In silico predictions of missense variants is important consideration when interpreting variants of uncertain significance (VUS) in the BRCA1 and BRCA2 genes. We trained and evaluated hundreds of machine learning algorithms based on results from validated functional assays to better predict missense variants in these genes as damaging or neutral. This new optimal ”BRCA-ML” model yielded a substantially more accurate method than current algorithms for interpreting the functional impact of variants in these genes, making BRCA-ML a valuable addition to data sources for VUS classification.
the receiver operating and precision-recall curves in Fig. 1, which demonstrates better performance of BRCA-ML compared with other prediction models. In particular, the high number of false negative calls in BRCA1 many of the models yielded low area under the precision-recall curves. BRCA-ML scores for every possible missense mutation caused by a single-nucleotide variation are also given in Supplementary Data Set 1.

Figure 2 shows the gene-level scores for every possible missense variant caused by a single-nucleotide variant in BRCA1 and BRCA2 using BRCA-ML and BayesDel9, a commonly used and highly accurate predictor. While BayesDel is correctly assigning higher scores to known functional domains, the higher scores are not much more than predicted benign variants across the gene. However, in BRCA-ML, the signal to noise ratio is considerably higher between damaging and neutral variants. This evidence suggests that, unlike BayesDel, changing the threshold for damaging mutations will not significantly affect the number of predicted damaging mutations in BRCA-ML.

**DISCUSSION**

We have shown that AutoML methods are efficient means to derive optimal ML models for predicting damaging missense mutations in BRCA1 and BRCA2. The final models derived for each gene, which we collectively term BRCA-ML, show marked improvements in MCC and other metrics with respect to individual missense prediction algorithms.

Even in the age of large-scale mutational scanning techniques like those from Findlay2 and Starita2,3, in silico mutation analysis will likely continue to be relevant. While the number of variants functionally tested is impressive for both studies (1056 and 3893, respectively), there are over 12,520 and 22,772 possible single-nucleotide variants in BRCA1 and BRCA2. Therefore, it could be several years before the technology exists to scale to all possible variants, hence a short term need for computational predictions.

It should be noted that there remain several limitations for these models. First, there are a limited number of known damaging mutations in BRCA1 and BRCA2 from which to build a model. The lack of damaging mutations limits the model ability to capture the complete variability of input data. Second, the training data are limited to characterized mutations in regions of the proteins known to be associated with impaired DNA damage repair. For example, the only missense variants in BRCA2 that are associated with disease are in the DNA-binding domain. It is not known if variants in other domains that we or others predict to cause damaging missense mutations are able to inhibit DNA repair. However, by keeping the test set isolated from the training data, this influence should be minimal. More known mutations in these genes will be necessary to quantify the amount of overfitting.

The data presented in this paper show that highly accurate prediction of missense variants in BRCA1 and BRCA2 are not only possible but simple to access (see Supplementary Data Set 1 for all possible SNVs in both genes). This improved performance in BRCA-ML should provide higher quality evidence to genetic counselors and researchers for interpreting deleteriousness of missense variants.

**METHODS**

AutoML

We employed the AutoML approach with the R (version 3.4.2) package h2o.ai (version: 3.16.0.2) to identify the optimal model for predicting the functional effect of missense variants in BRCA1 and BRCA2. Variants were loaded in the following order: Hart2, Starita2, Fernandes4, and Findlay; keeping only variants not observed in the previous studies. We also included new BRCA2 functional data for 15 neutral (V2527A, G2544S,
Locations and changes are relative to the hg19/GRCh37 human genome build. Since not all variants could unequivocally be assigned to a given class, we selected variants for inclusion if they satisfied the following criteria: “FUNC” (neutral) or “LOF” (damaging)\(^3\), HDR score \(\leq 0.33\) (damaging) or \(\geq 0.77\)\(^3\), or International Agency for Research on Cancer classes 0,1 (neutral) and 4,5 (damaging)\(^4\). Variants were excluded if they were not observed in known functional domains in \(\text{BRCA1}\) (BRCT: amino acids 1–109, RING: amino acids 1642–1855) or \(\text{BRCA2}\) (DNA Binding: amino acids 2479–3192). This left 1902 variants (\(n = 259\) damaging) for \(\text{BRCA1}\) and 202 variants (\(n = 74\) damaging) for \(\text{BRCA2}\).

