Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/btaa943
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genome Analysis

Calibrating variant-scoring methods for clinical decision making

Silvia Benevenuta [1], Emidio Capriotti [2,*] and Piero Fariselli [1,*]

[1] Department of Medical Sciences, University of Torino, Via Santena, 19, 10126 Torino, Italy, [2] BioFolD Unit, Department of Pharmacy and Biotechnology (FaBiT), University of Bologna, Via Selmi 3, 40126 Bologna, Italy.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Contact: emidio.capriotti@unibo.it and piero.fariselli@unito.it
Supplementary information: Supplementary data are available at *Bioinformatics* online.

Abstract

Identifying pathogenic variants and annotating them is a major challenge in human genetics, especially for the non-coding ones. Several tools have been developed and used to predict the functional effect of genetic variants. However, the calibration assessment of the predictions has received little attention. Calibration refers to the idea that if a model predicts a group of variants to be pathogenic with a probability *P*, it is expected that the same fraction *P* of true positive is found in the observed set. For instance, a well-calibrated classifier should label the variants such that among the ones to which it gave a probability value close to 0.7, approximately 70% actually belong to the pathogenic class. Poorly calibrated algorithms can be misleading and potentially harmful for clinical decision-making.

1 Introduction

One of the main challenges in human genetics is to predict the functional effect of genetic variants (Capriotti *et al.*, 2019; Lappalainen *et al.*, 2019; Capriotti *et al.*, 2012). Knowing whether or not a variant is potentially pathogenic can lead to better diagnosis and the implementation of more effective treatment strategies which have a significant impact on clinical settings. In their day-to-day life, physicians can rely on different tools to estimate the impact of a variant, but it is not an easy task to select the most appropriate one. A common strategy to select the most reliable method consists in reading articles reporting the results of critical assessment experiments. Unfortunately, in many cases evaluation metrics do not include model calibrations (Cheng *et al.*, 2019; Drubay *et al.*, 2018).

Even when the tools have good discrimination power, they may be unreliable if they are uncalibrated (Van Calster *et al.*, 2016, Van Calster *et al.*, 2019). Calibration is a relevant measure referring to the concept that if we take a group of variants, all predicted by a "calibrated" model to be pathogenic with a probability score *P* (e.g. 0.7), it is expected that the fraction of truly pathogenic variants in that group is exactly *P* (70% of true positive is found in the observed set). A recent review found that calibration is assessed far less often than discrimination (Christodoulou *et al.*, 2019) , which is problematic since poor calibration can make predictions misleading (Van Calster and Vickers, 2015). For its high impact on the interpretability of the prediction, calibration has been addressed as the Achilles heel of predictive analytics (Shah *et al.*, 2018; Van Calster *et al.*, 2019). The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) guidelines for prediction modelling studies recommend the reporting on calibration performance (Collins *et al.*, 2015). When predictions are used in support of decision-making diagnoses and prognoses, the calibration is even more relevant as observed in the case of the cancer prediction models (Yala *et al.*, 2019).

In this letter, we evaluate the calibration of the state-of-the-art methods for scoring the impact of the variants: CADD (Kircher *et al.*, 2014), DANN (Quang *et al.*, 2015), Eigen (Ionita-Laza *et al.*, 2016), DeepSea (Zhou and Troyanskaya, 2015), FATHMM-MKL (Shihab *et al.*, 2015) and PhD-SNP^{g} (Capriotti and Fariselli, 2017). We calculated the calibration and the Brier scores (Brier, 1950) of six tools on a dataset of 2,066 single nucleotide variants, both coding and non-coding. We observed that the top classifiers, which were not necessarily well-calibrated, may lead to an incorrect interpretation of the functional effect of the genetic variant.

2 AUC performance of the different predictors

One of the most commonly used metrics for assessing the performance of the classification methods is the AUC-ROC (Area Under the Receiver Operating Characteristic Curve). The ROC curve is the plot of the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings and it illustrates the discrimination ability of a binary classifier as its discrimination threshold varies (Supplementary Materials Section 5).

