Choosing variant interpretation tools for clinical applications: context matters

Josu Aguirre[1,†], Natàlia Padilla[1,†], Selen Özkan[1], Casandra Riera[1], Lidia Feliubadaló[2], and Xavier de la Cruz.*[1,3]

[1] Research Unit in Clinical and Translational Bioinformatics, Vall d'Hebron Institute of Research (VHIR), Universitat Autònoma de Barcelona, P/Vall d'Hebron, 119–129; 08035, Barcelona, Spain.

[2]Hereditary Cancer Program, Catalan Institute of Oncology (ICO), ONCOBELL-IDIBELL, CIBERONC, Barcelona, Spain

[3] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

[†]These two authors have contributed equally to this work

***Corresponding author;** Xavier de la Cruz. Email: xavier.delacruz@vhir.org

Abstract

Motivation: Our inability to solve the Variant Interpretation Problem (VIP) has become a bottleneck in the biomedical/clinical application of Next-Generation Sequencing. This situation has favored the development and use of *in silico* tools for the VIP. However, choosing the optimal tool for our purposes is difficult because of a fact usually ignored: the high variability of clinical context/scenarios across and within countries, and over time.

Results: We have developed a computational procedure, based on the use of cost models, that allows the simultaneous comparison of an arbitrary number of tools across all possible clinical scenarios. We apply our approach to a set of pathogenicity predictors for missense variants, showing how differences in clinical context translate to differences in tool ranking.

Availability: The source code is available at:

https://github.com/ClinicalTranslationalBioinformatics/clinical_space_partition

30 **BACKGROUND**

31 The clinical application of Next-Generation Sequencing (NGS) is presently limited by
32 our inability to fully interpret its results (Starita *et al.*, 2017; Shendure, Findlay and Snyder,
33 2019; Berrios *et al.*, 2021): we still cannot establish, with 100% accuracy, if the sequence
34 variants identified by this technique are benign or pathogenic. This problem, known as the
35 Variant Interpretation Problem (VIP) (Starita *et al.*, 2017), has important consequences in
36 terms of patient lives and economic cost, and is considered one of the challenges determining
37 the future of Genomic Medicine (Shendure, Findlay and Snyder, 2019).

38 A promising option to address the VIP is the experimental characterization of massive
39 amounts of variants using large-scale functional assays (Starita *et al.*, 2017; Shendure,
40 Findlay and Snyder, 2019). However, despite some good results, functional data still lack for
41 many proteins. There is an alternative option for variant annotation, in which the VIP is
42 posed as a classification problem and solved using pathogenicity predictors, which are *in*
43 *silico* tools based on artificial intelligence algorithms. Fast and inexpensive, these tools
44 require almost no training, and their use is contemplated in the ACMG/AMP (Richards *et al.*,
45 2015) guidelines for variant interpretation. However, the growing number of these methods
46 (>50) and their varied performances (Özkan *et al.*, 2021) pose a problem when deciding
47 which of these tools is the optimal choice for our application of interest, a non-trivial problem
48 that becomes apparent, e.g., when adapting ACMG/AMP guidelines to specific diseases
49 (Amendola *et al.*, 2016; Ghosh, Oak and Plon, 2017).

50 The procedure for predictor selection is common to any scientific/technological field:
51 (i) rank all available tools according to some performance measure, and (ii) choose the tool at
52 the top. In the case of VIP, experts often utilize summary parameters (Feliubadaló *et al.*,
53 2021; Fortuno *et al.*, 2021), like the Area Under the Roc Curve (AUC) or the Matthews
54 Correlation Coefficient (MCC) for the ranking step. However, because these parameters are

55 blind to the deployment context (Hand, 2012), their use may result in choices that are not
56 universally optimal. This situation is particularly plausible in the case clinical applications,
57 where context includes all the healthcare stakeholders involved in a medical problem, their
58 needs and actions, and the amount and nature of resources at their disposal. All these
59 components are characterized by a large national and international variability that can make
60 what is optimal in a scenario to become sub-optimal in another. For example, tools with high
61 sensitivity and low specificity may be adequate when working with the average population
62 but detrimental when applied to frail populations, for which tools with balanced
63 sensitivity/specificity are more desirable. However, quantifying this type of context
64 considerations in a manner that helps researchers and clinicians to identify the best tool for
65 each scenario, is a difficult challenge. Here we use cost models (Adams and Hand, 1999;
66 Pepe, 2003; Drummond and Holte, 2006; Hernández-Orallo, Flach and Ferri, 2012), a
67 formalism commonly applied in medical decision processes (Pepe, 2003; Hunink *et al.*,
68 2014), to address this problem. A convenient characteristic of cost models is that we can
69 encode the main features of the application context in a few numerical parameters, thus
70 allowing a systematic analysis of how context affects tool selection. We focus on this aspect
71 for the VIP, applying/extending the standard cost formalism (Adams and Hand, 1999;
72 Drummond and Holte, 2006; Hernández-Orallo, Flach and Ferri, 2012) to variant annotation
73 tools. In particular, we develop a computational procedure that allows comparing these tools
74 across the whole range of clinical scenarios. We apply our methodology to a set of sixteen
75 pathogenicity predictors for missense variants, showing how different methods may be
76 preferred depending on the clinical context.
77
78 **MATERIALS AND METHODS**
79 **Variant dataset**

