# CLINICAL likelihood ratios and balanced accuracy for 44 in silico tools AGAINST multiple large-scale functional assays of cancer susceptibility genes

C. Cubuk1\*, A. Garrett1\*, S. Choi\*1, L. King1, C. Loveday1, B. Torr1, G.J. Burghel2 ,  M. Durkie3, A. Callaway4, 5, R. Robinson6, J. Drummond7, I. Berry6, A. Wallace2, D. Eccles 5,8, M. Tischkowitz ,7,19, N. Whiffin 10,11,12, J.S. Ware 11,12, H. Hanson1, 13, C. Turnbull1, 14 and CanVIG-UK

\*these authors contributed equally to the work

1 Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, UK

2 Manchester Centre for Genomic Medicine and NW Laboratory Genetics Hub, Manchester University NHS Foundation Trust, Manchester, UK

3 Sheffield Diagnostic Genetics Service, Sheffield Children's NHS Foundation Trust, Sheffield, UK

4 Wessex Regional Genetics Laboratory, Salisbury Hospital NHS Foundation Trust, Salisbury, UK

5 Human Genetics and Genomic Medicine, Faculty of Medicine, University of Southampton, Southampton, UK

6 Yorkshire and North East Genomic Laboratory Hub, Leeds Teaching Hospitals NHS Trust, Leeds, UK

7 East Genomic Laboratory Hub, Cambridge University Hospitals Genomic Laboratory, Cambridge University Hospitals, Cambridge, UK

8 Cancer Sciences, Faculty of Medicine, University of Southampton, Southampton, UK

9 Department of Medical Genetics, National Institute for Health, Research Cambridge Biomedical Research Centre, University of Cambridge, Cambridge, UK

10 Cardiovascular Genetics and Genomics, National Heart and Lung Institute, Imperial College London, London, UK

11 NIHR Royal Brompton Cardiovascular Biomedical Research Unit, Royal Brompton & Harefield Hospitals & Imperial College London, London, UK;

12 The Wellcome Centre for Human Genetics, Oxford, UK

13 Department of Clinical Genetics, St. George's University Hospitals NHS Foundation Trust, London, UK

14 Cancer Genetics Unit, Royal Marsden NHS Foundation Trust, London, UK

Correspondence to Professor Clare Turnbull, Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton SM2 5NG, UK; [clare.turnbull@icr.ac.uk](mailto:clare.turnbull@icr.ac.uk)

## Word Count:

Abstract:191

Manuscript: 3976

## Abstract

### Purpose

For concordancy of multiple in silico tools, the American College of Medical Genetics (ACMG)/ AMP (Association of Molecular Pathology) framework affords ‘supporting’ evidence towards pathogenicity or benignity, equivalent to a likelihood ratio of ~2. However, tool performance may be over-estimated using clinical truth-sets also utilised in training of tools.

### Methods

We created a truth-set of 9436 missense variants classified as deleterious or tolerated in clinically-validated high-throughput functional assays for *BRCA1*, *BRCA2*, *MSH2*, *PTEN* and *TP53* to evaluate predictive performance for 44 recommended/commonly-used in silico tools.

### Results

Over two thirds of the tool-threshold-combinations examined had specificity of <50%, thus substantially over-calling deleteriousness. Revel scores of 0.8-1.0 had a PLR of 6.74 (5.24-8.82) compared to scores <0.7 and scores of 0-0.4 had a NLR of 34.3 (31.5-37.3) compared to scores of >0.7. For Meta-SNP, the equivalent PLR=42.9 (14.4-406) and NLR=19.4 (15.6-24.9).

### Conclusions

Against this clinically-validated truth-set, not previously used for tool training, there was wide variation across between commonly-used in silico tools in predictive perfromance. Overall, Revel and Meta-SNP had best balanced accuracy and might potentially be used at stronger evidence weighting than current ACMG/AMP prescription, in particular for predictions of benignity.

## Introduction

### Variant Interpretation

For more than three decades, sequence analysis of constitutional DNA has informed diagnosis and prediction of human Mendelian diseases. Robust identification of the causative pathogenic variant enables accurate prediction of the clinical course of disease and implementation of measures for prevention and early detection. Through technological advances, clinical genome sequencing is now routine, typically revealing in comparison to a reference genome in excess of 4 million variants in the average human1. Through concerted efforts within the clinical community to reduce erroneous assignation of variants as pathogenic, common frameworks for variant interpretation have been evolved, such as that of the American College of Medical Genetics/Association of Molecular Pathology (ACMG/AMP). Within this system, points are tallied up from quasi-orthogonal lines of evidence, such as clinical case series, segregation data, phenotypic specificity and laboratory assays2.

