Received: 14 February 2020 | Accepted: 23 February 2020

DOI: 10.1002/mgg3.1206

ORIGINAL ARTICLE

Molecular Genetics & Genomic Medicine Open Access WILEY

Problems in variation interpretation guidelines and in their implementation in computational tools

Mauno Vihinen

Department of Experimental Medical Science, Lund University, Lund, Sweden

Correspondence
Mauno Vihinen, Department of Experimental Medical Science, BMC B13, Lund University, SE-22184 Lund, Sweden.
Email mauno.vihinen@med.lu.se

Funding information
This work was supported by Vetenskapsrådet (Swedish Research Council, VR 2015-02510), Swedish Cancer Society (CAN 2017/699), and Alfred Österlunds Stiftelse.

Abstract

Background: ACMG/AMP and AMP/ASCO/CAP have released guidelines for variation interpretation, and ESHG for diagnostic sequencing. These guidelines contain recommendations including the use of computational prediction methods. The guidelines per se and the way they are implemented cause some problems.

Methods: Logical reasoning based on domain knowledge.

Results: According to the guidelines, several methods have to be used and they have to agree. This means that the methods with the poorest performance overrule the better ones. The choice of the prediction method(s) should be made by experts based on systematic benchmarking studies reporting all the relevant performance measures. Currently variation interpretation methods have been applied mainly to amino acid substitutions and splice site variants; however, predictors for some other types of variations are available and there will be tools for new application areas in the near future. Common problems in prediction method usage are discussed. The number of features used for method training or the number of variation types predicted by a tool are not indicators of method performance. Many published gene, protein or disease-specific benchmark studies suffer from too small dataset rendering the results useless. In the case of binary predictors, equal number of positive and negative cases is beneficial for training, the imbalance has to be corrected for performance assessment. Predictors cannot be better than the data they are based on and used for training and testing. Minor allele frequency (MAF) can help to detect likely benign cases, but the recommended MAF threshold is apparently too high. The fact that many rare variants are disease-causing or -related does not mean that rare variants in general would be harmful. How large a portion of the tested variants a tool can predict (coverage) is not a quality measure.

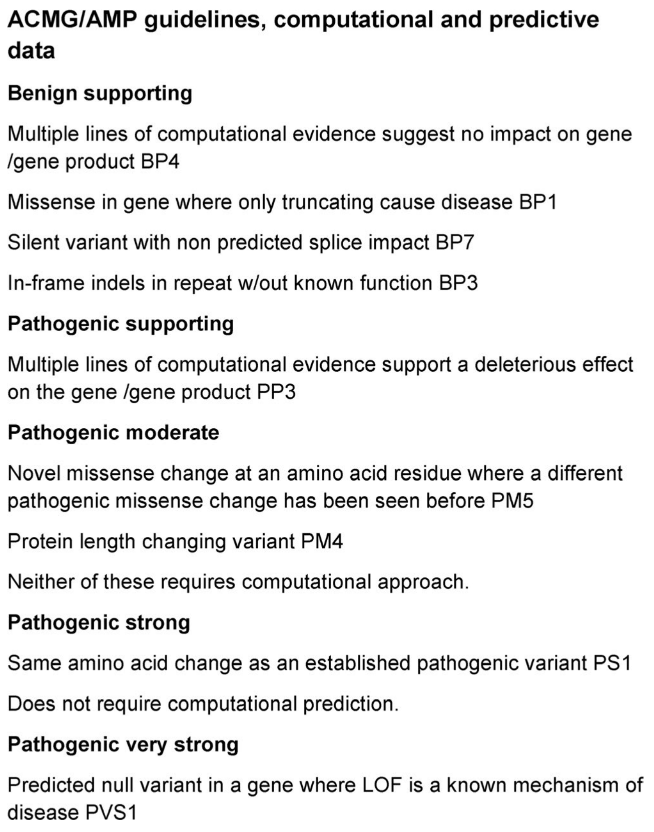
Conclusion: Methods used for variation interpretation have to be carefully selected. It should be possible to use only one predictor, with proven good performance or a limited number of complementary predictors with state-of-the-art performance. Bear in mind that diseases and pathogenicity have a continuum and variants are not dichotomic i.e. either pathogenic or benign, either.

KEYWORDS
ACMG/AMP guidelines, benchmark datasets, continuum of disease, majority vote, pathogenicity model, pathogenicity prediction, prediction methods, variation interpretation

1 | BACKGROUND

Variation interpretation has become the bottleneck of using genetic information. Sequencing methods are highly efficient and increasingly accurate. Guidelines and standards have been published for the use and interpretation of variation data by the American College of Medical Genetics and Genomics, and the Association for Molecular Pathology (ACMG/AMP) (Richards et al., 2015) and of AMP, the American Society of Clinical Oncology, and the College of American Pathologists (AMP/ASCO/CAP) (Li et al., 2017). The European Society for Human Genetics (ESHG) has released guidelines for diagnostic next-generation sequencing and they include also variation interpretation (Matthijs et al., 2016). These guidelines for systematic schemes and especially the ACMG/AMP guideline, although an American recommendation, are widely followed in many countries and laboratories.

Variation interpretation is a difficult task that brings together many and different kinds of data, methods, and expertise. Computational methods have been used for many years to predict the disease relevance, pathogenicity, or tolerance of variants. Computational solutions are essential due to the vast numbers of variants and as with experimental methods it is impossible to investigate significance of all the millions of variants that every genome contains in relation to a reference genome sequence.