80 For each predictor, we estimated the three performance parameters used for cost
81 models, sensitivity (s_e), specificity (s_p), and coverage (α), in a set of benign and pathogenic
82 variants retrieved from the dbNSFP database (Liu *et al.*, 2016), version 4.0b1a, release:
83 December 8, 2018. We imposed three filters. First, the variants should not affect a splicing
84 site. Second, the ClinVar Review Status of the variants should be: 'Practice guideline', or
85 'Expert Panel' or 'Criteria provided, multiple submitters, no conflicts'. And, third, we unified
86 the clinical significance classes as follows: 'Benign' and 'Likely_benign' variants were labeled
87 as 'benign'; 'Pathogenic' and 'Likely_pathogenic' variants were labeled as 'pathogenic'. The
88 resulting dataset was constituted by 11093 variants, 3752 pathogenic and 7341 benign, from
89 1914 proteins.
90
91 **Pathogenicity predictors**
92 We have used sixteen pathogenicity predictors: PolyPhen2-HDIV (Adzhubei *et al.*,
93 2010), PolyPhen2-HVAR (Adzhubei *et al.*, 2010), SIFT (Kumar, Henikoff and Ng, 2009),
94 CADD (Rentzsch *et al.*, 2019), MutationTaster2 (Schwarz *et al.*, 2014), MutationAssessor
95 (Reva, Antipin and Sander, 2011), REVEL (Ioannidis *et al.*, 2016), FATHMM (Shihab *et al.*,
96 2013), LRT (Chun and Fay, 2009), PROVEAN (Choi *et al.*, 2012), MetaLR (Dong *et al.*,
97 2015), MetaSVM (Dong *et al.*, 2015), VEST(Carter *et al.*, 2013), MutPred (Pejaver *et al.*,
98 2020), PON-P2 (Niroula, Urolagin and Vihinen, 2015), and PMut (López-Ferrando *et al.*,
99 2017). For each variant in our dataset, we retrieved the pathogenicity prediction of these tools
100 from the dbNSFP (Liu *et al.*, 2016) database, except for PON-P2 and PMut, for which we
101 used the corresponding website.
102
103 **Sensitivity, specificity, and coverage**
104 For each predictor, we computed the s_e, s_p, and α as follows:

$$s_e = \frac{TP}{N_p} \qquad s_p = \frac{TN}{N_b} \qquad \alpha = \frac{N}{N_{tot}}$$

where N_{tot} is the total number of variants in the dataset (11093), and N is the number of predictions generated by the tool ($N \leq N_{tot}$). N is the sum of N_p and N_b, the respective numbers of pathogenic and benign variants for which the predictor gave an output. TP (true positives) and TN (true negatives) are the numbers of pathogenic and benign variants successfully predicted by the predictor, respectively.

Sensitivity and specificity are related to the two misclassification errors of pathogenicity predictors: $1\text{-}s_e$, the fraction of pathogenic variants misclassified as benign, and $1\text{-}s_p$, the fraction of benign variants misclassified as pathogenic.

Coverage is related to a different limitation of predictors: $1\text{-}\alpha$ is the fraction of variants for which a predictor does not provide an output.

Cost computations

The computations associated with the two cost models used in this work, MISC and MISC+NOA (see below), are coded into separate Python scripts, CSP-noco and CSP-co, respectively.

The code is available at:

https://github.com/ClinicalTranslationalBioinformatics/clinical_space_partition

RESULTS

Before describing the results, we first give an intuitive explanation of the application of costs models to the VIP in the clinical setting. Although we focus on the case of

129 pathogenicity predictors, the reasoning is valid for any variant annotation process, whether
130 computational or experimental.

131 In different healthcare problems, *in silico* variant annotations are combined with
132 diverse amounts of clinical data. A common characteristic of these problems, e.g., in
133 diagnostics, is that one must identify the correct/best answer among several options. Each
134 answer is associated to a series of medical actions and costs to patients and their families,
135 healthcare institutions, etc. The medical decision process aims at minimizing these costs
136 (Hunink *et al.*, 2014). Accurate pathogenicity predictions contribute to this goal, whereas
137 misclassification errors (MISC) -benign variants predicted as pathogenic, or vice versa- favor
138 incorrect decisions. Even the apparently harmless situation in which a predictor produces no
139 answer (which we will call NOA) has consequences. It may call for further clinical testing in
140 diagnosis, limit the reach of screenings, etc.

141 In summary, both MISC and NOA are associated with costs affecting healthcare
142 stakeholders. Reducing this impact is thus a guiding principle when ranking annotation tools
143 for clinical applications. Here, we advance a rigorous approach, based on cost models
144 (Adams and Hand, 1999; Pepe, 2003; Drummond and Holte, 2006; Hernández-Orallo, Flach
145 and Ferri, 2012), to address this challenge. First, we present two frameworks that correspond
146 to the aforementioned situations: MISC, in which only misclassification errors are
147 considered, and MISC+NOA, in which we also consider the case when tools give no answer.
148 The underlying formalism, a combination of computational and mathematical results, is given
149 in Appendices 1 and 2. After presenting the cost models, we report their application to study
150 how the diversity of clinical scenarios affects the prioritization of pathogenicity predictors.
151