### Emergence of in silico tools

However, rare missense variants are frequently identified on clinical genetic testing that have not previously been reported or for which existing clinical and laboratory data are sparse. In these scenarios, evaluation of variant pathogenicity/benignity is largely reliant on predicted alteration of protein function using features such as:

**Homology in sequence alignment between divergent species.** Orthologs are gene sequences derived from the same ancestral gene present in two species' last common ancestor. Where an amino acid is highly conserved across multiple orthologs, this indicates that a change in that amino acid will be of deleterious consequence for protein function.

**Physiochemical differences between amino acids.** Amino acids are characterised by their composition, polarity, and molecular volume. A large ‘physiochemical distance’ for a substitution would be predicted to have a greater impact on protein function3.

**Disruption to 3D protein structure.** Amino acid substitutions are more likely to be deleterious if they alter tertiary protein structure, including folding, bonds and binding site shapes4. Whilst protein structure can be visualised directly using X-ray crystallography and nuclear magnetic resonance (NMR), most predictions are largely on modelling.

Over the last twenty years, computational biologists have developed a number of in silico prediction algorithms or “tools” variously leveraging these features. In addition to knowledge of biological principles, in silico tools may be trained against a “truth-set” in which impact of variants on protein function is already quantified5. The performance of the tool may then be “validated” or “tested” against other independent “truth-sets”. Early training truth-sets typically included prokaryote assays of broad cell function: given the divergence between humans and yeast and the complex cellular functions of many human disease-associated genes, such predictions are likely to relate only loosely to clinical pathogenicity of human disease genes6. However, more recently, large-scale databases of clinical and neutral population variation have been made publicly available, such as ClinVar, Human Gene Mutation Database (HGMD), SwissProt and ExAC/gnomAD7-10.

There has been a surge in release of new tools and ‘meta-predictors’ (tool combinations) largely trained on these same datasets. With multiple elements incorporated into sophisticated machine learning algorithms, training on restricted datasets has potential to result in overfitting, that is recognition of features present within the training dataset which are due to random variation, rather than those which are useful for prediction in new datasets11-13. In the case of ‘meta-predictors’, this may result in the constituent tools appearing to perform better and therefore being allocated excess weight within the overall agorithm14. The field has been further confused by inter-tool comparisons using the same datasets upon which they were trained11

In addition, some ‘clinical’ databases may offer less reliable variant classifications: earlier instances of HGMD, for example, largely assigned as pathogenic any variant detected in an individual with the relevant phenotype, whilst only more recently has ClinVar curation tackled erroneous community submissions15,16. Such databases may also be biased towards variants with features which are more easily detectable by prediction software, as conclusive clinical classifications are less frequently established for more “challenging” variants11. Furthermore, differences in tools performance have been observed across different populations and between different variant types (gain of function vs loss of function and pathogenic vs benign)17-19.

We summarise the data inputs, methodologies and testing/training data underpinning 45 in silico tools widely used in clinical practice and/or available from amalgamation sites 20,21 (**Table 1, Supplementary Table 1**).

### Clinical application of in silico predictions

Whilst primarily developed to support genomic research, in silico predictions have also been widely used by clinical diagnostic laboratories to supplement clinical data for variant classification. However, as the substantial discordancy between tools and high rates of false positive predictions has become more evident, greater caution has been applied. Indeed, in the 2015 ACMG/AMP framework for variant interpretation, it is recommended that evidence from in silico tools should only be used when multiple tools are concordant and only to provide ‘supporting’-level evidence towards assessment of pathogenicity or benignity2. Using a quantitative Bayesian translation of the ACMG/AMP framework, supplementary evidence equates to a likelihood ratio of only 2-fold22. Tools most widely used clinically include PolyPhen2, SIFT and MutationTaster, due in large part to their inclusion within commercially-developed interfaces 2,4,6,23.

### Large-scale assays of cancer susceptibility gene function

Reliable assays of cellular function that correlate well with clinical pathogenicity have long been awaited by those working in genetic variation interpretation. The majority of early published experimental assays feature only a handful of variants, have been conducted in a post-hoc and/or piecemeal fashion and often fail on reproducibility. Leveraging improved capability in gene editing technology, data from robust, systematic, high-throughput saturation genome editing experiments have recently become available for key cancer susceptibility genes, which have been shown to correlate strongly with well-curated orthogonally-generated clinical classifications (Supplementary Tables 2, 3, 4).

These ‘new-generation’, clinically-validated functional assays provide large ‘fresh’ truth-sets for unbiased evaluation of in silico tools. We thus sought to evaluate against functional assays of *BRCA1*, *BRCA2, MSH2, PTEN* and *TP53* individually and in combination, predictive performance for 45 widely-used in silico tools (72 tool-threshold-combinations).