ACMG/AMP guidelines, computational and predictive data

Benign supporting

Multiple lines of computational evidence suggest no impact on gene /gene product BP4

Missense in gene where only truncating cause disease BP1

Silent variant with non predicted splice impact BP7

In-frame indels in repeat w/out known function BP3

Pathogenic supporting

Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3

Pathogenic moderate

Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5

Protein length changing variant PM4

Neither of these requires computational approach.

Pathogenic strong

Same amino acid change as an established pathogenic variant PS1

Does not require computational prediction.

Pathogenic very strong

Predicted null variant in a gene where LOF is a known mechanism of disease PVS1

FIGURE 1 ACMG/AMP guidelines for the use of computational and prediction methods, taken from Richards et al. (2015). Of the individual items, BP4, BP7, and PP3 include the use of computational tools

Since their release, a number of changes have been suggested to the ACMG/AMP guidelines including changes to criterion PVS1 (Abou Tayoun et al., 2018), to the entire process (Nykamp et al., 2017), allele frequency thresholds (Kelly et al., 2018; Kobayashi et al., 2017; Nykamp et al., 2017; Rim et al., 2019; Whiffin et al., 2017), etc.

Here, I discuss problems related to the recommendations about the use of computational methods, including problems with the schemes per se and problems on how they are implemented in practice. The way the guidelines are implemented and followed today hampers optimal variation interpretation and thereby does not serve the best of the patients as the full power of predictions, together with other evidence, is not taken into account. Therefore, pathogenicity of a substantial portion of the variants is not considered due to overly conservative recommendations and practices.

2 | GUIDELINES FOR USE OF COMPUTATIONAL PREDICTORS IN VARIATION INTERPRETATION

In this article, the focus is on the use of prediction methods for variation interpretation. The ACMG/AMP guidelines for using computational and predictive data are shown in Figure 1. In the scheme, the criteria are divided into two categories for benign (supporting and strong) and into four categories for pathogenic classification (supporting, moderate, strong, and very strong). Of the items listed in Figure 1, only BP4, BP7, and PP3 require and use computational predictors. According to the guidelines, prediction methods should be considered as supporting, when multiple lines of computational evidence support the effect. Furthermore, predictions should *be used carefully and not as a sole source of evidence to make a clinical assertion.* Quotations from the guidelines are indicated in italics, along with the source. This one is from ACMG/AMP.

Some available prediction methods are listed in all three guidelines, in ACMG/AMP in Table 2, in AMP/ASCO/CAP in Table 2, and in ESHG guidelines in Supplementary Table S1. These methods include predictors for amino acid substitutions, splice sites, and sequence conservation.

Computational methods are widely used because of their availability, as they can handle large numbers of variations in exome and genome scale, and because they can provide rather reliable predictions. Functional studies would be the optimal choice, however, they are costly, time consuming, and often not reimbursed. Every genome and exome still contains a huge number of unique variants, analysis of which is not experimentally feasible.

Comparison of the implementation and obtained diagnosis with ACMG/AMP guidelines in nine laboratories indicated that the PP3 criterion was the second most widely used (39% of cases) and BP4 was used in 16% of cases, thus, totally 55% of the investigated variants were interpreted based on computational predictions, along with other evidence (Amendola et al., 2016). Therefore, it is important to implement these methods in a proper way. As discussed below, there are currently a number of practices that prevent optimal use of predictive data. With the suggested alteration to the practices, it would be possible to use computational tools for interpretation of numerous additional variations. In the case of exome- or genome-wide sequencing this will make a big difference.

3 | PROBLEMS WITH THE GUIDELINES AND IN THEIR IMPLEMENTATION

According to the guidelines, multiple predictors should agree, and if they do not, prediction methods should not be used at all for the assessment of that variant. This recommendation opens for severe problems. First, the methods with the poorest performance get the largest impact. Second, clinical and research laboratories are not necessarily experts in bioinformatics and thereby the choice of the methods has often not been optimal and even been problematic. Third, computational tools mentioned in the guidelines have been considered as recommendations when clearly better tools are available. Fourth, if the same or similar methods are used together, the outcome is biased. Fifth, the schemes implicitly promote an idea of dichotomic distribution of variations either to the benign or the pathogenic class.

3.1 | Poor methods overrule good ones

Critical Assessment of Genome Interpretation (CAGI, https://genomeinterpretation.org/) challenges have shown that prediction methods tend to agree with each other more than with the reality. On the other hand, analysis of the concordance of predictors indicated that 18 tested methods agreed on ClinVar data only for about 5% of benign cases and 39%–47% of pathogenic cases, depending on the dataset (Ghosh, Oak, & Plon, 2017). In a benchmark study of 10 methods with about 40,000 variants there was not a single variant that all the methods agreed upon (Thusberg, Olatubosun, & Vihinen, 2011). In another analysis with almost 60,000 variants, 10%–45% of predictions were contradictory depending on the choice of methods (de la Campa, Padilla, & Cruz, 2017).

The more methods, the smaller are the chances that they all agree. Thereby the coverage of variants that can be predicted reduces quickly when several methods are used together.