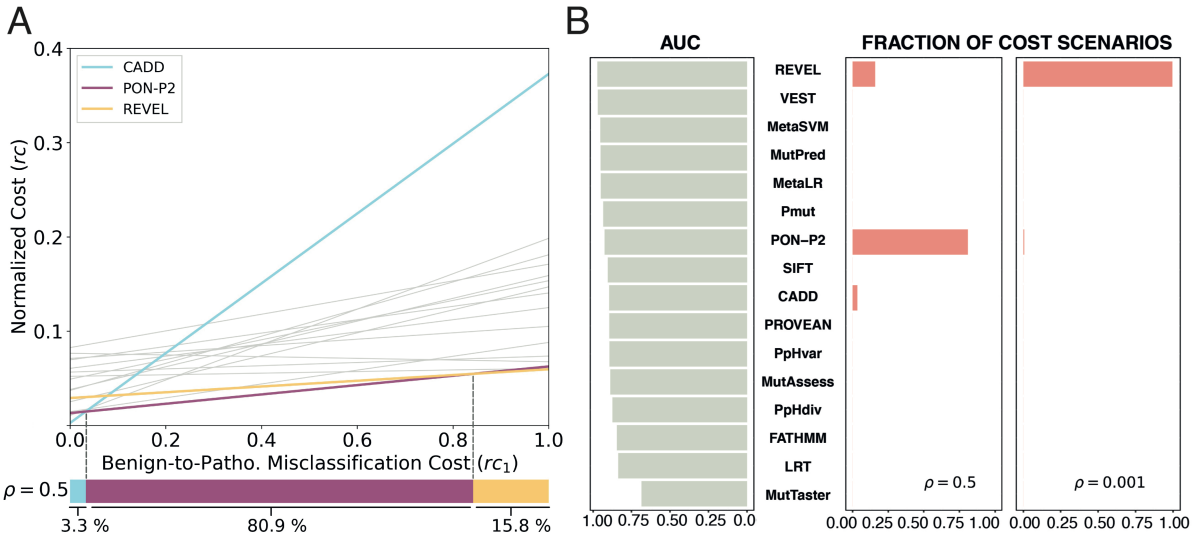
152 **Cost model for MISC**

153 This model is appropriate in settings where the consequences of NOA are negligible.
154 For example, when working with proteins for which pathogenicity predictors have very good
155 coverage; when searching for combinations of methods that optimally cover all clinical
156 settings; etc.
157 In this case (Appendix 1, Supplementary Materials), we can compute rc, the
158 normalized average cost of misclassification errors (Adams and Hand, 1999; Drummond and
159 Holte, 2006), applying the standard cost model (Adams and Hand, 1999; Pepe, 2003;
160 Drummond and Holte, 2006; Hernández-Orallo, Flach and Ferri, 2012):
161
162 (1) $rc = [(1-\rho)(1-s_p) - \rho(1-s_e)]rc_1 + \rho(1-s_e)$
163
164 where s_e and s_p are the sensitivity and specificity of the predictor, respectively. rc_1 is the
165 normalized cost of misclassifying benign variants as pathogenic. Each value of rc_1
166 corresponds to a clinical context encoded in terms of cost. The set of rc_1 values covers all
167 possible clinical contexts; we will refer to it as 'clinical space'. The equivalent cost for the
168 misclassification of pathogenic variants as benign is implicit in the formula (see Appendix 1).
169 Finallly, ρ is the proportion of pathogenic variants expected in the deployment context. Its
170 value is comprised between 0 and 1, and varies with the genome region sequenced (gene
171 panel, whole exome, etc.) and with the population of individuals tested (e.g., the individuals
172 attending a specific hospital unit, ethnic group, etc.). Below, we explore how different values
173 of ρ affect the cost-ranking of predictors. For some of our analyses, to reflect the biasing
174 effect of working with patient populations, we chose ρ=0.5.
175 We can use rc for ranking an arbitrary number of predictors in terms of cost. For a
176 single clinical scenario (defined by its rc_1) the procedure is: (i) compute the rc of each

177 method using equation (1); and (ii) sort the resulting values. The cost-optimal method will be

178 the one with the lowest rc.

179



180

181 **Fig. 1. Application of the cost model MISC to sixteen pathogenicity predictors. a,** Each

182 predictor is represented by a line in which each point corresponds to the normalized cost of

183 the predictor (y-axis) at a given value of the benign-to-pathogenic misclassification error (rc_1,

184 x-axis). We have colored the lines corresponding to PON-P2, REVEL, and CADD. The

185 colored bar below the x-axis shows the fraction of cost scenarios corresponding to each of

186 these predictors. **b**, Ranking of the sixteen pathogenicity predictors according to AUC (grey

187 bars, left half) and fraction of cost scenarios for which each predictor is optimal (pink bars,

188 right half; results shown for two ρ values, 0.5 and 0.001).

189

190

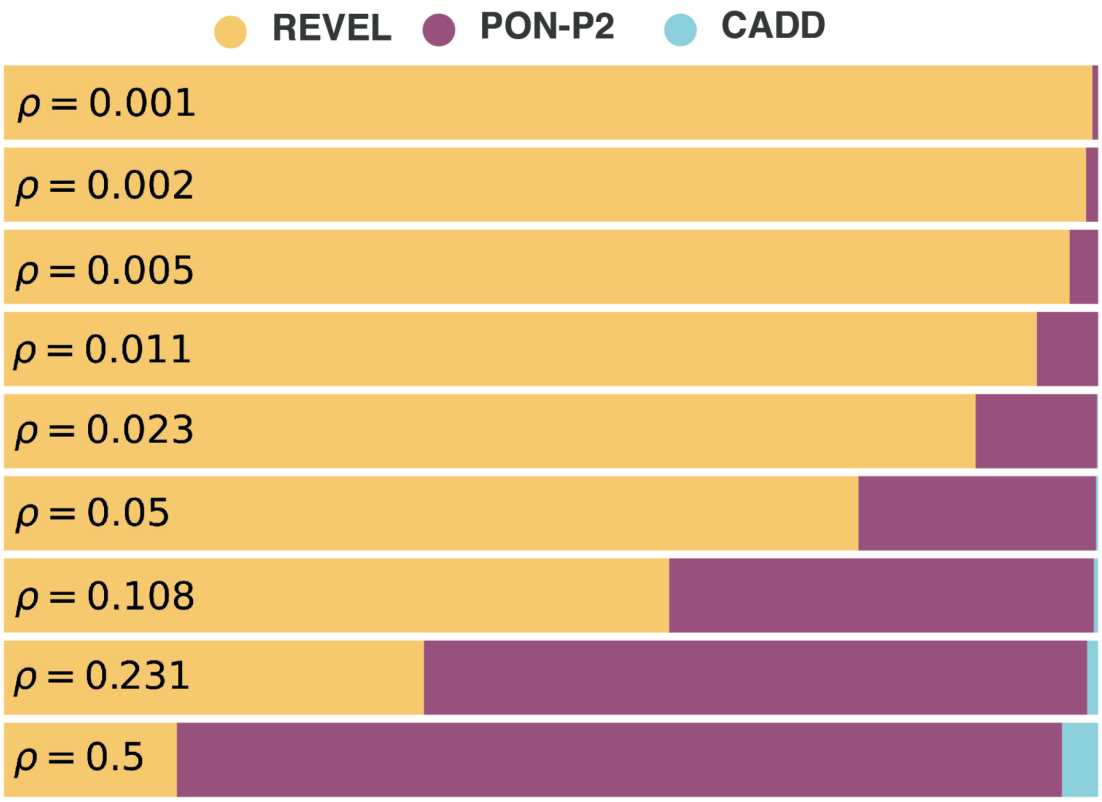
191 To explore the dependence of the previous ranking on clinical context we could use a

192 brute force approach: varying rc_1 between 0 and 1 systematically and repeating the ranking at

193 each value. This approach will produce a division of the clinical space into regions within

194 which only a single predictor is cost-optimal. However, brute force may be computationally

195 demanding and inaccurate when the number of predictors is large. To avoid these problems,
196 we have developed a fast alternative; an algorithm (CSP-noco, Appendix 1, Supplementary
197 Materials) that produces the exact division of the clinical space using some formal properties
198 of the problem. Below, we illustrate the application of this algorithm to the comparison of a
199 set of pathogenicity predictors across the whole clinical space (Fig. 1) and show, also, how
200 the results obtained depend on ρ (Fig. 2).