## Materials and Methods

### Generation of functional-truth-sets of BRCA1, BRCA2, MSH2, PTEN and TP53 variants

For *BRCA1,* we used data on 2321 non-synonymous variants generated by Findlay et al in which *BRCA1* function was assessed via assay of cellular fitness of HAP1 for the 13 exons comprising the RING and BRCT functional domains generated via saturation genome editing24. For *BRCA2* function, we used data generated by Couch et al, who performed a homology-directed DNA break repair (HDR) assay in *BRCA2*-deﬁcient cells, assessing 237 variants in the *BRCA2* DNA binding domain introduced via site-directed mutagenesis 25-28. For *MSH2* we used data for 5212 single-base-substitution variants introduced by saturation mutagenesis from HAP1 survival following treatment with 6-TG, which induces lesions unrepairable by the MMR machinery29. For *PTEN* we integrated data for 7244 variants generated on phosphatase activity in an artificial humanised yeast model with data from Variant Abundance by Massively Parallel Sequencing (VAMP-seq) in which *PTEN* protein expression in a human cell line was quantified for 4112 *PTEN* variants, from which 2380 non-synonymous variants overlapped with the phosphatase activity data30,31. As per specification of the ClinGen*TP53* expert group for clinical variant classification, we integrated data from (i) yeast-based transactivation assays performed eight-fold for variants introduced by site-directed mutagenesis and (ii) survival of isogenic *TP53*-wild-type and *TP53*-null cell populations treated with Nutlin-3 and/or etoposide for variants generated using Mutagenesis by Integrated TilEs (MITE), from which there were 2314 overlapping variants32-35.

Each gene-specific functional-truth-set was curated to include only missense variants, described in accordance to HGVS nomenclature for GRCh37 transcripts ENST00000357654 (*BRCA1*), ENST00000380152 (*BRCA2*), ENST00000233146 (*MSH2)*, ENST00000371953 (*PTEN*), and ENST00000269305 (*TP53*). The potentially spliceogenic exonic variants at the two bases flanking the intron-exon boundary were also excluded. Missense variants were classified as non-functional (deleterious, DEL) or functional (tolerated, TOL) in accordance with functional assay specifications (**Supplementary Table 2**). Variants with results discordant between constituent assays (*PTEN* and *TP53)* were excluded from the functional-truth-sets, as were variants with intermediate assay activity for *BRCA1*, *BRCA2* and *MSH2*. Of 12623 non-synonymous variants for which assay data were available, 11212 were missense in suitable regions, of which 9436 gave results of deleterious/tolerated (1641 in *BRCA1*, 188 in *BRCA2*, 4783 in *MSH2*, 957 in *PTEN* and 1867 in *TP53*). 1413 variants were non-functional (deleterious) and 8023 functional (tolerated) (**Supplementary Table 3**).

### Generation of ClinVar-truth-sets of BRCA1, BRCA2, MSH2, PTEN and TP53 variants

We also assembled available ClinVar classifications for these 9436 missense variants, retaining those with ClinVar classifications of ‘pathogenic’/‘likely pathogenic’ (267 variants) or ‘benign’/’likely benign’ (≥1 star rating) (66 variants). These were assigned in the ‘ClinVar-truth-set’ as deleterious (DEL) and tolerated (TOL) respectively (Supplementary Table 4).

Tool evaluations were primarily focused on the functional-truth-sets, as many of the tools had been trained/evaluated using ClinVar data and/or tool predictions constituting part of the ClinVar classification.

### In silico Tools

45 in silico tools were selected on the basis of inclusion in publicly available/commercial variant-interpretation resources and/or reported use in clinical diagnostics (Table 1, Supplementary Table 1)2,20,21. The parameters/thresholds for tool predictions as deleterious (DEL) or tolerated (TOL) were based on default author recommendations (Supplementary Table 5). Where there was variation from default author recommended settings reported in the literature or commonly used in practice, additional tool-threshold-combinations were included, resulting in 72 in total. For example, for Revel we specified three tool-threshold-combinations: Revel\_a: <0.4 predicted-TOL; >0.7 predicted-DEL, Revel\_b: <0.7 predicted-TOL; ≥0.7 Predicted-DEL, Revel\_c: <0.5 predicted-TOL; ≥0.5 predicted-DEL. 57/72 tool-threshold-combinations involved binary categorisation above or below a cut-off; 15/72 were non-binary, involving exclusion of an indeterminate scoring set of variants. We excluded from subsequent analysis tool-threshold-combinations for which predictions (i) produced no discrimination (one exclusion: Integrated\_fitCons\_b: all calls deleterious) (ii) were generated for <25% of variants examined (one exclusion: SNP3D (calls for <3% of variants))). Following exclusions, we examined 70 tool-threshold-combinations in total representing 44 tools. We also examined under a ‘full concordancy model’ (i.e. discordant calls were excluded) (i) pair-wise combination twelve of the tools with best balanced accuracy and (ii), three-way combination of (a) SIFT, Polyphen2 and MutationTaster and (b) Rebel b, P Mut and rfPred (**Supplementary Table 5**).

