant was assigned by a version of the model that did not include that variant in the training set. Therefore, no additional filtering of the data sets against the training data set of each tool was performed.

Results

Recently published tools can provide up to strong evidence for pathogenicity

Our local posterior probability-based calibration approach enabled the estimation of score thresholds for AlphaMissense, ESM1b, and VARITY_R that corresponded to distinct evidential strength levels within the ACMG/AMP variant classification guidelines. We found that all 3 tools were able to reach at least the *Moderate* level for benignity, with VARITY_R reaching *Strong* (BP4), and the *Strong* level of evidence for pathogenicity (PP3) (Table 1, Figure 1A). However, the score thresholds at which these were achieved were more stringent than the thresholds recommended by the tool developers. In fact, the recommended thresholds for AlphaMissense (0.564) and ESM1b (−7.5) do not meet the *Supporting* level of evidence for pathogenicity or benignity based on our calibration. Overall, all 3 tools exhibited similar behavior to the 4 best-performing tools from our previous study, even when considering newer intervals between *Moderate* and *Strong* according to the ACMG/AMP point-based system (Table 1). When we attempted to calibrate EVE, it nominally appeared to reach the *Moderate* level of evidential strength for both pathogenicity and benignity. Score thresholds for *Supporting* and *Moderate* were 0.684 and 0.845, respectively, for pathogenicity, and 0.137 and 0.209, respectively, for benignity. However, EVE predictions were available only for a subset of genes in our calibration set, leaving about half of the benign/likely benign variants unscored. Furthermore, unscored genes showed a marked skew in ratio of pathogenic to benign variants. Because of potential sampling bias, we lack confidence in the applicability of the measured thresholds, rendering us currently unable to recommend their use in clinical variant classification.

Clinical calibration shows modest improvements over existing computational predictors

We assessed the validity of our calibration by using the score thresholds estimated in Table 1 to group variants from the *ClinVar 2020* (not used in calibration) and *gnomAD* data

Table 1 Estimated threshold intervals for all tools in this and our previous study according to the ACMG/AMP recommendations for sequence variant interpretation

	Benign (BP4)					Pathogenic (PP3)			
Method	Strong (−4)	(−3)	Moderate (−2)	Supporting (−1)	Indeterminate (0)	Supporting (+1)	Moderate (+2)	(+3)	Strong (+4)
BayesDel[a]	-	≤ −0.520	[−0.519, −0.360]	[−0.359, −0.180]	[−0.179, 0.129]	[0.130, 0.269]	[0.270, 0.409]	[0.410, 0.499]	≥0.500
MutPred2[a]	≤0.010	[0.011, 0.031]	[0.032, 0.197]	[0.198, 0.391]	[0.392, 0.736]	[0.737, 0.828]	[0.829, 0.894]	[0.895, 0.931]	≥0.932
REVEL[a]	≤0.016	[0.017, 0.052]	[0.053, 0.183]	[0.184, 0.290]	[0.291, 0.643]	[0.644, 0.772]	[0.773, 0.878]	[0.879, 0.931]	≥0.932
VEST4[a]	-	≤0.077	[0.078, 0.302]	[0.303, 0.449]	[0.450, 0.763]	[0.764, 0.860]	[0.861, 0.908]	[0.909, 0.964]	≥0.965
AlphaMissense[b]	-	≤0.070	[0.071, 0.099]	[0.100, 0.169]	[0.170, 0.791]	[0.792, 0.905]	[0.906, 0.971]	[0.972, 0.989]	≥0.990
ESM1b[b]	-	≥8.8	[−3.1, 8.7]	[−6.3, −3.2]	[−10.6, −6.4]	[−12.1, −10.7]	[−13.9, −12.2]	[−23.9, −14.0]	≤ −24.0
VARITY_R[b]	≤0.036	[0.037, 0.063]	[0.064, 0.116]	[0.117, 0.251]	[0.252, 0.674]	[0.675, 0.841]	[0.842, 0.914]	[0.915, 0.964]	≥0.965