201

202 **Fig. 2. Dependency on ρ of the distribution of pathogenicity predictors between cost**
203 **scenarios.** Each bar represents a version of the clinical space for a ρ value. The size of the
204 colored regions corresponds to the fraction of cost scenarios where a single predictor is cost-
205 optimal. The three methods represented (REVEL, PON-P2, CADD) are those that are cost-
206 optimal in at least some scenarios.

207

208 **Cost model for MISC+NOA**

209 We have extended MISC to include the cost of NOA (see Appendix 2, Supplementary
210 Materials), the case when predictors do not produce an output, i.e., when they have an
211 incomplete coverage, a relatively frequent situation. For example, it may happen when
212 combining several pathogenicity predictors within the ACMG/AMP guidelines (Richards *et*
213 *al.*, 2015; de la Campa, Padilla and de la Cruz, 2017; Vihinen, 2020): when the outputs
214 disagree, computational evidence is rejected. The fraction of cases for which this happens is
215 $(1-\alpha)$, where α is the coverage of the method (see Materials and Methods).

216 In this new scenario, rc is equal to:

217

218 (2) $rc = [\alpha\rho(1-s_e)+\alpha-1]rc_0+[\alpha(1-\rho)(1-s_p)+\alpha-1]rc_1+1-\alpha$

219

220 where rc_0, rc_1, and ρ are the same as in the previous section. α is the coverage of the method
221 (see Materials and Methods). The cost associated to NOA, rc_2, is implicit in the formalism
222 (see Appendix 2). The set of (rc_0, rc_1) pairs comprises all possible clinical scenarios when
223 misclassification and no answer costs are considered. This set defines a triangular region in
224 the rc_0-rc_1 plane (Fig. 3a) to which we will refer as 'clinical space'.

225

226

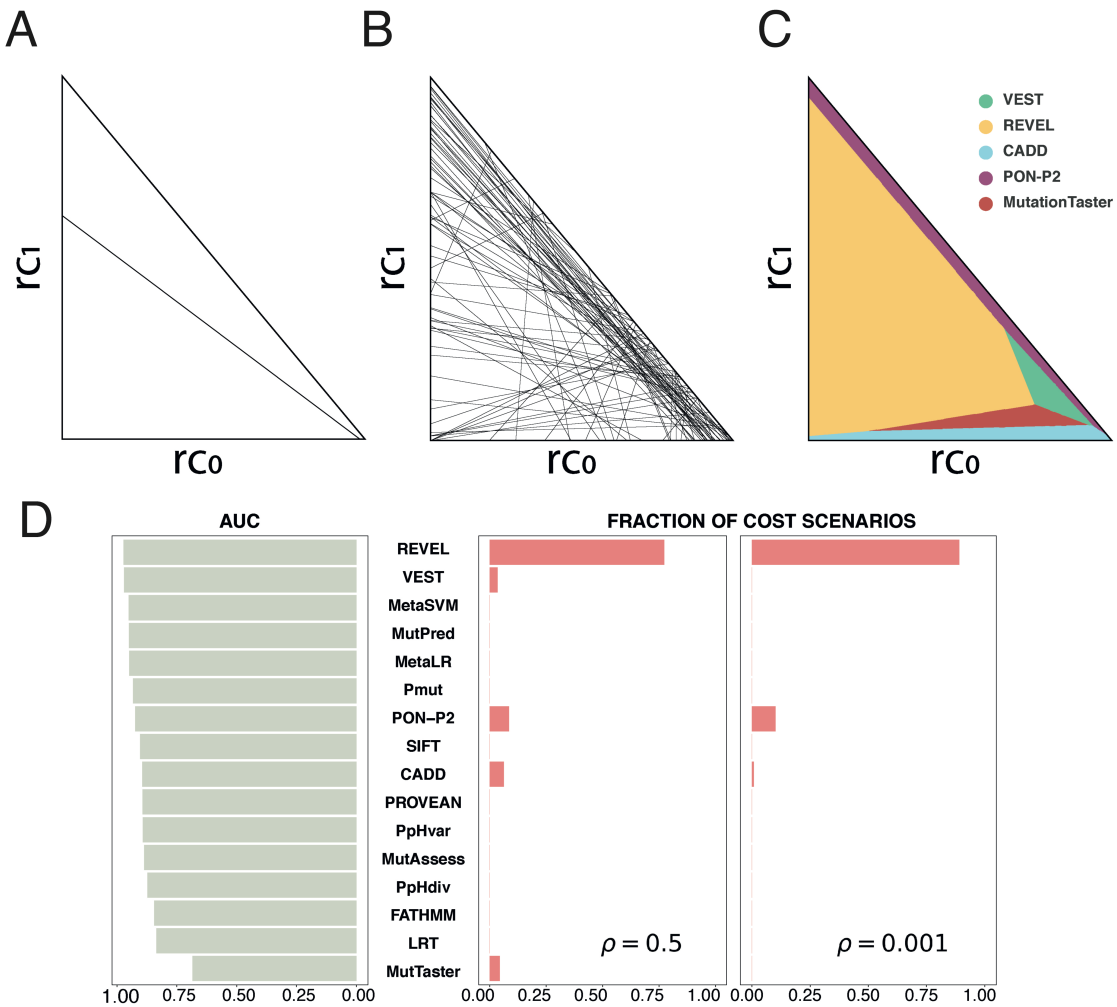
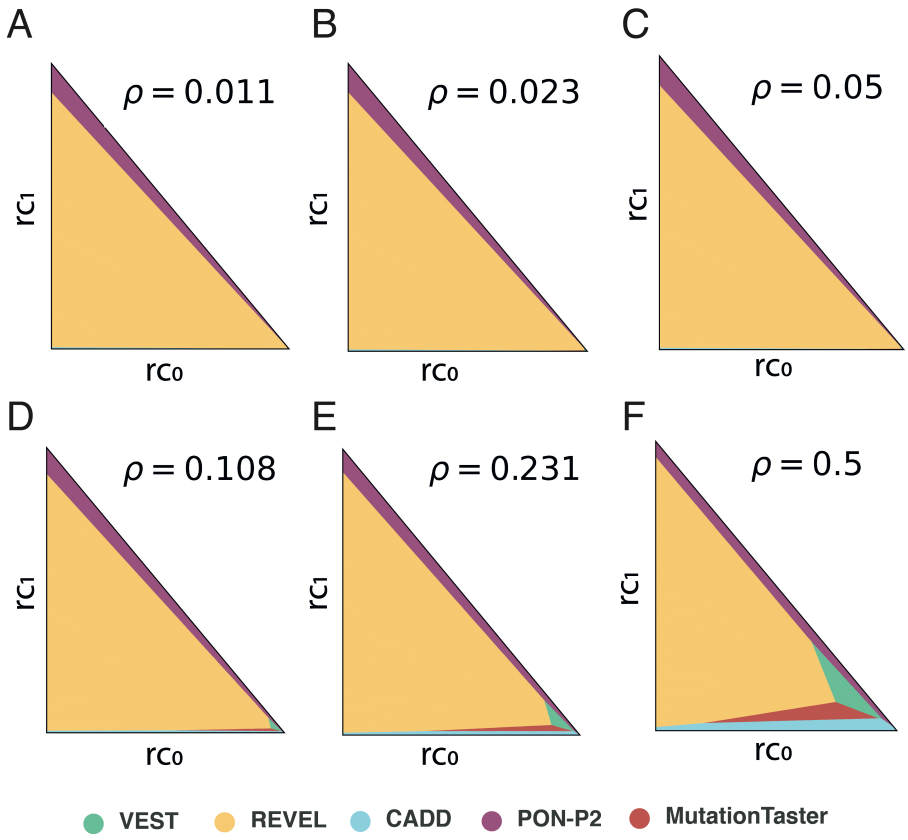