The most important outcome of the guidelines is that they give the decision power to the poorest methods (Figure 2). The methods with low-performance disagree most often with the well-performing ones and therefore they drag the overall performance down. Thus, the requirement for the predictors to agree overemphasizes the significance of poor methods and prevents the use of predictions for cases on which the high-quality methods agree, but which are not correctly predicted by the poor one(s).

3.2 | Lacking expertise in method choice and use

Amino acid substitutions have been the most studied variation type, and there are well over 100 predictors described in literature. Services like dbNSFP (Liu, Jian, & Boerwinkle, 2011) and VarCards (Li et al., 2018) contain predictions for more than 20 tools making it easy to obtain the predictions. Domain knowledge is required to choose the best tools. The most widely used methods are about 10 or 20 years old. Nobody would use 20-year-old sequencer, but the same people happily resort to old prediction methods, although the advancements have been even bigger in the interpretation field than in sequencing technology. The old methods are simply old methods and far behind in the performance in comparison to many new ones. As the benchmark studies mentioned above have indicated, there are 30% or even larger differences between the state-of-the-art tools and some others. In no other field in medicine it is possible to use methods with such low performance.

Therefore, when using multiple predictors (see Figure 2), one has to choose methods that have state-of-the-art performance and which are not based on too similar principles. To do this in a best way, it is necessary to understand how the methods work and how they have been implemented and trained. This expertise is lacking in many diagnostic laboratories.

3.3 | Named methods are considered as recommendations

Many laboratories seem to consider the methods mentioned in the guidelines as recommendations. This has apparently,

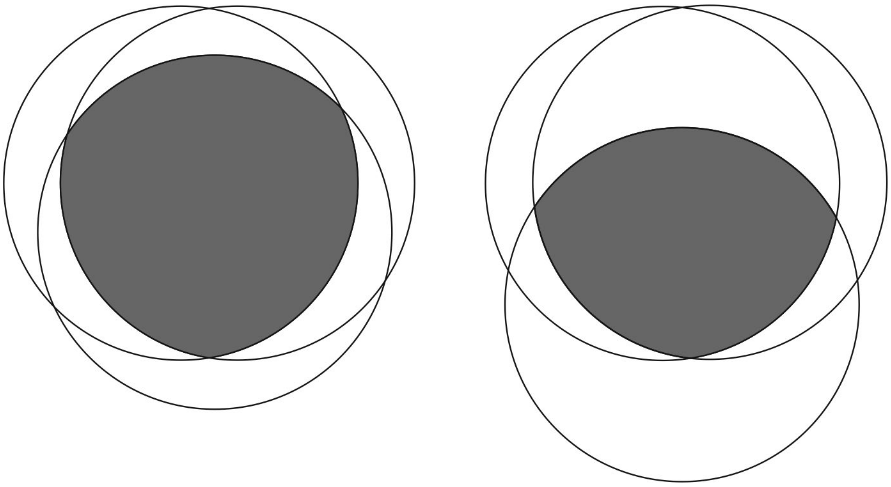


FIGURE 2 Visualization of the problem when requiring several predictions methods to agree. The Venn diagram to the left indicates that prediction methods with high performance agree on many variations and therefore can predict large portion of cases. The intersection indicated in grey is for predictions that all the tools agree on. If the set of tools contains a poor method, as to the right, the number of cases that the tools agree is significantly reduced and the reduction is due to the poor tool

or at least hopefully, not been the intention. One could even question how the listed methods have been chosen as many of them did not represent the state of the art at the time that the guidelines were published, or not even when the methods were released.

3.4 | Problematic majority vote

A further complication emerges when the predictions are combined. The majority vote approach, as implemented in the guidelines, may introduce problems, see Vihinen (2014). If similar or the same predictions are combined, the outcome is biased. In addition to including constituent methods and meta-predictors that utilize the same predictions, this may appear when methods with similar foundations are combined. For example, if one combines evolutionary data-based predictions together, the outcome will likely be concordant but definitely biased.

Methods are often chosen ad hoc, or taking those mentioned in the guidelines as recommendations, or using the same method as last time, or for last 10 years. Popularity does not equal performance. It requires expertise to pick the best-performing methods and combinations that do not suffer from bias. As the field is moving fast, it is not relevant to name methods in here, instead to refer to benchmark studies (see below).

3.5 | Pathogenic-benign binarism

"*Efforts to resolve the classification of the variant as pathogenic or benign should be undertaken*" (ACMG/AMP) and "*A community activity is needed to collect and share the available information, with the aim to definitely classify the variants into pathogenic or benign*" (ESHG, Statement 28) indicate the common but flawed thinking to divide variants into binary, mutually exclusive categories of benign and pathogenic. The reality is that in every condition there is a continuum, discussed, and described in a pathogenicity model (Vihinen, 2017). There are variations that cannot be placed in either category because of variable phenotype, incomplete penetrance, and other factors. This heterogeneity means that some variants can be benign for an individual but pathogenic or unclassifiable for another or pathogenic for an individual at later time. Therefore, it will not be possible to group all variants into only the two groups.

ACMG guidelines state that *Most tools … are not reliable at predicting missense variants with milder effect*. This is indeed a more difficult prediction task than distinguishing pathogenic from neutral variants, but this area has obtained attention and the first tool is available for generic variant phenotype severity prediction, called PON-PS (Niroula & Vihinen, 2017).