### Statistical Analysis

Tool predictions were generated as per resources/versions specified in Supplementary Table 5. These predictions were compared to five gene-specific functional-truth-sets, the combined-functional-truth-set of 9436 variants (ALL) and the ClinVar-truth-set of 333 variants. For each of the 70 tool-threshold-combination, predictions of DEL, TOL or missing were generated for each of the 9436 missense variants. Missing predictions resulted in diminution of the total number of predictions where (i) the tool failed to make a prediction for the variant (indeterminate) (ii) the prediction lay in the range between the defined thresholds for TOL or DEL (indeterminate or VUS, e.g. prediction range 0.4-0.7 for Revel\_a). Each prediction was assigned True Positive (TP) where predicted-DEL and classified DEL in the truth-set, True Negative (TN) where predicted-TOL and classified TOL in the truth-set, False Positive (FP) where predicted-DEL and classified TOL in truth-set, or False Negative (FN) where predicted-TOL and classified DEL in truth-set recall (Supplementary Table 6). For each functional-truth-set, the overall prevalence, detection prevalence, sensitivity (recall), specificity, positive predictive value (PPV, precision) and negative predictive value (NPV) were calculated. Balanced accuracy (BA), which combines sensitivity and specificity, was presented as the primary pan-performance metric36. We also calculated the Matthews correlation coefficient (MCC), which combines TP, TN, FP and FN, the area under the curve (AUC) and the F1, which combines precision and recall (Supplementary Tables 7, 8)37. To adjust for differing contribution of the five gene-specific functional-truth-sets, the mean of the five outputs was also generated.

Positive likelihood ratios were generated comparing values above the threshold to those below; negative likelihood ratios generated by comparing values below the threshold to those above (Supplementary Table 9). For REVEL and MetaSNP, we undertook a ‘banded’ analysis, examining the positive likelihood ratios for pathogenicity for various scoring bands above 0.7 against tool prediction<0.7; for positive likelihood ratios for benignity (negative likelihood ratios for pathogenicity), we examined various scoring bands below 0.7 and compared each to tool prediction>0.738. Where zero fields precluded generation of a PLR/NLR, we performed a Haldane correction (addition of 0.5 to each cell) (Supplementary Tables 10, 11).

To determine the optimal cut-off value for each tool, we used as the reference metric balanced accuracy (BA) calculated using the dichotomized scores, and iterated in 2% intervals from the lowest value. The optimisation process was terminated when the tested cut-off value became higher than the maximum variant effect score of the tool evaluated. We then took the mean of the optimised thresholds for the five functional truthsets. We evaluated BA against this mean threshold for the six truthsets (five gene-specific and the combined-functional-truth-set (ALL)).

Analyses were performed using R v.3.6.2 and STATAv15 (Timberlake Analytics).

## Results

### Variant Inclusion

We used 9436 variants in the combined-functional-truth-set which overall has a sensitivity of 0.91 and specificity of 0.95 for ClinVar calls (Supplementary Table 4). We included 70 tool-thresholds representing 44 tools. Of these, 9/70 tool-threshold-combinations generated predictions for <80% of the variants (variant inclusion, Supplementary Tables 8). For example, for Revel\_a, the threshold setting recommended for clinical application in the UK Association for Clinical Genomic Science (UK-ACGS) guidance, scores of <0.4 are predicted as tolerated and scores of >0.7 are predicted as deleterious. For Revel\_a, 42.4% of the 9436 variants score 0.4-0.7 such that they are indeterminate and not classified39,40. Notably, within this indeterminate Revel range, whilst the true-positives cluster towards the higher end, the distribution for the true-negatives is relatively even (**Supplementary Figure 1**).

### Overall performance

The true prevalence of deleterious variants in the combined-functional-truth-set was 15% (1413/9436). However, the detection prevalence (i.e. the total proportion ***called*** by the in silico tool as deleterious) was >50% for 56/70, >75% for 28/70 and >90% for 11/70 of the tool-threshold-combinations. Thus, whilst sensitivity was generally high (>80% for 56/70 tool-threshold-combinations) this tended to be at a cost of poor specificity and PPV.

Based on mean BA across the five gene-specific truth-sets, the best performing tool-threshold-combinations were meta-tools Revel\_b and MetaSNP. Revel\_b (tolerated ≤0.7, deleterious >0.7) exhibited BA=79%, reflecting sensitivity of 89% and specificity of 68%, whilst MetaSNP (tolerated ≤0.5, deleterious >0.5) exhibited BA=79%, reflecting sensitivity of 92% and specificity of 66% (Figure 1 and Supplementary Table 5). Also strongly performing were PMut, MutPred\_b, and meta-tools rfPred and VEST3\_c (tolerated ≤0.5; deleterious >0.5). Strong performances for some tool-threshold-combinations, such as PANTHER and Eigen-PC\_b (tolerated < 0; deleterious > 0.5) must be caveated by their levels of variant exclusion (54% and 31% respectively). The tools most widely used clinically, SIFT, PolyPhen2\_HumVar and MutationTaster ranked in positions 17th, 23rd and 45th for BA: their high sensitivities (96-98%) came at the cost of poorer specificities (20-38%) (Supplementary Table 8).