The intervals correspond to 3 pathogenic, 1 indeterminate, and 3 benign intervals (*Very Strong* not shown because it was never reached) in the current guidelines. The ACMG/AMP guidelines are expected to transition to a point-based system,[5] and the numbers in parentheses in the header indicate point values corresponding to each evidential strength interval in this system. Although the 2015 guidelines do not include a strength level between *Moderate* (2 points) and *Strong* (4 points), intervals for the 3-point strength of evidence are also reported because 3-point evidence will be recommended for future editions of the guidelines. A "–" implies that the given tool did not meet the likelihood ratio (posterior probability; Supplemental Materials and Methods) threshold for that strength.

ACMG, American College of Medical Genetics and Genomics; *AMP*, Association for Molecular Pathology.

[a]All intervals are the same as those reported in our previous study,[3] with additional columns for the interval corresponding to the indeterminate range and ±3 points as per the point-based system.

[b]Calibrated in this study.

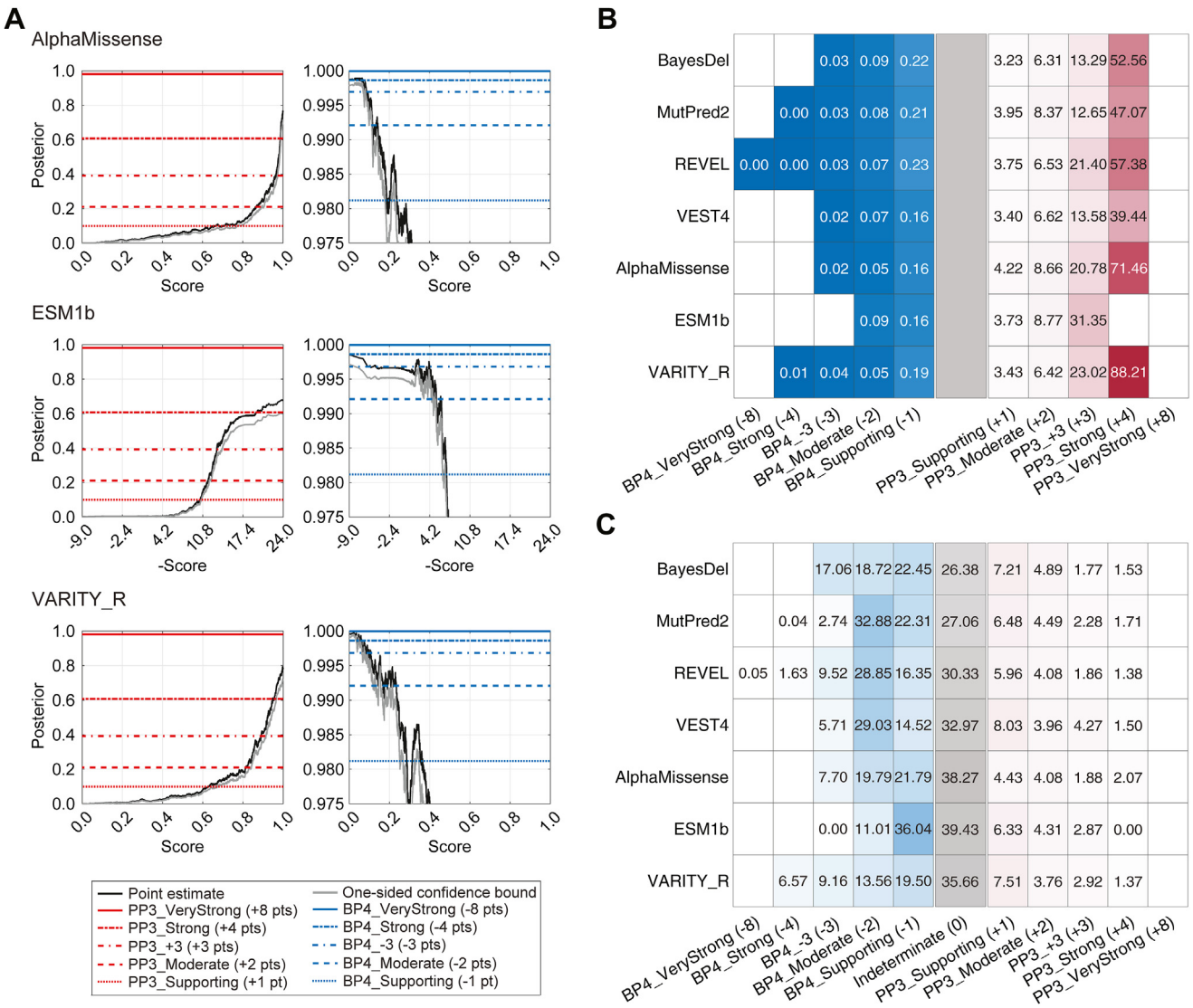

Figure 1 Local posterior probability curves and comparison with previously calibrated tools. A. Pairs of curves for AlphaMissense, ESM1b and VARITY_R. For each tool, the curve on the left is for pathogenicity (red horizontal lines), and the curve on the right is for benignity (blue horizontal lines). The horizontal lines represent the posterior probability thresholds for *Supporting*, *Moderate*, *Strong*, and *Very Strong* evidence as per current ACMG/AMP guidelines. A horizontal line representing the 3-point strength of evidence is also shown. The black curves represent the posterior probability estimated from the ClinVar 2019 set. The gray curves represent 1-sided 95% CIs (in the direction of more stringent thresholds), calculated from 10,000 bootstrap samples of this data set. The points at which the gray curves intersect the horizontal lines represent the thresholds for the relevant intervals. B. The likelihood ratios within each interval on the independent ClinVar 2020 set. Darker colors indicate higher values for pathogenicity and lower values for benignity (because these are positive likelihood ratios). The limits for the color gradients are asymmetric, with ranges set between 0 and 1 for benignity, and 1 and 100 for pathogenicity. A gray