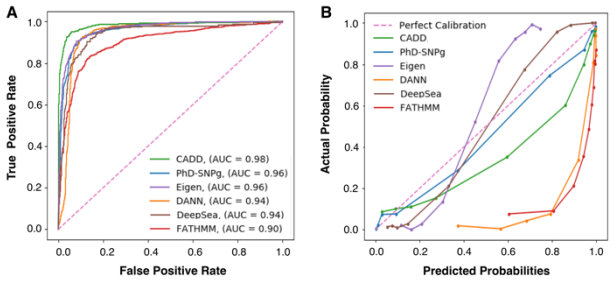

Fig. 1. (A) ROC curves of PhD-SNP^g, FATHMM-MKL, CADD, DANN and Eigen on the complete dataset (both coding and non-coding variants). DeepSea has been evaluated only on the subset of non-coding variants, since it has been developed only to score them. AUCs for coding and non-coding variants are reported in Table S2. True and false-positive rates are defined in Supplementary Materials. (B) Calibration curves of the predictors on coding and non-coding variants. CADD and Eigen scores have been modified using a sigmoid transformation ($1/ (1+exp(-A - x + B))$). The best parameters were: A=1, B=2.5 for CADD and A=1, B=1.63/0.05 for Eigen (coding and non-coding variants were transformed separately, since Eigen provides two different sets of scores).

An ideal classifier would have an AUC-ROC of 1, while a completely random classifier would have an AUC-ROC of 0.5. AUC is an efficient way to reject tools that fail to differentiate between pathogenic and benign variants.

From Fig. 1A we can see that all predictors perform quite well as discriminators on the selected dataset (All the predictions of the methods are reported in Supplementary File 1). However, none of the predictors has been validated for its calibration. Using an ill-calibrated classifier could lead to an incorrect interpretation of the functional effect of the genetic variant (it could be over-estimated or under-estimated).

3 Calibration evaluation

A standard way to examine whether or not a predictor is calibrated is to plot the calibration curve or using the Brier score (Supplementary Materials). The calibration curve shows whether the predicted probabilities agree with the observed probabilities. If the calibration curve lies on the diagonal, the predictor is perfectly calibrated, and it requires no further investigation. The deviation from the diagonal indicates the miscalibration. Brier score is a numerical value that ranges from zero to one (one being totally uncalibrated, zero being perfect calibration).

To evaluate the calibration of a method that returns a probability score we compared its outputs with the observed class frequency. For Eigen and CADD, which provided only raw scores, we transformed their outputs using an optimal sigmoid function (Fig. 1B). From Figs. 1A and 1B, we observed that despite showing similar AUCs, the tested tools have significantly different calibration curves. Indeed, PhD-SNP^g is the best-calibrated method, while DeepSea and DANN resulted in the least calibrated predictions. However, all the presented methods can be calibrated using the isotonic-regression, which transforms the output of a non-calibrated classifier in a very well-calibrated one (Niculescu-Mizil and Caruana, 2005). The effect of this kind of transformation is reported in Table 1 (and Fig.S4), where the isotonic-regression mapping is computed using a 10-fold cross-validation procedure. The cross-validation procedure is necessary to evaluate the calibration on never-seen-before data (with at least 500-1000 datapoints). The sigmoid calibration, although it requires very few data-points, was less effective and not all the methods can be calibrated (Fig. S6-S7).