Tool performance for variants excluded due to being in the intermediate range of the assays for *BRCA1*, *BRCA2* and *MSH2* is shown in **Supplementary Figure 2** and **Supplementary Table 12**. Median scores for the functionally-intermediate variants largely lay between medians for the deleterious and tolerated group, but with little evidence of graded correlation.

### Consistency between gene-specific truth-sets

The five individual-gene functional truth-sets varied in dataset-size, gene-representativeness and proportion of deleterious variants. The 4783 *MSH2*, 957 *PTEN* and 1867 TP53 variants included spanned the full gene, whilst the 1641 *BRCA1* were restricted to the RING/BRCT domains and the 188 *BRCA2* variants likewise all lay within the DNA-binding domain. Interestingly, examination for all missense variants across the *BRCA1* gene showed a higher median Revel score for variants outside of any domain (0.57) than for variants within the BRCT domain (0.48) (**Supplementary Figure 3**). The proportion of deleterious variants in the *MSH2* dataset (8%) was much lower than for *BRCA1* (23%), *PTEN* (20%) and *TP53* (22%) and in particular *BRCA2* (34%); this metric influences PPV and NPV. Based on ordinal rankings for mean BA, there was broad consistency across the five individual-gene functional-truth-sets for tool-threshold-combinations with binary cut-offs (Supplementary Table 8). There was greater heterogeneity across the five individual-gene functional datasets for tools with non-binary thresholds, on account of the proportion of TP/TN excluded in the indeterminate range.

*Predictions for loss-of-function versus dominant-negative effects*

Pathogenic missense variants in *BRCA1*, *BRCA2* and *MSH2* are understood largely to act via a two-hit (loss-of-function) mechanism. For variants in *PTEN* and *TP53*, pathogenic effect can be conferred by either loss-of-function or dominant negative (gain-of-function) effect. For *TP53*, performance on the Nutlin-tp53WT assay would be predicted to select for variants acting by dominant negative effect (DNE).

The BA for *TP53* was above the mean BA for the five gene-specific truth-sets for all 20/20 of the top performing tools. For *BRCA1*, *BRCA2*, *MSH2* and *PTEN* this proportion was 11/20, 8/20, 6/20 and 9/20 respectively. Additionally, for each of *TP53* and *PTEN*, Revel scores for variants at the known dominant-negative hot-spots exceeded the median scores across all other deleterious variants (**Supplementary Figure 4**). Moreover, across the *TP53* variant set, there was strong correlation (p<2.2 x 10-16) between the Revel Score and the p53WT Nutlin-3 z-score (**Supplementary Figure 5**).

### Combinations of tools

Despite Revel\_b being a meta-predictor encompassing twelve component tools, its mean BA across the five genes could be improved from 79% to up to 84% by concordance combination with other high performing tools such as Meta-SNP, VEST3, rfPred and MutPred, although with drop-out of discordant variants ranging from 6-26% (Supplementary Table 8).

### Performance against ClinVar

The prediction parameters for the 70 tool-threshold-combinations against the ClinVar-truth-set, generally exceed performance against the mean of the functional-truth-sets, likely reflecting the direct or indirect relationship between ClinVar classifications and tool training11. Overall there was consistency in the ordinal performance of most tools between the mean of the functional-truth-sets and the ClinVar-truth-set, with Revel\_b ranking second for BA against the ClinVar-truth-set. Performance against the ClinVar-truth-set appeared disproportionately better for tools trained exclusively on ClinVar, such as ClinPred, compared to tools trained on different, mixed datasets.

### Positive and negative likelihood ratios

Particularly relevant metrics for clinical classification are the positive likelihood ratio for calling deleterious (true positive rate/false positive rate) and negative likelihood ratio for calling deleterious (or positive likelihood ratio for calling benignity, the true negative rate/false negative rate). Tool-threshold-combinations performing well on BA tended to exhibit strong but balanced positive and negative likelihood ratios, for example Revel\_b had mean PLR= 3.13 (2.75-3.58) and NLR= 7.20 (6.27-8.33) whilst Meta-SNP exhibited PLR= 2.79 (2.49-3.14) and NLR= 9.98 (8.81-11.3) (Supplementary Table 9).

Tool-threshold-combinations with high sensitivity and low specificity typically exhibited poor PLR but strong NLR, driven by low rates of false negatives. Tool-threshold-combinations with high specificity but weaker sensitivity exhibited poor NLR but much stronger PLR, driven by lower rates of false positives. Using the mean of the five functional-truth-sets, PLRs and NLRs at different thresholds of Revel and MetaSNP were calculated (Supplementary Tables 10, 11).

## Discussion

We present predictive parameters and positive/negative likelihood ratios for 44 in silico tools and 70 tool-threshold-combinations, examining 9436 missense variants generated from systematic functional assays for five genes, which have been validated against clinical pathogenicity.