Fig. 3. Application of the cost model MISC+NOA (misclassification errors+no answer) to sixteen pathogenicity predictors. a, Comparison of two predictors. The clinical space is represented by a triangle encompassing all cost scenarios -all (rc_0,rc_1) pairs. The line divides the triangle into two regions within which a single method is cost-optimal. **b**, As we add more predictors, a pattern of polygons forms; inside each polygon, a single predictor is cost-optimal. **c,** Identification and unification of the polygons in regions corresponding to the same predictor, using our approach (CSP-co). **d,** Ranking of the sixteen pathogenicity predictors according to AUC (grey bars, left half) and fraction of cost scenarios for which each predictor is optimal (pink bars, right half; results for ρ=0.5 and ρ=0.001). This fraction is the size of the predictor's area in the triangle, divided by 0.5 (the whole triangle area).

238 Because rc has the same meaning as in MISC, it can be used to systematically rank
239 predictors in multiple clinical scenarios, in terms of cost. However, for this task we cannot
240 utilize the above algorithm, CSP-noco, because the clinical space is now two-dimensional. To
241 solve this problem, we have developed a Breadth First Search-based algorithm (CSP-co,
242 Appendix 2, Supplementary Materials) that divides the clinical space into regions within
243 which only a single predictor is cost-optimal. As before, this division depends on ρ, an aspect
244 that we explore below (Fig. 4).
245



246

247 **Fig. 4. Dependency on ρ of the distribution of pathogenicity predictors between cost**
248 **scenarios.** Each triangle represents the clinical space for a ρ value. **a,** ρ=0.011; **b,** ρ=0.023; **c,**
249 ρ=0.05; **d,** ρ=0.108; **e,** ρ=0.231; **f,** ρ=0.5. We chose ρ values such that each order of
250 magnitude is evenly sampled. The colored regions correspond to the cost scenarios where the
251 same predictor is cost-optimal.

252

253 **Application of cost models to the comparison of pathogenicity predictors across cost**

254 **scenarios**

255 In this section, we show the application of MISC and MISC+NOA, using CSP-noco

256 and CSP-co, respectively, to the comparison of sixteen selected pathogenicity predictors

257 across clinical space. The goal is to study the dependence of the cost-optimal method on

258 context.

259

260 **1. Analysis using MISC**

261 In MISC, each pathogenicity predictor is characterized by a line relating normalized

262 cost (rc) and clinical scenario (represented by an rc_1 value).

263 In Fig. 1a, we show the lines for the sixteen predictors analyzed, for ρ=0.5. The

264 relative positions of these lines at a given rc_1 value determine which predictor is preferable in

265 terms of cost. This comparison is extended to all possible rc_1 values with our algorithm CSP-

266 noco (Appendix 1, Supplementary Materials). The result obtained (Fig. 1a, bottom) shows

267 that changes in rc_1, i.e., in the clinical scenario, may affect the tool of choice for

268 pathogenicity prediction. More precisely, we see that in 81% of the cases PON-P2 is the cost-

269 optimal method; in the remainder, REVEL and CADD prevail in 16% and in 3% of the cases,

270 respectively. This picture contrasts with the fixed view conveyed by the AUC analysis (Fig.

271 1b), in which REVEL has the highest AUC and would always be the tool of choice,

272 regardless of the clinical context, if predictor selection is based only on AUC. When

273 reproducing the same analysis with MCC instead of AUC, we find (Supplementary Figure

274 S2a) that while choosing methods using MCC is not cost-optimal for some scenarios, on

275 average, it is better than using AUC, in this case.

276 Using cost models identifies ρ, the proportion of pathogenic variants, as a factor that
277 we may want to consider when choosing our preferred predictor. For example, if we change ρ
278 from 0.5 to 0.001, the ranking of the methods changes (Fig. 1b), and REVEL predominates in
279 99.5% cost scenarios and PON-P2 in 0.5%. In Fig. 2, we explore more systematically the
280 effect of ρ on the rankings, describing the transition between REVEL and PON-P2, an effect
281 due to the fact that as ρ grows methods with high sensitivities will be favored.
282 Finally, we would like to note that the previous analyses can be repeated as often as
283 required, e.g., when clinical costs change with time, or when we are working with specific
284 patient cohorts, etc., because our algorithm CSP-noco is computationally inexpensive.
285
286 **2. Analysis using MISC+NOA**
287 Again, our goal is to describe the distribution of optimal predictors across the clinical
288 space (which is now two-dimensional) obtained with our algorithm CSP-co (Appendix 2,
289 Supplementary Materials).
290 Once applied, the algorithm divides the clinical space into a mosaic of polygons (Fig.
291 3b) that is processed to give the distribution of predictors sought (Fig. 3c). We see that
292 introducing coverage considerations into the election of predictors, i.e., using MISC+NOA,
293 does not affect, overall, our previous result: deployment context affects the ranking of
294 predictors. There are, however, some quantitative/qualitative differences in the distribution of
295 the predictors. For example, for ρ=0.5 five tools (CADD, MutationTaster2, PON-P2,
296 REVEL, and Vest), instead of three, constitute the optimal division of the clinical space.
297 Again, this division is uneven, with REVEL prevailing in 77.3% scenarios, PON-P2 in 8.5%,
298 etc.
299 Comparison of these results with those of the AUC ranking (Fig. 3d) confirms that
300 there are clinical scenarios for which the predictor with top AUC is not cost-optimal. NOA