Table 1. Brier scores of the methods on the dataset

Predictor	BS_{Coding}	$BS_{Non\text{-}Coding}$	BS_{All}
PhD-SNP^g	0.10 / 0.10	0.03 / 0.03	0.07 / 0.07
DANN	0.24 / 0.09	0.27 / 0.05	0.25 / 0.07
FATHMM	0.17 / 0.15	0.07 / 0.04	0.14 / 0.12
DeepSea	-	0.43 / 0.08	-
Eigen*	0.14 / 0.07	0.06 / 0.04	0.11 / 0.06
CADD*	0.06 / 0.05	0.04 / 0.03	0.05 / 0.05

Brier scores (BS) of the methods before and after isotonic calibration. * Uncalibrated scores for Eigen and CADD are obtained after sigmoid transformation

4 Conclusion

Despite showing comparable AUCs, different methods may have significantly different calibration curves. Usually, the AUC is taken as the only evaluation criterion to assess the validity of the model. Thus, a model is chosen without checking its calibration. Nonetheless, its scores might still be used and interpreted as a measure of the "*pathogenicity*" of the variants. This assumption could lead to an incorrect interpretation of the functional effects and their probability meaning.

According to our analysis, from the user standpoint, we suggest selecting a method based both on the classification and calibration performances.

In particular, for the end-users, who do not want to process the predictor outputs, we suggest to use PhD-SNP^g, as the ready-on-the-shelf method that is both accurate and naturally calibrated (Fig.1). For developers, and expert users who prefer other tools (such as CADD or FATHMM), we recommend calibrating the predictor before its usage. The calibration can be performed using suitable software such as *scikit-learn* (Pedregosa, F. *et al.*, 2011) calibration suite, which transforms the predictor outputs as shown in Supplementary Materials (Figs. S4, S14-S19).

Acknowledgements

P.F. thanks the Italian Ministry for Education, University and Research under the programme “Dipartimenti di Eccellenza 2018-2022 D15D18000410001”.

Funding

E.C. and P.F. are supported by funds of the Italian Ministry of Education, University and Research (MIUR-PRIN-201744NR8S).

Conflict of Interest: none declared.

References

Brier,G.W. (1950) Verification of Forecasts Expressed in Terms of Probability,. *Mon. Weather Rev.*, **78**, 1–3.

Capriotti,E. *et al.* (2012) Bioinformatics for personal genome interpretation. *Brief Bioinform*, **13**, 495–512.

Capriotti,E. *et al.* (2019) Integrating molecular networks with genetic variant interpretation for precision medicine. *Wiley Interdiscip Rev Syst Biol Med*, **11**, e1443.

Capriotti,E. and Fariselli,P. (2017) PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Res*, **45**, W247–W252.

Cheng,N. *et al.* (2019) Comparison and integration of computational methods for deleterious synonymous mutation prediction. *Brief Bioinform*.

Christodoulou,E. *et al.* (2019) A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*, **110**, 12–22.

Collins,G.S. *et al.* (2015) Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *Eur Urol*, **67**, 1142–1151.

Drubay,D. *et al.* (2018) A benchmark study of scoring methods for non-coding mutations. *Bioinformatics*, **34**, 1635–1641.

Ionita-Laza,I. *et al.* (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet*, **48**, 214–20.

Kircher,M. *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, **46**, 310–5.

Lappalainen,T. *et al.* (2019) Genomic Analysis in the Age of Human Genome Sequencing. *Cell*, **177**, 70–84.

Niculescu-Mizil,A. and Caruana,R. (2005) Predicting good probabilities with supervised learning., pp. 625–632.

Pedregosa, F. *et al.* (2011) Scikit-learn: Machine Learning in Python. *JMLR*, **12**, 2825–2830.

Quang,D. *et al.* (2015) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, **31**, 761–3.

Shah,N.D. *et al.* (2018) Big Data and Predictive Analytics: Recalibrating Expectations. *JAMA*, **320**, 27–28.

Shihab,H.A. *et al.* (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, **31**, 1536–43.

Van Calster,B. *et al.* (2019) Calibration: the Achilles heel of predictive analytics. *BMC Med*, **17**, 230.

Van Calster,B. and Vickers,A.J. (2015) Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Mak.*, **35**, 162–9.

Yala,A. *et al.* (2019) A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology*, **292**, 60–66.

Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods*, **12**, 931–4.