We demonstrated that most widely-used in silico tools have high sensitivity, that is they are unlikely to miscall a truly deleterious variants. However, many of the tools maintain high sensitivity at the expense of very high false positive call-rates, as reflected by the 56/70 tools which called more than 50% of the variants as deleterious (true frequency 15%). For the tools widely used in clinic at their specified thresholds, across the five functional-truth-sets mean PPV was 30% for SIFT, 28% for Polyphen HumVAR and 26% for MutationTaster 11.

Because tools are generally calibrated to overcall as pathogenic, their negative predictive value is typically good: 44/70 tool-threshold-combinations had NPV>95%. Furthermore, NPV is dependent on the prevalence of true pathogenic variants; the NPV would further improve in the context of a clinical laboratory in which prevalence of true pathogenic variants is typically lower than the 15% in the combined-functional-truth-set. These data argue against current equivalence within the ACMG/AMP framework for in silico tool prediction of pathogenicity and benignity. These data replicate similar observations reported in previous analyses using ClinVar truth-sets17. For example, 2361 variants have Revel score <0.5: of these 2328 are true negatives and only 33 are false negatives.

As tool thresholds are typically set for high sensitivity, it is the corresponding specificity which drives our rankings for BA. At specified thresholds, Revel\_b, MetaSNP, PMut, MutPred\_b, rfPred and VEST3\_c all perform well, with BA≥73%, AUC≥83% and MMC≥41%. Notably Revel, Meta-SNP and rfPred are all meta-predictors, that is they have been developed using machine learning optimised amalgamation of component algorithms (Supplementary Table 1).

Although included as it is a widely-used tool, we would caveat generalisability of performance of Align-GVGD, as not only have sequence alignments been especially well-curated for the genes analysed, but the tool was trained on *BRCA1/2* classifications and *TP53* functional datasets32,41.

Against the five functional-truth-sets, for concordant calls of deleterious for SIFT, Polyphen2\_HumVar and MutationTaster, the mean positive Likelihood Ratio is only 1.21 (1.16-1.27), with 39% of variants dropping out due to discordant calls (Supplementary Table 9)11. More broadly, for any non-binary tool-threshold-combinations or combining of tools using a concordance model, any apparent boost in calculated BA must be caveated by the inevitable exclusion of a substantial proportion of the ‘difficult’ indeterminate/discordant variants. As the 2015 ACMG/AMP framework does not specify which in silico tools are allowable, it is duly conservative in offering only supplementary evidence weighting (Likelihood Ratio ~2) and only where multiple tools are concordant. Using Revel at the dichotomous threshold of 0.7, offers PLR of 3.13 (2.75-3.58) and NLR of 7.20 (6.27-8.33), but higher evidence weighting may be warranted for scores at the extreme tails. For example the LR of 6.74 (5.24-8.82)) for Revel or 42.9 (14.4-406) for Meta-SNP for scores of 0.8-1.0 would comfortably constitute stronger evidence towards pathogenicity, as would the LR of 34.3 (31.5-37.3) for Revel or 19.4 (15.6-24.9) for Meta-SNP for scores of <0-0.4 towards benignity (Table 2)22,42. Our data would thus overall support calibrated use of high-performing meta-tools for clinical variant interpretation, rather than ad-hoc combinations of multiple tools. Provided the tool has not been trained on the functional data, as in silico predictions are derived from orthogonal data to functional assays, we would support the two evidence types being separately counted towards a variant classification.

The relationship between functional assay results, clinical classifications and true underlying pathogenicity remains elusive. Imperfect correlation of the assay data to clinical classification may in part reflect erroneous clinical classifications resident on public databases. Clinical classifications are indeed not sacrosanct and are only as good as the comprehensiveness and accuracy of available clinical information as well as the validity of classification schema employed 43,44. The functional assays assessed are relatively recent; their incorporation into clinical classification, ClinVar and other resources will further confound data benchmarking.

Inflation of tool performance against publicly available datasets (in particular ClinVar), over-fitting and the shortcomings of clinically-derived classifications have been well described previously11. Indeed, many ‘truth-sets’ previously used for tool evaluations (i) have overlapped with the datasets upon which tools were trained or (ii) correlated poorly with true clinical pathogenicity, comprising population data and/or prokaryotic cell models and/or in vitro assays of functions and/or ‘clinical’ classifications of poor quality11,45. Thus, whilst the functional assays we have used are unlikely to perfectly recapitulate true human pathogenesis, given their powerful correlation against clinical classifications, size of data and systematic generation, arguably they represent leading truth-sets for unbiased evaluation of tool performance.