301 plays an important role in this result. PON-P2, which has moderate coverage, prevails in cost
302 scenarios near the diagonal (Fig. 3c). That is, for those cases for which the cost of NOA is
303 negligible. REVEL, on the contrary, because of its higher coverage, predominates over most
304 of the clinical space. As a consequence, using AUC for choosing predictors gives more cost-
305 optimal results than for MISC. Reproducing the same analysis with MCC instead of AUC
306 shows (Supplementary Figure S2b) that choosing predictors using MCC is not cost-optimal
307 for all scenarios either, as for MISC.

308 Characterization of the impact of different ρ values (Fig. 4) complete the description
309 of this section. We find that the overall picture does not vary so much between different ρ's
310 because only two to five methods out of sixteen participate in the final division of the clinical
311 space. Their identities change slightly. All across the ρ range, REVEL predominates, with
312 PON-P2 also appearing, near the diagonal. As ρ increases, three other methods (CADD,
313 MutationTaster2, and VEST) start occupying the area nearby the bottom-right vertex of the
314 clinical space. The overall predominance of REVEL has different origins: higher coverage
315 relative to PON-P2 (0.99 vs. ~0.50); better specificity relative to CADD and MutationTaster2
316 (0.88 vs. 0.25 and 0.60, respectively); and a mixture of factors relative to VEST.

317

318 **DISCUSSION**

319 Nowadays, pathogenicity predictors are widely utilized, in combination with other
320 sources of evidence, to address the VIP (Richards *et al.*, 2015). However, the growing
321 number of predictors available and their varied performances (Özkan *et al.*, 2021) pose a
322 challenge to interested users, who have to identify the optimal tool for their needs among tens
323 of options. Usually, this is done ranking pathogenicity predictors on the basis of their AUCs,
324 or MCCs, etc., parameters that are blind to the deployment context. Here, we show,
325 applying/extending cost models used in healthcare (Pepe, 2003; Hunink *et al.*, 2014), that this

326 procedure is not optimal and that changes in clinical scenarios can indeed affect the tool of
327 choice. Our work is inspired (i) by the use of cost in the similar problem of deciding between
328 candidate tests for the same task (Pepe, 2003), e.g., for Elisa assays, etc., and (ii) by the use
329 of cost to compare the performance of binary classifiers in the machine learning field (Adams
330 and Hand, 1999; Drummond and Holte, 2006). We have explored two options: MISC, in
331 which only misclassification errors are contemplated, and MISC+NOA, where, in addition,
332 the situation in which the predictor does not produce an outcome is also contemplated. In
333 both cases, clinical context is encoded using a few cost-related parameters (two for MISC and
334 three for MISC+NOA). For each model, we have developed an algorithm (CSP-noco and
335 CSP-co, respectively) that allows the comparison of an arbitrary number of predictors across
336 the whole set of clinical scenarios. We have then applied these results to a set of sixteen
337 pathogenicity predictors used for the VIP. The views resulting from MISC (Fig. 1) and
338 MISC+NOA (Fig. 3) coincide in that cost-optimal methods may vary depending on the
339 region of the clinical space considered. This view contrasts with the unique classification of
340 predictors emanating from the use of AUC (Figs. 1b, 3d) or MCC (Supplementary Figure S2)
341 and indicates that context considerations are valuable when choosing pathogenicity predictors
342 for specific applications. In clinical settings where estimating the cost parameters (rc_0, rc_1,
343 and rc_2) is challenging, users may explore the cost-division of the clinical space using the
344 more accessible ratios of these parameters (Adams and Hand, 1999). Or, they can replace
345 exact values by ranges of values.

346 An advantage of cost models is that they provide an application-aware approach to
347 integrating two key components of the tool ranking problem, misclassification error and
348 coverage. For this reason, they can be used in the design of variant annotation strategies, to
349 shed new light on their applicability. For example, we know that identification of germline
350 pathogenic variants in TP53 plays a significant role in the diagnosis of the Li-Fraumeni

351 syndrome (Schneider *et al.*, 1999). To improve the accuracy of this procedure, Fortuno et al.
352 (Fortuno *et al.*, 2018) test a group of pathogenicity predictors to establish which one is
353 preferable for the scoring of missense variants. Using MCC as primary criterion, these
354 authors conclude that the combination of predictors Align-GVGD+BayesDel (A+B) is better
355 than the other options, closely followed by Align-GVGD+REVEL (A+R). Based on this
356 work, the guidelines for the scoring of TP53 variants in the clinical recommend A+B as the
357 tool of choice (Fortuno *et al.*, 2021) for the generation of computational evidence on
358 missense variants. When we apply cost models to this problem we see a different situation
359 (Fig. 5). With MISC, which only considers misclassification errors, the picture obtained with
360 the MCC is now less clear: A+B predominates over most of the clinical space (80.7%), but
361 A+R is optimal in 19.3% of the scenarios. The situation changes even more when using
362 MISC+NOA, which also takes into account discordances between paired predictors (NOA).
363 The distribution of the clinical space experiences a significant change (Fig. 5) relative to the
364 MISC case. The accumulated size of A+B and A+R drops from 100% to 31.1%, and two new
365 regions appear, corresponding to BayesDel (50%) and REVEL (18.9%), respectively. These
366 new regions correspond to clinical contexts where NOA costs are high relative to
367 misclassification costs. For example, these could be scenarios in which surveillance programs
368 for pediatric patients with an unclear Li-Fraumeni diagnosis are more expensive than usual,
369 etc. In these settings, rejecting *in silico* evidence when Align-GVGD and its companion tool
370 (REVEL or BayesDel) disagree may be a worse option than using REVEL or BayesDel
371 alone. We are not advocating the use of these two predictors in these clinical space regions;
372 we want to emphasize that cost models identify a problem in the election of in silico tools.
373 There are different options to address this problem, some of which are directly suggested by
374 the cost model (equation (2)). For example, we can focus on the method-dependent
375 parameters, sensitivity (s_e), specificity (s_p), and coverage (α), and look for combinations of

predictors that guarantee good s_e and s_p while increasing α. This is a realistic alternative since presently there are well over 50 pathogenicity predictors available (see Özkan et al. (Özkan *et al.*, 2021), and references therein). Alternatively, we can aim to increase α by developing methods specific for variants with conflicting predictions (de la Campa, Padilla and de la Cruz, 2017). We can also try to act directly on the cost parameters; e.g., in the case of rc_2 we could find strategies that reduce the costs of managing patients with no diagnosis, etc. In summary, the use of cost models helps identify key issues in the choice of predictors and provides insight on how to solve them.