4 | COMMENTS TO ESHG GUIDELINES

Supplementary guidelines, Statement 18: *The bioinformatics pipeline must be tailored for the technical platform used. During pipeline validation the diagnostic specifications must be measured by assessing analytical sensitivity and specificity.* This is a good recommendation, however, not sufficient for complete performance assessment; see the discussion below about measures to be used for performance assessment. More comprehensive analysis of the performance should be made.

Statement 20: *The diagnostic laboratory has to validate all parts of the bioinformatics pipeline (public domain tools or commercial software packages) with standard datasets whenever relevant changes (new releases) are implemented.* This recommendation is reasonable and necessary. There are a number of issues to be considered when testing and there are guidelines for doing that (Vihinen, 2012, 2013). The dataset used for testing should not contain cases used for training the included methods; this should be checked every time the assessment is done. Furthermore, these kinds of datasets are valuable but laborious to generate, therefore they should be made available, preferably via the VariBench database (Nair & Vihinen, 2013; Sarkar, Yang & Vihinen, 2020). Some methods utilize sequence conservation information based on blasting of sequence databases, which are changing constantly. In such cases, evaluations should be done at frequent intervals.

5 | HOW TO TEST AND BENCHMARK PREDICTORS?

The first issue when using prediction methods is to choose suitable tools because for many variation types and effects numerous tools are available and based on many different principles. For example, for amino acid substitutions there are three wide categories of tools including evolutionary conservation-based approaches, machine learning methods, and meta-predictors, see Niroula and Vihinen (2016). The choice of methods should be based on their proven performance. For this purpose a number of benchmark studies have been performed for different types of variations and predictors including protein substitution tolerance/pathogenicity (Bendl et al., 2014; Grimm et al., 2015; Masica & Karchin, 2016; Niroula, Urolagin, & Vihinen, 2015; Riera, Padilla, & Cruz, 2016; Thusberg et al., 2011; Sarkar et al. 2020), variants affecting protein stability (Khan & Vihinen, 2010; Potapov, Cohen, & Schreiber, 2009), protein localization (Laurila & Vihinen, 2009), protein disorder (Ali, Urolagin, Gurarslan, & Vihinen, 2014), protein solubility (Yang, Niroula, Shen, & Vihinen, 2016), benign variants (Niroula & Vihinen, 2019), variants in transmembrane proteins (Orioli & Vihinen, 2019), alternative splicing (Desmet, Hamroun, Collod-Beroud, Claustres, & Beroud, 2010; Jian, Boerwinkle, & Liu, 2014), and phenotypes of amino acid substitutions (Anderson & Lassmann, 2018).

Benchmark studies have shown huge differences in performances. A recent analysis of benign variants indicated that the best tool had accuracy of about 0.96 while some other widely used methods misclassified one of three cases (Niroula & Vihinen, 2019). Similar results have been obtained in other benchmark studies.

Benchmark studies have to be based on gold standard cases with known, experimentally studied outcome. Such datasets are available in VariBench (Nair & Vihinen, 2013) and VariSNP (Schaafsma & Vihinen, 2015) and have been widely used. Recently, VariBench was updated with 419 new datasets (Sarkar et al. 2020). Benchmarks have to fulfill a number criteria including relevance, representativeness, nonredundancy, containing experimentally verified cases both with positive and negative effect, and be scalable and reusable (Nair & Vihinen, 2013).

To obtain a full picture of performances of binary prediction methods at least six measures have to be provided and many other reporting requirements have to be fulfilled when publishing such methods (Vihinen, 2012, 2013). These guidelines are also followed outside bioinformatics and medicine, that is, in technology. The recommendations include reporting method description, used datasets, performance assessment, and implementation. If the same or similar data are used for both method training and testing, the performance is inflated. Circularity (Grimm et al., 2015) has been common in publications. Method assessments have to be based on a substantial number of cases. Performance assessments are statistical in nature, therefore, a handful of instances are not sufficient.

6 | PROBLEMS IN METHOD PEFORMANCE ASSESSMENT AND SELECTION

Despite the guidelines for performance assessment at available, misconceptions and wrong practices are common. The number of features used for method training or the number of variation types predicted by a tool does not tell anything about its performance. Many gene, protein, or disease-specific benchmark studies suffer from too small dataset rendering the results useless. In the case of binary predictors, there should be equal number of positive and negative cases or the imbalance has to be corrected.

6.1 | Number of features and predictions

The number of features used in a predictor or a number of predictions made does not correlate in any way with the performance of the method. Machine learning methods are trained on features that are related to the investigated phenomenon. Method developers do not initially know which features are relevant, therefore, feature selection step usually is implemented. During this process the most important features are identified and then used to train the final predictor. There are even some tools for which feature selection has not been made.

The reason for performing feature selection is that the number of informative features is typically rather small. With an increasing number of features comes the so-called curse of dimensionality. The more there are features, the larger dataset is needed to represent the space of all possible combinations of features. The number of required cases grows exponentially along with the number of features. Therefore, it is beneficial to use the smallest possible number of relevant features. It makes also the predictions faster and more reliable as many weak and less informative features can reduce the method performance.

The number of prediction types made is no quality indicator, either. It is possible to include large numbers of predictions, however, from the end-user point of view it is relevant how good those predictions are and whether they have been systematically benchmarked. There are only a few variation prediction areas where systematic analyses have been conducted and have been possible. The limiting factor is the availability of suitable test datasets.