rectangle is introduced at the center for comparability with (C). C. The percentage of variants predicted to be within the interval in the gnomAD set. Blue and red distinguish the evidential strength intervals for benignity from pathogenicity, respectively, with the indeterminate interval colored gray. The color gradient corresponds to the value in the cells, regardless of color. Darker colors indicate higher proportions. A white cell without a value indicates that the tool did not reach thresholds corresponding to that interval. The indeterminate interval also included variants without any scores. ACMG, American College of Medical Genetics and Genomics; AMP, Association for Molecular Pathology.

sets, while also comparing them with the 4 previously calibrated tools (Figure 1B and C). For the *ClinVar 2020* set, we calculated likelihood ratios within each interval defined by these thresholds, reflective of true- and false-positive rates for the classification of pathogenic variants. All tools met or exceeded (or, for benignity, were less than) the expected likelihood ratio values corresponding to each interval. The only exception to this was that some of the previously calibrated tools did not meet the thresholds for the 3-point intervals (Figure 1B). VARITY_R and

AlphaMissense resulted in higher likelihood ratios in the interval corresponding to *Strong* for PP3 than the 4 previously calibrated tools. However, it is unclear to what extent this is driven by the small number of variants in this interval relative to other intervals. No variant in the *ClinVar 2020* set received an ESM1b score of −24.0, effectively capping the maximal strength for benignity achieved by ESM1b at *Moderate*. For the gnomAD set, we calculated the proportion of variants within each interval to assess how evidential strength is distributed for each tool in variants from the population (Figure 1C). VARITY_R and AlphaMissense behaved as expected, in a manner similar to the 4 previously calibrated tools, with the proportion of variants in the *Strong* interval for pathogenicity being within the estimated prior probability of pathogenicity (0.0441). However, AlphaMissense classified the smallest proportion of variants as being within all 3 pathogenic intervals (0.125), slightly lower than REVEL (0.133). It is unclear if this results from AlphaMissense being trained on variants from gnomAD as a proxy for nonpathogenic variants.