Although we excluded the two intron-flanking exonic variants on account of potential spliceogenic effect, other spliceogenic exonic variants resulting in a null protein will have been called as deleterious by most assays: we have assumed this group to be small in number and roughly consistent between genes. Although our analysis has focused predominantly on missense variants for which pathogenesis is via loss-of-function; our data for *TP53/PTEN* support the tools discriminating comparably for DNE. This observation is consistent with previous reports in which prediction for DNE compared to loss-of-function (LOF) variants was poorer for older tools (SIFT, Polyphen) but equivalent using newer algorithms such as Revel17,18.

The observed variation in tool performance between the five individual-gene truth-sets likely reflects heterogeneity in composition of pathogenic variants types (LOF versus DNE), varying accuracy of assay in recapitulation of true pathogenicity and sampling variation (chance). It has been argued that, there will be systematic differences gene-by-gene in how tools perform, on account of innate gene-specific differences in the genomic context of pathogenic and benign variants. Data from a broader range of MAVES examining the full spectrum of coding variants would enable further exploration of such hypotheses. However, whilst there are rapid advances in technology for high throughput gene editing and assay readout, expansion of MAVES to additional genes has been limited on account of lack of availability of clinical truth-sets by which to validate assays and limited understanding of mutational mechanism of clinical pathogenicity46.

A perennial issue in the arena of variant interpretation is that of intermediate penetrance/effect. The clinical model of dichotomous classification as pathogenic or benign imperfectly accommodates underlying continuity of clinical penetrance and a corresponding more continuous distribution of in vivo and in vitro cellular function. To simplify the tool assessment, we removed from our functional-truth-sets all variants scored as intermediate for assay performance. However, better quantitation of variants of intermediate effect will require study of these intermediate assay scores and will necessitate use of continuous measurement rather than binary categories for both in silico predictions and functional assays.

Performance of combinations for high-performing tools indicate room for improvement in algorithms. Furthermore, whilst we focused predominantly on established/author-provided tool thresholds, generation of new thresholds optimised for BA against these functional datasets indicated potential for substantially improved tool performance, in particular for MetaSNP and MutPred (Supplementary Table 13). Tools could be further evolved using more advanced machine learning approaches with weighting of contribution of these functional truth-set to optimise tool combination, performance and variant inclusion.

There is a growing preponderance of in silico tools. As many previous authors have found, many of these tools used at their recommended thresholds have very poor specificity and PPV. Where ClinVar was also used to train the tools, tool evaluation against ClinVar may misrepresent performance due to over-fitting. The cautious ACMG/AMP evidence weights may still be overly-generous for many in silico tools. However, evaluation against a large systematically-generated ‘clinical-grade’ truth-set of functional assay data allows unbiased identification of the more predictive tools and discriminatory thresholds. Our data suggest that greater weights of evidence towards pathogenicity/benignity might be afforded for specific tools such as Revel and MetaSNP, with potential for evidence calibration by absolute score. Using a Bayesian conversion, the respective relevant positive and negative likelihood ratios can be incorporated into the ACMG/AMP framework22,42.

## Acknowledgments

C.C., A.G., S.C., L.K., B.T. and H.H. are supported by CRUK Catalyst Award CanGene-CanVar (C61296/A27223). N.W. is currently supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 220134/Z/20/Z).

## Author Information

Conceptualization: C.T., C.C., M.D., A.C., G.J.B., R.R., J.D., I.B., A.W.; Data curation: C.C., S.C., L.K., A.G., H.H..; Formal Analysis: C.C., A.G., C.T., C.L., J.W., N.W; Funding acquisition: C.T., M.T., D.E.; Project administration: B.T.; Visualization: C.C., C.T.; Writing – original draft: C.T., C.C., A.G.; Writing – review & editing: all authors

## Ethics Declaration

This analysis made use of publicly available datasets only. The human variant data used was all de-identified and therefore IRB approval was not required.

## Data Availability

The data analysed are all publicly available as per references/URLs provided. While this manuscript does not contain primary research data, materials and data developed during this study will be made available upon request to the corresponding author.

## Disclosure

The authors declare no conflict of interest.

**Figure 1: Balanced Accuracy for 70 tool-threshold combinations for seven truth-sets.** Rates of true positive (TP), false negative (FN), true negative (TN) and false positive (FP) tool calls against functional truth-sets also shown, along with rates of tool calls of deleterious (DEL, pink), tolerated (TOL, blue) or indeterminate (grey)

## References

1. Firth, H.V. & Hurst, J.A. *Oxford Desk Reference: Clinical Genetics and Genomics*, (Oxford University Press, 2017).

2. Richards, 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* **17**, 405-24 (2015).

3. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862-4 (1974).

4. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).

5. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**, S3 (2013).

6. Sim, N.L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* **40**, W452-7 (2012).

7. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506-d515 (2019).

8. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).

9. Stenson, P.D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* **136**, 665-677 (2017).

10. Landrum, M.J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-d1067 (2018).

11. Gunning, A.C. *et al.* Assessing performance of pathogenicity predictors using clinically relevant variant datasets. *J Med Genet* (2020).