6.2 | Too small datasets

During the last few years there has been a flurry of papers where variant prediction methods have been tested on individual genes/proteins/diseases or on small groups of diseases. Most of these studies do not have any predictive power since the tests have been based on too small numbers of cases. The smallest number this author has seen was nine cases! As benchmarking is a statistical approach, a substantial number of cases is needed; the actual number depends on the application and test. In practice, there has to be at least in the order of 100 variants of certain effect/type to achieve reasonable performance and reliability, see the analysis of Riera et al. (2016).

Often the same authors who analyses method performance use the same too small dataset to train a novel predictor. The best generic tolerance/pathogenicity predictions are most of the time better than gene/protein-specific predictors (Riera et al., 2016). However, the performance of predictors varies for different genes/proteins, and there are examples where specific predictors are better. There are a number of reasons for generic methods to outperform specific tools. They are based on larger datasets and thereby can generalize more reliably. The mechanisms of variation effects are similar in all genes and proteins and thereby generic methods have a benefit of including variations in different contexts.

As an example we can look at single amino acids substitutions. There are altogether 380 alterations of which 150 (39.5%) are possible by a single nucleotide change and are thus much more likely to occur. These variants appear in many different contexts in proteins and have different effects. Thus, ideally each variant type should be included several times and in different environments, diseases, and sequence/structure contexts. This means that for most genes and proteins there are not large enough datasets available.

6.3 | Imbalance of positive and neutral variants

Disease-causing variants are in many instances in minority compared with all the possible variants. When developing machine learning methods, the positive (having effect) and negative (not having an effect) cases should appear in equal numbers (Wei, Wang, Wang, Kruger, & Dunbrack, 2010). When testing method performance, there should be either equal numbers of positive and negative cases or the difference has to be somehow mitigated. Some of the performance scores are sensitive for class imbalance and can give highly misleading scores if not taken care of (Vihinen, 2012). It is quite common in literature that the class imbalance has not been mitigated when method performances are compared. Thereby some scores may be severely affected.

7 | OTHER PROBLEMS AND ISSUES IN COMPUTATIONAL VARIATION INTERPRETATION

There are several other issues related to variation interpretation in general, not just about the guidelines. These are described and justified briefly. Predictors cannot be better than the data they are based on and used for training and testing. Minor allele frequency (MAF) can help to detect likely benign cases. The recommended MAF threshold is apparently too high. The fact that many rare variants are disease-causing or -related cannot be extrapolated to mean that rare variants in general would be harmful. How large a portion of the tested variants a tool can predict (coverage) is not a quality measure. Finally, the clinical and research communities should pay more attention to the language used for describing and naming variations.

7.1 | Training data quality

Prediction methods cannot be better than the data they are based on. High-quality variant datasets of substantial size are laborious and costly to collect as typically many manual steps are needed. Most of the existing variation datasets are available in VariBench (Nair & Vihinen, 2013; Sarkar et al. 2020) and VariSNP (Schaafsma & Vihinen, 2015).

As discussed above, representativeness is one of the criteria for benchmark data. Representativeness means that cases in a dataset cover the possible space. None of the 24 tested datasets used for training machine learning methods were well representative when analyzing distribution and coverage of cases in chromosomes, protein structures, CATH domains and classes, Pfam families, Enzyme Commission (EC) categories, and Gene Ontology annotations (Schaafsma & Vihinen, 2018). By considering representativeness it would be possible to achieve better predictor performance by covering the event space of variations better.

Relevance of a dataset means that the cases are indeed for the investigated phenomenon. The datasets used in the representativeness analysis contained <2% of disease-causing variants among the benign training sets (Schaafsma &

Vihinen, 2018). A minor allele frequency between 0.01 and 0.25 in any of the ExAC populations (Lek et al., 2016) was used to reveal benign variants.

7.2 | Variant data quality

Variation interpretation does not increase the quality of sequencing data, therefore, variant calls have to be reliable. Sequencing methods have their problems and biases. It all starts with the quality of the sample (Chen, Liu, Evans, & Ettwiller, 2017).

Recently an approach was presented for blacklisting variants common in private cohorts but not in public databases (Maffucci et al., 2019). This is an intriguing idea, however, their approach does not detect many such variants; instead, it cleans sequencing and data processing errors, mainly in low complexity regions and repeats, where disease-related variants are rather rare. It is a useful addition to filtering approaches and facilitates identification of several pathogenic variants. It may be necessary to recalibrate blacklists once changes are made to sequencing techniques or data management and analysis pipelines, as the sequencing artifacts may not remain the same.

7.3 | Minor allele frequency

Frequent variants in populations are considered to be benign. Allele frequency is the most often used criteria in variation interpretation (Amendola et al., 2016). ACMG/AMP guidelines suggest a threshold of 0.05 to distinguish frequent and thus likely not disease-related variants. This threshold is ultraconservative. In many studies, including variation tolerance predictor development, a threshold of 0.01 has been used (see Niroula & Vihinen, 2019).