Discussion

In this study, we calibrated 3 recently published computational tools to align with the ACMG/AMP guidelines for clinical variant classification. These tools reached clinically useful evidential strength levels. However, in all cases, their recommended (default) thresholds did not meet even the *Supporting* level of evidence for variant pathogenicity. These tools largely behaved similarly to 4 tools that we previously calibrated and at best offer modest improvements in the strength of evidence that can be applied while minimizing the number of false-positive predictions in the *Supporting* and *Moderate* categories. We extended our previous study to include intervals corresponding to 3 classification points, anticipating the use of the point-based system to weight evidence in the next version of the ACMG/AMP standards. This calibration shares the limitations of our previous study, including those related to the representativeness of data, potential circularities, prior probability estimation, and gene- or disease-specific applicability and variability.[6,24] We did not calibrate methods that incorporate allele frequency (AF) as an explicit or strong implicit feature for 2 reasons. First, the use of a predictor incorporating AF will limit use of evidence codes dependent upon AF in variant classification, such as BA1. Such methods would be impractical to use in most clinical classification pipelines. Second, methods using AF require distinct calibrations for different AF thresholds (or once for the most stringent AF group), for which we currently lack sufficient data.[21]

The development of more advanced computational VIPs has often been motivated by the idea that no computational method can yet "be relied on alone for genetic diagnosis."[25] However, this is an inappropriate and unachievable benchmark for utility because no single source of evidence other than high AF—computational or otherwise—can presently be the sole criterion to determine the role of a variant in disease. Clinical standards for the classification of rare genetic variants always require the integration of multiple lines of evidence, a principle fundamental to the ACMG/AMP clinical classification framework.[1] Consequently, the term "likely pathogenic" in AlphaMissense's claim of classifying "32% of all missense variants as likely pathogenic" is inconsistent with terminology used in clinical variant classification.

Historically, computational tools have been trained or calibrated to predict various proxies for variant pathogenicity that do not necessarily meet these clinical standards. Consequently, their utility in clinical variant classification was initially limited to providing *Supporting* evidence. Our calibration provides a means to reconcile this misalignment of developer and clinical perspectives by providing data-driven, tool-specific guidance on use in clinical variant classification. We found that the AlphaMissense and ESM1b developers' proposed thresholds did not reach the threshold for supporting-level evidence; higher thresholds were necessary to reach *Supporting*. At even higher thresholds, AlphaMissense and VARITY_R can reach *Moderate* and *Strong* pathogenicity evidence for some variants. This underscores the importance of independent calibration of tools used in clinical variant classification, just as critical assessments (such as Critical Assessment Structure Prediction[26] and CAGI[22]) have revealed how developers' subtle knowledge of their methods and data inadvertently influence the results of their own assessments. Together with the ability to provide *Supporting* and *Moderate* benign evidence, we recommend these calibrated tools as potential alternatives alongside the previously recommended tools.

Our results continue to suggest increasingly important roles for computational predictors of variant impact in interpreting genomic data for clinical diagnosis and screening. The initial releases of this new generation of tools performed comparably to the best predecessors, suggesting potential for their future improvement. Moreover, these distinct approaches may offer independent information valuable for meta-predictors. Relative to most other lines of evidence, computational tools have an outsized role because they can be readily applied to every relevant genomic variant. The continued development of enhanced in silico variant impact prediction methods augurs promising advances in clinical variant classification.

Data Availability

Data sets described in the Materials and Methods are available on Zenodo: https://zenodo.org/records/13766399; intermediate result files and code to calculate local posterior probabilities, estimate thresholds, and plot figures in the paper are available here: https://github.com/pejaverlab/clingen-svi-comp_calibration. Machine-parsable calibration thresholds are available in VIPdb: https://genomeinterpretation.org/vipdb

Acknowledgments

The authors thank the developers of all the methods calibrated here for making their prediction scores, tools, and/or training set information available. This work has been posted as a preprint at https://doi.org/10.1101/2024.09.17.611902.