For training each gene, 80% of variants were selected and trained to maximize the per class accuracy, with robustness assessed using fivefold cross-validation. Input features were missense prediction models from dbNSFP (version 3.4)\(^7\), including SiftScore, Polyphen2HdivScore, Polyphen2HvarScore, LrtScore, MutationtasterScore, FathmmScore, ProveanScore, Vest3Score, MetasvmScore, MetalrScore, MCapScore, RevelScore, MutpredScore, CaddRaw, DanniScore, FathmmMcKodialCodingScore, GenocanyonScore, IntegratedFitconsScore, Gm12878FitconsScore, H1HescFitconsScore, HuvecFitconsScore, BayesDel, AlignGVGD\(^\text{P}\), AlignGVGD\(^\text{P}\), and BayesDel\(^\text{P}\) were also added using the BioR framework\(^10\). Optimal cutpoints for each of the individual input features (\(n = 25\)) from dbNSFP, AlignGVGD, and BayesDel were determined using the same training data as used in AutoML so as to make a fair comparison.

**Evaluation**

For the test set evaluation, statistical measures of sensitivity, specificity were computed with the caret package\(^11\). The MCC is used throughout as an optimal metric for gauging the performance of a binary classifier, as it represents a singular value that takes into consideration the proportion of each class. The values of MCC range from \(-1\) to \(1\), where \(-1\) represents the worst possible agreement and 1 representing perfect agreement. We also present traditional measures of performance for ML models such as receiver operating curves and precision-recall curves. All chromosomal locations and changes are relative to the hg19/GRCh37 human genome build.

**REFERENCES**

1. Olson, R. S., Cava, W. L., Mustahsan, Z., Varik, A. & Moore, J. H. Data-driven advice for applying machine learning to bioinformatics problems. In Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 192–203 (World Scientific, 2018).
2. Starita, L. M. et al. A multiplex homology-directed DNA repair assay reveals the impact of more than 1,000 BRCA1 missense substitution variants on protein function. Am. J. Hum. Genet. 103, 498–508 (2018).
3. Findlay, G. M. et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature 562, 217–222 (2018).
4. Fernandes, V. C. et al. Impact of amino acid substitutions at secondary structures in the BRCT domains of the tumor suppressor BRCA1: implications for clinical annotation. J. Biol. Chem. https://doi.org/10.1074/jbc.RA118.005274 (2019).
5. Hart, S. N. et al. Comprehensive annotation of BRCA1 and BRCA2 missense variants by functionally validated sequence-based computational prediction models. Genet. Med. https://doi.org/10.1016/s1476-4438(17)30169-x (2018).
6. CRAN—Package h2o. https://cran.r-project.org/web/packages/h2o/index.html. (Accessed 29 Jan 2018).
7. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum. Mutat. 37, 235–241 (2016).
ACKNOWLEDGEMENTS

This work was supported by the Breast Cancer Research Foundation; National Institutes of Health grants [CA192393, CA176785, CA116167]; and a National Cancer Institute Specialized Program of Research Excellence (SPORE) in Breast Cancer to Mayo Clinic (PS0 CA116201).

AUTHOR CONTRIBUTIONS

H.S. generated the functional data. S.N.H. designed the analysis, analyzed the data, and drafted the paper. F.J.C., S.Y., and E.C.P. contributed a number of key suggestions for the present version. All authors edited and approved the paper.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41523-020-0159-x.

Correspondence and requests for materials should be addressed to S.N.H.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.