Lower allele frequencies have been suggested based on numerous publications, see, for example, the study of *BRCA1* and *BRCA2* variants as well as alterations in additional 77 genes (Kobayashi et al., 2017), hearing loss-related genes (Rim et al., 2019), and cardiomyopathy variants (Kelly et al., 2018). There is even a scheme to calculate the threshold (Whiffin et al., 2017), and for taking the quality and abundance of the data and inheritance pattern into account (Nykamp et al., 2017). Common to all these papers is that the allele frequency threshold could be safely significantly lower than the recommended 0.05.

7.4 | Rare variant ≠ disease-causing variant

It has been claimed that most rare amino acid substitutions would be disease related (Kryukov, Pennacchio, & Sunyaev, 2007). Rare variants contain disease-related alterations more often than frequent alleles (Marth et al., 2011). However, both experimental studies and predictions indicate that in many genes many variants, which are mainly rare, are benign or do not have a strong phenotype (Schaafsma & Vihinen, 2017). Massively parallel reporter assays have been used to investigate almost all variants or a large portion of them in some genes and proteins and found that there is always a large portion of variants that are benign or without phenotype, for example, in PPARG (Majithia et al., 2016) and BRCA1 RING domain (Starita et al., 2015).

7.5 | Coverage of predictions

None of the existing prediction methods is capable of predicting all possible variants and even if they could the quality and performance would suffer. There are numerous reasons for the paucity of predictions, most often due to some essential parameter missing. This is typical in the case of variants in genes or proteins that are either unique for humans or shared only with a small number of other organisms. Evolutionary data, which practically all methods use, cannot be obtained in such cases.

High coverage is not indicative of high performance, either. Methods that classify variants to more than two categories can be more realistic and take the continuum of pathogenicity into account (Vihinen, 2017). In the end, only highly reliable data, including predicted outcomes, can be used for clinical assertion. This applies to any parameter used for the assessment and means that it will not be possible to predict and assert all variants.

7.6 | Terminology

ACMG/AMP guidelines have largely contributed toward terminology. Specifically, "variation" is now widely used instead of the fuzzy and often negative "mutation" that has several meanings which can confuse a meaning. The same applies to "polymorphism". However, all the guidelines are still using some problematic wording, such as missense when describing amino acid substitutions (the sense in missense refers to RNA variation). Other common problematic terms, include nonsense, frameshift, indel, and functional variation. For the problems and their remedies, see Vihinen (2015a).

8 | SUGGESTIONS

The ACMG/AMP, AMP/ASCO/CAP, and ESHG guidelines are useful and allow harmonization of variation interpretation over numerous laboratories and countries, as they are followed in many places. However, as indicated in several contributions before and in the discussion above, there are certain issues that would benefit from amendments. As prediction methods along with allele frequency information are the most widely used criteria (Amendola et al., 2016) in variation interpretation, it is important that these data are used in the most efficient and best possible way.

The major issue concerning computational predictions according to ACMG/AMP and AMP/ASCO/CAP guidelines is the requirement that multiple predictors have to agree. This means that usually the poorest performing method dictates the outcome of the assessment. Furthermore, tools based on the same or similar characteristics usually agree. Thereby it is not a surprise that evolutionary methods have high concordance, however, it does not mean that the prediction would be correct. When using many similar methods the outcome is biased.

The requirement for concordant predictions does not serve and benefit the patient, not at least in the way the guidelines are most often implemented. The situation is analogous to US National Comprehensive Cancer Network guidelines for genetic testing, which, according to a recent study, miss up to 50% of actionable variants (Beitsch et al., 2019).

My suggestion is therefore that even just one predictor, with proven good performance, could be used. Alternatively, a limited number of predictors could be used. Each of them should represent state-of-the-art performance and should be complementary, not based on the same principles and reusing the same data and predictions (Vihinen, 2014). Details for method development as well as the used data should be available. This is often a problem with commercial solutions, which typically do not reveal details and performance and essentially sell a pig in a poke (Vihinen, 2015b). Authorities should refrain from mentioning prediction and other methods as better methods will be released during years. However, if tools are mentioned, they should represent the best current performance instead of methods that may be widely used but which have poor performance.

The choice of the prediction method(s) should be based on systematic benchmarking studies reporting all the relevant performance measures. Laboratories should consult bioinformaticians when choosing prediction methods and this should happen rather frequently as new and improved methods are published every now and then. In addition to amino acid substitutions and splice site variants, there are now tools for many other types and effects of variations. The spectrum of predicted features and variation types will expand in the near future along with the availability of increasing experimental information.

ACKNOWLEDGMENTS

Financial support from Vetenskapsrådet, Swedish Cancer Society, and Alfred Österlunds Stiftelse is gratefully acknowledged.

CONFLICT OF INTEREST

None.

ORCID

Mauno Vihinen https://orcid.org/0000-0002-9614-7976

REFERENCES

Abou Tayoun, A. N., Pesaran, T., DiStefano, M. T., Oza, A., Rehm, H. L., Biesecker, L. G., & Harrison, S. M. (2018). Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Human Mutation*, *39*, 1517–1524. https://doi.org/10.1002/humu.23626

Ali, H., Urolagin, S., Gurarslan, O., & Vihinen, M. (2014). Performance of protein disorder prediction programs on amino acid substitutions. *Human Mutation*, *35*, 794–804. https://doi.org/10.1002/humu.22564

Amendola, L. M., Jarvik, G. P., Leo, M. C., McLaughlin, H. M., Akkari, Y., Amaral, M. D., … Rehm, H. L. (2016). Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the Clinical Sequencing Exploratory Research Consortium. *American Journal of Human Genetics*, *98*, 1067–1076. https://doi.org/10.1016/j.ajhg.2016.03.024

Anderson, D., & Lassmann, T. (2018). A phenotype centric benchmark of variant prioritisation tools. *NPJ Genomic Medicine*, *3*, 5.