Funding

This work was supported by the National Institutes of Health grants R00LM012992, R01HG013350, and U01HG012022. This work was also supported in part through the computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards grant UL1TR004419 from the National Center for Advancing Translational Sciences. Research reported in this publication was also supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD026880 and S10OD030463. L.G.B. was supported by HG200388-10 and HG200387-10 and A.O'D.-L. by U01HG011755. Entry of data into VIPdb is supported by U24HG007346. This work was conducted as part of the ClinGen Sequence Variant Interpretation Working Group. ClinGen is primarily funded by the National Human Genome Research Institute with cofunding from the National Cancer Institute, through the following grants: U24HG009649 (to Baylor/Stanford), U24HG006834 (to Broad/Geisinger), and U24HG009650 (to UNC/Kaiser). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

Conceptualization: V.P., P.R., S.E.B.; Data Curation: T.B., V.P.; Formal Analysis: T.B., V.P.; Interpretation, discussion of results, and oversight: M.S.G., S.M.H., S.V.T., A.O.-L., L.G.B., P.R., S.E.B., V.P.; Writing-original draft: T.B., V.P.; Critical evaluation of manuscript drafts, writing-review and editing: S.L.S., E.A.W.N., A.B.B., M.S.G., S.M.H., S.V.T., A.O.-L., L.G.B., P.R., S.E.B., V.P.

Ethics Declaration

This work does not report a clinical study or experiment with human subjects.

Conflict of Interest

Leslie G. Biesecker is a member of the Illumina Medical Ethics Committee and receives research support from Merck, Inc and royalties from Wolters-Kluwer. Vikas Pejaver and Predrag Radivojac participated in the development of some of the tools assessed in this study. Anne O'Donnell-Luria receives research support from PacBio and is a consultant for Addition Therapeutics and on the SAB for Congenica Inc. All other authors declare no conflicts of interest.

Members of the ClinGen Sequence Variant Interpretation Working Group

Ahmad A. Tayoun, Anne O'Donnell-Luria, Garry R. Cutting, Heidi L. Rehm, Izabela Karbassi, Jessica Mester, Jonathan S. Berg, Leslie G. Biesecker, Marc S. Greenblatt, Natasha T. Strande, Peter Kang, Rachel Karchin, Scott Topper, Sean V. Tavtigian, Sharon E. Plon, Sian Ellard, Steven E. Brenner, Steven M. Harrison, and Tina Pesaran.

Additional Information

The online version of this article (https://doi.org/10.1016/j.gim.2025.101402) contains supplemental material.

Affiliations

[1]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY; [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA; [3]Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA; [4]Department of Medicine and University of Vermont Cancer Center, University of Vermont, Larner College of Medicine, Burlington, VT; [5]Ambry Genetics, Aliso Viejo, CA; [6]Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT; [7]Center for Precision Health Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; [8]Khoury College of Computer Sciences, Northeastern University, Boston, MA; [9]Department of Plant and Microbial Biology and Center for Computational Biology, University of California, Berkeley, CA; [10]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY

References

1. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-424. http://doi.org/10.1038/gim.2015.30