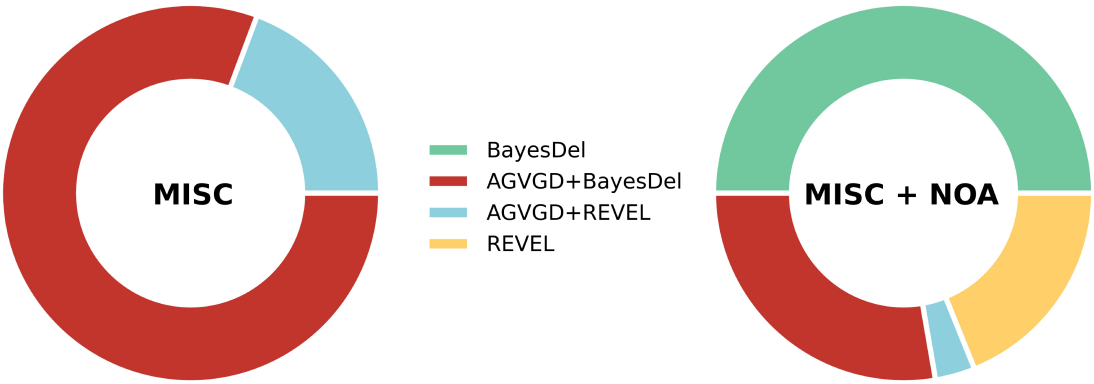


Fig. 5. Application of the cost models MISC and MISC+NOA to the predictors used in Fortuno et al. (Fortuno *et al.*, 2018) **study for TP53 variants.** The rings corresponding to MISC (left) and MISC+NOA (right) are divided according to the percentages of cost scenarios corresponding to each predictor.

Finally, it is worth noting that the methodology presented for comparing multiple tools using cost models is general and independent of the nature of the tools compared (we only need to know their s_e and s_p, for MISC, plus α for MISC+NOA). That is, it can be applied to compare pathogenicity predictors, pathogenicity predictors with functional assays,

395 or sets of bioinformatics tools used in other biomedical/clinical fields. The latter may be an
396 increasing need in medical imaging (Liu *et al.*, 2019), diagnosis/prognosis of covid-19
397 (Wynants *et al.*, 2020), etc., where dozens of classifiers are developed, published, and their
398 performance compared to that of human healthcare experts. We believe that our algorithms
399 provide objective and easy-to-use ways to compare the performance of human experts and
400 their artificial intelligence counterparts over the whole clinical application range.
401

402 **Funding**
403 This work was supported by research grants SAF2016-80255-R from the Spanish Ministerio
404 de Economía y Competitividad (MINECO), PID2019-111217RB-I00 from the Spanish
405 Ministerio de Ciencia e Innovación, and by the European Regional Development Fund
406 (ERDF) through the Interreg program POCTEFA (Pirepred, EFA086/15).
407

408 **Acknowledgements**
409 The authors acknowledge comments on the work from members of the Pirepred european
410 consortium.
411

412 **References**
413 Adams, N. M. and Hand, D. J. (1999) 'Comparing classifiers when the misallocation costs
414 are uncertain', *Pattern Recognition*, 32(7), pp. 1139–1147. doi: 10.1016/S0031-
415 3203(98)00154-X.
416 Adzhubei, I. A. *et al.* (2010) 'PolyPhen-2 : prediction of functional effects of human
417 nsSNPs', *Nat. Methods*, 7(4), pp. 248–249. doi: 10.1038/nmeth0410-248.
418 Amendola, L. M. *et al.* (2016) 'Performance of ACMG-AMP Variant-Interpretation
419 Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research

420 Consortium', *The American Journal of Human Genetics*, 98(6), pp. 1067–1076. doi:
421 10.1016/j.ajhg.2016.03.024.
422 Berrios, C. *et al.* (2021) 'Challenges in genetic testing: clinician variant interpretation
423 processes and the impact on clinical care', *Genetics in Medicine*, pp. 1–11. doi:
424 10.1038/s41436-021-01267-x.
425 Carter, H. *et al.* (2013) 'Identifying Mendelian disease genes with the variant effect scoring
426 tool.', *BMC genomics*, 14 Suppl 3(Suppl 3), p. S3. doi: 10.1186/1471-2164-14-s3-s3.
427 Choi, Y. *et al.* (2012) 'Predicting the functional effect of amino Acid substitutions and
428 indels.', *PloS one*, 7(10), p. e46688. doi: 10.1371/journal.pone.0046688.
429 Chun, S. and Fay, J. C. (2009) 'Identification of deleterious mutations within three human
430 genomes', *Genome Research*, 19(9), pp. 1553–1561. doi: 10.1101/gr.092619.109.
431 Dong, C. *et al.* (2015) 'Comparison and integration of deleteriousness prediction methods for
432 nonsynonymous SNVs in whole exome sequencing studies', *Human Molecular Genetics*,
433 24(8), pp. 2125–2137. doi: 10.1093/hmg/ddu733.
434 Drummond, C. and Holte, R. C. (2006) 'Cost curves: An improved method for visualizing
435 classifier performance', *Machine Learning*, 65(1), pp. 95–130. doi: 10.1007/s10994-006-
436 8199-5.
437 Feliubadaló, L. *et al.* (2021) 'A Collaborative Effort to Define Classification Criteria for
438 ATM Variants in Hereditary Cancer Patients', *Clinical Chemistry*, 67, pp. 518–533. doi:
439 10.1093/clinchem/hvaa250.
440 Fortuno, C. *et al.* (2018) 'Improved, ACMG-compliant, in silico prediction of pathogenicity
441 for missense substitutions encoded by TP53 variants', *Human Mutation*, 39(8), pp. 1061–
442 1069. doi: 10.1002/humu.23553.
443 Fortuno, C. *et al.* (2021) 'ClinGen TP53 Variant Curation Expert Panel. Specifications of the
444 ACMG/AMP variant interpretation guidelines for germline TP53 variants', *Human Mutation*,

42(3), pp. 223–236. doi: 10.1002/humu.24152.Specifications.