Beitsch, P. D., Whitworth, P. W., Hughes, K., Patel, R., Rosen, B., Compagnoni, G., … Nussbaum, R. L. (2019). Underdiagnosis of hereditary breast cancer: Are genetic testing guidelines a tool or an obstacle? *Journal of Clinical Oncology*, *37*, 453–460. https://doi.org/10.1200/jco.18.01631

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., … Damborsky, J. (2014). PredictSNP: Robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Computational Biology*, *10*, e1003440. https://doi.org/10.1371/journal.pcbi.1003440

Chen, L., Liu, P., Evans, T. C. Jr, & Ettwiller, L. M. (2017). DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, *355*, 752–756.

de la Campa, E. A., Padilla, N., & 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*, 569. https://doi.org/10.1186/s12864-017-3914-0

Desmet, F., Hamroun, G., Collod-Beroud, G., Claustres, M., & Beroud, C. (2010). Bioinformatics identification of splice site signals and prediction of mutation effects. In R. M. Mohan (Ed.), *Research advances in nucleic acids research* (pp. 1–16). Kerala: Global Reseach Network.

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

Grimm, D. G., Azencott, C. A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., … Borgwardt, K. M. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation*, *37*, 1013–1024. https://doi.org/10.1002/humu.22768

Jian, X., Boerwinkle, E., & Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome.

Nucleic Acids Research, *42*, 13534–13544. https://doi.org/10.1093/nar/gku1206

Kelly, M. A., Caleshu, C., Morales, A., Buchan, J., Wolf, Z., Harrison, S. M., … Funke, B. (2018). Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: Recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel. *Genetics in Medicine*, *20*, 351–359. https://doi.org/10.1038/gim.2017.218

Khan, S., & Vihinen, M. (2010). Performance of protein stability predictors. *Human Mutation*, *31*, 675–684. https://doi.org/10.1002/humu.21242

Kobayashi, Y., Yang, S., Nykamp, K., Garcia, J., Lincoln, S. E., & Topper, S. E. (2017). Pathogenic variant burden in the ExAC database: An empirical approach to evaluating population data for clinical variant interpretation. *Genome Medicine*, *9*, 13. https://doi.org/10.1186/s13073-017-0403-7

Kryukov, G. V., Pennacchio, L. A., & Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. *American Journal of Human Genetics*, *80*, 727–739. https://doi.org/10.1086/513473

Laurila, K., & Vihinen, M. (2009). Prediction of disease-related mutations affecting protein localization. *BMC Genomics*, *10*, 122. https://doi.org/10.1186/1471-2164-10-122

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*, 285–291. https://doi.org/10.1038/nature19057

Li, J., Shi, L., Zhang, K., Zhang, Y., Hu, S., Zhao, T., … Sun, Z. (2018). VarCards: An integrated genetic and clinical database for coding variants in the human genome. *Nucleic Acids Research*, *46*, D1039–D1048. https://doi.org/10.1093/nar/gkx1039

Li, M. M., Datto, M., Duncavage, E. J., Kulkarni, S., Lindeman, N. I., Roy, S., … Nikiforova, M. N. (2017). Standards and guidelines for the interpretation and reporting of sequence variants in cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *The Journal of Molecular Diagnostics*, *19*, 4–23.

Liu, X., Jian, X., & Boerwinkle, E. (2011). dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation*, *32*, 894–899. https://doi.org/10.1002/humu.21517

Maffucci, P., Bigio, B., Rapaport, F., Cobat, A., Borghesi, A., Lopez, M., … Itan, Y. (2019). Blacklisting variants common in private cohorts but not in public databases optimizes human exome analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 950–959. https://doi.org/10.1073/pnas.1808403116

Majithia, A. R., Tsuda, B., Agostini, M., Gnanapradeepan, K., Rice, R., Peloso, G., … Altshuler, D. (2016). Prospective functional classification of all possible missense variants in PPARG. *Nature Genetics*, *48*, 1570–1575. https://doi.org/10.1038/ng.3700

Marth, G. T., Yu, F., Indap, A. R., Garimella, K., Gravel, S., Leong, W. F., … Gibbs, R. (2011). The functional spectrum of low-frequency coding variation. *Genome Biology*, *12*, R84. https://doi.org/10.1186/gb-2011-12-9-r84

Masica, D. L., & Karchin, R. (2016). Towards increasing the clinical relevance of in silico methods to predict pathogenic missense variants. *PLoS Computational Biology*, *12*, e1004725. https://doi.org/10.1371/journal.pcbi.1004725

Matthijs, G., Souche, E., Alders, M., Corveleyn, A., Eck, S., Feenstra, I., … Bauer, P. (2016). Guidelines for diagnostic next-generation sequencing. *European Journal of Human Genetics*, *24*, 2–5. https://doi.org/10.1038/ejhg.2015.226

Nair, P. S., & Vihinen, M. (2013). VariBench: A benchmark database for variations. *Human Mutation*, *34*, 42–49. https://doi.org/10.1002/humu.22204