2. Biesecker LG, Harrison SM, ClinGen Sequence Variant Interpretation Working Group. The ACMG/AMP reputable source criteria for the interpretation of sequence variants. *Genet Med*. 2018;20(12):1687-1688. http://doi.org/10.1038/gim.2018.42
3. Harrison SM, Biesecker LG, Rehm HL. Overview of specifications to the ACMG/AMP variant interpretation guidelines. *Curr Protoc Hum Genet*. 2019;103(1):e93. http://doi.org/10.1002/cphg.93
4. Brnich SE, Abou Tayoun AN, Couch FJ, et al. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med*. 2019;12(1):3. http://doi.org/10.1186/s13073-019-0690-2
5. Lin YJ, Menon AS, Hu Z, Brenner SE. Variant Impact Predictor database (VIPdb), version 2: trends from three decades of genetic variant impact predictors. *Hum Genomics*. 2024;18(1):90. http://doi.org/10.1186/s40246-024-00663-z
6. Pejaver V, Byrne AB, Feng BJ, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet*. 2022;109(12):2163-2177. http://doi.org/10.1016/j.ajhg.2022.10.013
7. Tavtigian SV, Greenblatt MS, Harrison SM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med*. 2018;20(9):1054-1060. http://doi.org/10.1038/gim.2017.210
8. Tavtigian SV, Harrison SM, Boucher KM, Biesecker LG. Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. *Hum Mutat*. 2020;41(10):1734-1737. http://doi.org/10.1002/humu.24088
9. Feng BJ. PERCH: a unified framework for disease gene prioritization. *Hum Mutat*. 2017;38(3):243-251. http://doi.org/10.1002/humu.23158
10. Pejaver V, Urresti J, Lugo-Martinez J, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat Commun*. 2020;11(1):5918. http://doi.org/10.1038/s41467-020-19669-x
11. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet*. 2016;99(4):877-885. http://doi.org/10.1016/j.ajhg.2016.08.016
12. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. *BMC Genomics*. 2013;14(suppl 3):S3. http://doi.org/10.1186/1471-2164-14-S3-S3
13. Rehm HL, Berg JS, Brooks LD, et al. ClinGen — the Clinical Genome Resource. *N Engl J Med*. 2015;372(23):2235-2242. http://doi.org/10.1056/NEJMsr1406261
14. Stenton SL, Pejaver V, Bergquist T, et al. Assessment of the evidence yield for the calibrated PP3/BP4 computational recommendations. *Genet Med*. 2024;26(11):101213. http://doi.org/10.1016/j.gim.2024.101213
15. Frazer J, Notin P, Dias M, et al. Disease variant prediction with deep generative models of evolutionary data. *Nature*. 2021;599(7883):91-95. http://doi.org/10.1038/s41586-021-04043-8
16. Wu Y, Li R, Sun S, Weile J, Roth FP. Improved pathogenicity prediction for rare human missense variants. *Am J Hum Genet*. 2021;108(10):1891-1906. http://doi.org/10.1016/j.ajhg.2021.08.012
17. Brandes N, Goldman G, Wang CH, Ye CJ, Ntranos V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet*. 2023;55(9):1512-1522. http://doi.org/10.1038/s41588-023-01465-0
18. Cheng J, Novati G, Pan J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*. 2023;381(6664):eadg7492. http://doi.org/10.1126/science.adg7492
19. IGVF Consortium. Deciphering the impact of genomic variation on function. *Nature*. 2024;633(8028):47-57. http://doi.org/10.1038/s41586-024-07510-0
20. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443. http://doi.org/10.1038/s41586-020-2308-7
21. Rastogi R, Chung R, Li S, et al. Critical assessment of missense variant effect predictors on disease-relevant variant data. *Hum Genet*. 2025. http://doi.org/10.1007/s00439-025-02732-2
22. Critical Assessment of Genome Interpretation Consortium. CAGI, the Critical Assessment of Genome Interpretation, establishes progress and prospects for computational genetic variant interpretation methods. *Genome Biol*. 2024;25(1):53. http://doi.org/10.1186/s13059-023-03113-6
23. Landrum MJ, Chitipiralla S, Brown GR, et al. ClinVar: improvements to accessing data. *Nucleic Acids Res*. 2020;48(D1):D835-D844. http://doi.org/10.1093/nar/gkz972
24. Tejura M, Fayer S, McEwen AE, Flynn J, Starita LM, Fowler DM. Calibration of variant effect predictors on genome-wide data masks heterogeneous performance across genes. *Am J Hum Genet*. 2024;111(9):2031-2043. http://doi.org/10.1016/j.ajhg.2024.07.018
25. Marsh JA, Teichmann SA. Predicting pathogenic protein variants. *Science*. 2023;381(6664):1284-1285. http://doi.org/10.1126/science.adj8672
26. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XV. *Proteins*. 2023;91(12):1539-1549. http://doi.org/10.1002/prot.26617