12. Hawkins, D.M. The problem of overfitting. *J Chem Inf Comput Sci* **44**, 1-12 (2004).

13. Subramanian, J. & Simon, R. Overfitting in prediction models - is it a problem only in high dimensions? *Contemp Clin Trials* **36**, 636-41 (2013).

14. Grimm, D.G. *et al.* The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum Mutat* **36**, 513-23 (2015).

15. Stenson, P.D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* **133**, 1-9 (2014).

16. Landrum, M.J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-5 (2014).

17. Ghosh, R., Oak, N. & Plon, S.E. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol* **18**, 225 (2017).

18. Flanagan, S.E., Patch, A.M. & Ellard, S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet Test Mol Biomarkers* **14**, 533-7 (2010).

19. Niroula, A. & Vihinen, M. How good are pathogenicity predictors in detecting benign variants? *PLoS Comput Biol* **15**, e1006481 (2019).

20. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).

21. Alamut Visual Software Application <https://www.interactive-biosoftware.com/alamut-visual/>.

22. Tavtigian, S.V. *et al.* Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med* **20**, 1054-1060 (2018).

23. Schwarz, J.M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods* **7**, 575-576 (2010).

24. Findlay, G.M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing <https://sge.gs.washington.edu/BRCA1/>. *Nature* **562**, 217-222 (2018).

25. Guidugli, L. *et al.* Assessment of the Clinical Relevance of BRCA2 Missense Variants by Functional and Computational Approaches <https://www.sciencedirect.com/science/article/pii/S0002929717305025?via%3Dihub>. *Am J Hum Genet* **102**, 233-248 (2018).

26. Guidugli, L. *et al.* Functional assays for analysis of variants of uncertain significance in BRCA2. *Hum Mutat* **35**, 151-64 (2014).

27. Richardson, M.E. *et al.* Strong functional data for pathogenicity or neutrality classify BRCA2 DNA-binding-domain variants of uncertain significance. *Am J Hum Genet* **108**, 458-468 (2021).

28. Hart, S.N. *et al.* Comprehensive annotation of BRCA1 and BRCA2 missense variants by functionally validated sequence-based computational prediction models. *Genet Med* **21**, 71-80 (2019).

29. Jia, X. *et al.* Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *Am J Hum Genet* **108**, 163-175 (2021).

30. Mighell, T.L., Thacker, S., Fombonne, E., Eng, C. & O'Roak, B.J. An Integrated Deep-Mutational-Scanning Approach Provides Clinical Insights on PTEN Genotype-Phenotype Relationships. *Am J Hum Genet* **106**, 818-829 (2020).

31. Matreyek, K.A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet* **50**, 874-882 (2018).

32. Kato, S. *et al.* Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis <https://www.pnas.org/content/100/14/8424.long>. *Proc Natl Acad Sci U S A* **100**, 8424-9 (2003).

33. Giacomelli, A.O. *et al.* Mutational processes shape the landscape of TP53 mutations in human cancer <https://www.nature.com/articles/s41588-018-0204-y>. *Nat Genet* **50**, 1381-1387 (2018).

34. Savage, S.A. TP53 Rule Specifications for the ACMG/AMP Variant Curation Guidelines. Vol. 2019 (ClinGen, clinicalgenome.org, 2019).

35. Fortuno, C. *et al.* Specifications of the ACMG/AMP variant interpretation guidelines for germline TP53 variants. *Hum Mutat* (2020).

36. Brodersen, K.H., Ong, C.S., Stephan, K.E. & Buhmann, J.M. The balanced accuracy and its posterior distribution. in *2010 20th International Conference on Pattern Recognition* 3121-3124 (IEEE, 2010).

37. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).

38. Centor, R.M. Estimating confidence intervals of likelihood ratios. *Med Decis Making* **12**, 229-33 (1992).

39. Ioannidis, N.M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877-885 (2016).

40. Ellard, S. *et al.* ACGS Best Practice Guidelines for Variant Classification 2020. (Association for Clinical Genetics Science (ACGS), 2020).

41. Tavtigian, S.V., Greenblatt, M.S., Lesueur, F. & Byrnes, G.B. In silico analysis of missense substitutions using sequence-alignment based methods. *Hum Mutat* **29**, 1327-36 (2008).

42. Garrett, A. *et al.* Combining evidence for and against pathogenicity for variants in cancer susceptibility genes: CanVIG-UK consensus recommendations. *J Med Genet* (2020).

43. Yang, S. *et al.* Sources of discordance among germ-line variant classifications in ClinVar. *Genetics in Medicine* **19**, 1118-1126 (2017).

44. Harrison, S.M. *et al.* Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. *Genet Med* (2017).

45. Mahmood, K. *et al.* Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Human Genomics* **11**, 10 (2017).

46. Gelman, H. *et al.* Recommendations for the collection and use of multiplexed functional data for clinical variant interpretation. *Genome Med* **11**, 85 (2019).