Niroula, A., Urolagin, S., & Vihinen, M. (2015). PON-P2: Prediction method for fast and reliable identification of harmful variants. *PLoS ONE*, *10*(2), e0117380. https://doi.org/10.1371/journal.pone.0117380

Niroula, A., & Vihinen, M. (2016). Variation interpretation predictors: Principles, types, performance, and choice. *Human Mutation*, *37*, 579–597. https://doi.org/10.1002/humu.22987

Niroula, A., & Vihinen, M. (2017). Predicting severity of disease-causing variants. *Human Mutation*, *38*, 357–364. https://doi.org/10.1002/humu.23173

Niroula, A., & Vihinen, M. (2019). How good are pathogenicity predictors in detecting benign variants? *PLoS Computational Biology*, *15*, e1006481. https://doi.org/10.1371/journal.pcbi.1006481

Nykamp, K., Anderson, M., Powers, M., Garcia, J., Herrera, B., Ho, Y. Y., … Topper, S. (2017). Sherloc: A comprehensive refinement of the ACMG-AMP variant classification criteria. *Genetics in Medicine*, *19*, 1105–1117. https://doi.org/10.1038/gim.2017.37

Orioli, T., & Vihinen, M. (2019). Benchmarking membrane proteins: Subcellular localization and variant tolerance predictors. *BMC Genomics*, *20*(Suppl 8, 547).

Potapov, V., Cohen, M., & Schreiber, G. (2009). Assessing computational methods for predicting protein stability upon mutation: Good on average but not in the details. *Protein Engineering, Design and Selection*, *22*, 553–560. https://doi.org/10.1093/protein/gzp030

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., … Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, *17*, 405–423. https://doi.org/10.1038/gim.2015.30

Riera, C., Padilla, N., & de la Cruz, X. (2016). The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions. *Human Mutation*, *37*, 1012–1024. https://doi.org/10.1002/humu.23048

Rim, J. H., Lee, J. S., Jung, J., Lee, J. H., Lee, S. T., Choi, J. R., … Gee, H. Y. (2019). Systematic evaluation of gene variants linked to hearing loss based on allele frequency threshold and filtering allele frequency. *Scientific Reports*, *9*, 4583. https://doi.org/10.1038/s41598-019-41068-6

Sarkar, A., Yang, Y, & Vihinen, M. (2020). Variation benchmark datasets: update, criteria, quality and applications. *Database*. baz117.

Schaafsma, G. C., & Vihinen, M. (2015). VariSNP, A benchmark database for variations from dbSNP. *Human Mutation*, *36*, 161–166. https://doi.org/10.1002/humu.22727

Schaafsma, G. C. P., & Vihinen, M. (2017). Large differences in proportions of harmful and benign amino acid substitutions between proteins and diseases. *Human Mutation*, *38*, 839–848. https://doi.org/10.1002/humu.23236

Schaafsma, G. C., & Vihinen, M. (2018). Representativeness of variation benchmark datasets. *BMC Bioinformatics*, *19*(1), 461. https://doi.org/10.1186/s12859-018-2478-6

Starita, L. M., Young, D. L., Islam, M., Kitzman, J. O., Gullingsrud, J., Hause, R. J., … Fields, S. (2015). Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics*, *200*, 413–422. https://doi.org/10.1534/genetics.115.175802

Thusberg, J., Olatubosun, A., & Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Human Mutation*, *32*, 358–368. https://doi.org/10.1002/humu.21445

Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics*, *13*(Suppl 4), S2. https://doi.org/10.1186/1471-2164-13-s4-s2

Vihinen, M. (2013). Guidelines for reporting and using prediction tools for genetic variation analysis. *Human Mutation*, *34*, 275–282. https://doi.org/10.1002/humu.22253

Vihinen, M. (2014). Majority vote and other problems when using computational tools. *Human Mutation*, *35*, 912–914. https://doi.org/10.1002/humu.22600

Vihinen, M. (2015a). Muddled genetic terms miss and mess the message. *Trends in Genetics*, *31*, 423–425. https://doi.org/10.1016/j.tig.2015.05.008

Vihinen, M. (2015b). No more hidden solutions in bioinformatics. *Nature*, *521*, 261. https://doi.org/10.1038/521261a

Vihinen, M. (2017). How to define pathogenicity, health, and disease? *Human Mutation*, *38*, 129–136. https://doi.org/10.1002/humu.23144

Wei, Q., Wang, L., Wang, Q., Kruger, W. D., & Dunbrack, R. L. Jr (2010). Testing computational prediction of missense mutation phenotypes: Functional characterization of 204 mutations of human cystathionine beta synthase. *Proteins*, *78*, 2058–2074.

Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A. H., Karczewski, K., Ing, A. Y., … Ware, J. S. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in Medicine*, *19*, 1151–1158. https://doi.org/10.1038/gim.2017.26

Yang, Y., Niroula, A., Shen, B., & Vihinen, M. (2016). PON-Sol: Prediction of effects of amino acid substitutions on protein solubility. *Bioinformatics*, *32*, 2032–2034. https://doi.org/10.1093/bioinformatics/btw066

How to cite this article: Vihinen M. Problems in variation interpretation guidelines and in their implementation in computational tools. *Mol Genet Genomic Med*. 2020;8:e1206. https://doi.org/10.1002/mgg3.1206