Ghosh, R., Oak, N. and Plon, S. E. (2017) 'Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines', *Genome Biology*, 18, p. 225. doi: 10.1186/s13059-017-1353-5.

Hand, D. J. (2012) 'Assessing the Performance of Classification Methods', *International Statistical Review*, 80, pp. 400–414. doi: 10.1111/j.1751-5823.2012.00183.x.

Hernández-Orallo, J., Flach, P. and Ferri, C. (2012) 'A unified view of performance metrics: Translating threshold choice into expected classification loss', *Journal of Machine Learning Research*, 13, pp. 2813–2869.

Hunink, M. G. M. *et al.* (2014) *Decision making in health and medicine: Integrating evidence and values, second edition*, *Decision Making in Health and Medicine: Integrating Evidence and Values, Second Edition*. Cambridge: Cambridge University Press.

Ioannidis, N. M. *et al.* (2016) 'REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants', *American Journal of Human Genetics*, 99(4), pp. 877–885. doi: 10.1016/j.ajhg.2016.08.016.

Kumar, P., Henikoff, S. and Ng, P. C. (2009) 'Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm', *Nature Protocols*, 4(7), p. 1073. doi: 10.1038/nprot.2009.86.

de la Campa, E. álvarez, Padilla, N. and de la Cruz, X. (2017) 'Development of pathogenicity predictors specific for variants that do not comply with clinical guidelines for the use of computational evidence', *BMC Genomics*, 18(Suppl 5), pp. 1–14. doi: 10.1186/s12864-017-3914-0.

Liu, X. *et al.* (2016) 'dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs', *Human Mutation*, 37(3), pp. 235–241. doi: 10.1002/humu.22932.

470 Liu, X. *et al.* (2019) 'A comparison of deep learning performance against health-care
471 professionals in detecting diseases from medical imaging: a systematic review and meta-
472 analysis', *The Lancet Digital Health*, 1, pp. e271–e297. doi: 10.1016/S2589-7500(19)30123-
473 2.
474 López-Ferrando, V. *et al.* (2017) 'PMut: A web-based tool for the annotation of pathological
475 variants on proteins, 2017 update', *Nucleic Acids Research*, 45(W1), pp. W222–W228. doi:
476 10.1093/nar/gkx313.
477 Niroula, A., Urolagin, S. and Vihinen, M. (2015) 'PON-P2: Prediction Method for Fast and
478 Reliable Identification of Harmful Variants', *Plos One*, 10(2), p. e0117380. doi:
479 10.1371/journal.pone.0117380.
480 Özkan, S. *et al.* (2021) 'The computational approach to variant interpretation: principles,
481 results, and applicability', in Lázaro, C., Lerner-Ellis, J., and Spurdle, A. (eds) *Clinical DNA*
482 *Variant Interpretation: Theory and Practice*. San Diego: Elsevier Inc./Academic Press, pp.
483 89–119.
484 Pejaver, V. *et al.* (2020) 'Inferring the molecular and phenotypic impact of amino acid
485 variants with MutPred2', *Nature Communications*, 11(1). doi: 10.1038/s41467-020-19669-x.
486 Pepe, M. S. (2003) *The Statistical Evaluation of Medical Tests for Classification and*
487 *Prediction*. Oxford: Oxford University Press.
488 Rentzsch, P. *et al.* (2019) 'CADD: Predicting the deleteriousness of variants throughout the
489 human genome', *Nucleic Acids Research*, 47(D1), pp. D886–D894. doi:
490 10.1093/nar/gky1016.
491 Reva, B., Antipin, Y. and Sander, C. (2011) 'Predicting the functional impact of protein
492 mutations: Application to cancer genomics', *Nucleic Acids Research*, 39(17), p. e118. doi:
493 10.1093/nar/gkr407.
494 Richards, S. *et al.* (2015) 'Standards and guidelines for the interpretation of sequence

495 variants: A joint consensus recommendation of the American College of Medical Genetics
496 and Genomics and the Association for Molecular Pathology', *Genetics in Medicine*, 17(5),
497 pp. 405–424. doi: 10.1038/gim.2015.30.
498 Schneider, K. *et al.* (1999) 'Li-Fraumeni syndrome', *Gene Reviews*, (University of
499 Washington, Seattle, Seattle (WA)).
500 Schwarz, J. M. *et al.* (2014) 'Mutationtaster2: Mutation prediction for the deep-sequencing
501 age', *Nature Methods*, 11(4), pp. 361–362. doi: 10.1038/nmeth.2890.
502 Shendure, J., Findlay, G. M. and Snyder, M. W. (2019) 'Genomic Medicine–Progress,
503 Pitfalls, and Promise', *Cell*, 177, pp. 45–57. doi: 10.1016/j.cell.2019.02.003.
504 Shihab, H. A. *et al.* (2013) 'Predicting the Functional, Molecular, and Phenotypic
505 Consequences of Amino Acid Substitutions using Hidden Markov Models', *Human*
506 *Mutation*, 34(1), pp. 57–65. doi: 10.1002/humu.22225.
507 Starita, L. M. *et al.* (2017) 'Variant Interpretation: Functional Assays to the Rescue',
508 *American Journal of Human Genetics*, 101(3), pp. 315–325. doi: 10.1016/j.ajhg.2017.07.014.
509 Vihinen, M. (2020) 'Problems in variation interpretation guidelines and in their
510 implementation in computational tools', *Molecular Genetics & Genomic Medicine*, 8, p.
511 e1206.
512 Wynants, L. *et al.* (2020) 'Prediction models for diagnosis and prognosis of covid-19:
513 Systematic review and critical appraisal', *The BMJ*, 369, p. m1328. doi: 10.1136/bmj.m1328.
514