genetics of kidney disease

Phenotype first: a data-driven approach to genetic penetrance

Jan Halbritter[1] and Eric Olinger[2]

Refers to: Forrest IS, Vy HMT, Rocheleau G, et al. Machine learning-based penetrance of genetic variants. *Science*. 2025;389:eadm7066.

Kidney International (2026) ■, ■–■; https://doi.org/10.1016/j.kint.2026.01.010

KEYWORDS: genotype; penetrance; phenotype; variants; VUS

Massively parallel sequencing has long entered clinical care for pediatric and adult nephrology patients, where a genetic etiology is considered after appropriate clinical evaluation.[1] However, technological advances in genetic testing have outpaced the capacity to interpret the wealth of genetic data generated. Currently, the pathogenicity of a given variant is assigned to 1 of 5 categories (from "benign" to "uncertain significance" to "pathogenic") using a framework established by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology, integrating population, computational, functional, and segregation data.[2] Despite advances in *in silico* meta-predictors (e.g., REVEL score[3]) and protein structure predictions (e.g., AlphaMissense[4]), the majority of reported variants remain of "uncertain significance" (47.8%, ClinVar Miner accessed January 2024) and are therefore not used for clinical diagnosis, management, prognostication, or genetic counseling, highlighting the limited clinical utility of many genetic results. Furthermore, penetrance, defined as the risk that carriers of a given genetic variant will develop a clinical phenotype, as well as phenotypic severity (expressivity), varies along a continuum rather than existing as categorical or dichotomous (case/control) traits. Currently, variant pathogenicity is treated as a fixed attribute without accounting for environmental, demographic, or polygenic context, all of which can influence disease expression.[5] Estimating disease penetrance is particularly critical in dominant disorders and in females with X-linked diseases. Relevant nephrology examples include *APOL1*-associated kidney disease, monoallelic *COL4A3/A4/A5* variants, and genes associated with cystic kidney disease.[6] Despite the importance of penetrance data in allowing for more widespread integration of genomic testing into clinical practice and precision medicine, information on penetrance is unavailable for most individual variants, especially rare ones. Recognizing this challenge, the ClinGen Low Penetrance/Risk Allele Working Group identified low penetrance variants as distinct from highly penetrant disease-causing variants, the classification frameworks of the latter being inadequate to capture the penetrance spectrum.[7] Large-scale population studies with detailed phenotyping are needed to improve the evidence sustaining the penetrance associated with any given genetic variant.

What did the study show?

Responding to this call, the study by Forrest *et al.*[8] used patient electronic health records to create machine-learning (ML) models for 10 selected autosomal dominant disorders with incomplete or age-dependent penetrance, including autosomal dominant polycystic kidney disease (ADPKD). Gradient-boosted, tree-based ML models were derived and internally validated in the Mount Sinai Data Warehouse cohort of approximately 1.3 million individuals, including 1792 polycystic kidney disease (PKD) cases. PKD cases were defined by ultrasound criteria or International Classification of Diseases, Tenth Revision (ICD-10) codes, including cystic kidney disease (Q61) and chronic kidney disease (CKD) (N18). The

[1]*Department of Nephrology and Medical Intensive Care, Charité-Universitätsmedizin Berlin, Berlin, Germany; and* [2]*Center for Human Genetics, Cliniques Universitaires Saint-Luc, Brussels, Belgium*

Correspondence: *Eric Olinger, Center for Human Genetics, Cliniques Universitaires Saint-Luc, Avenue Hippocrate, 10, Brussels 1200, Belgium. E-mail: eric.olinger@saintluc.uclouvain.be*

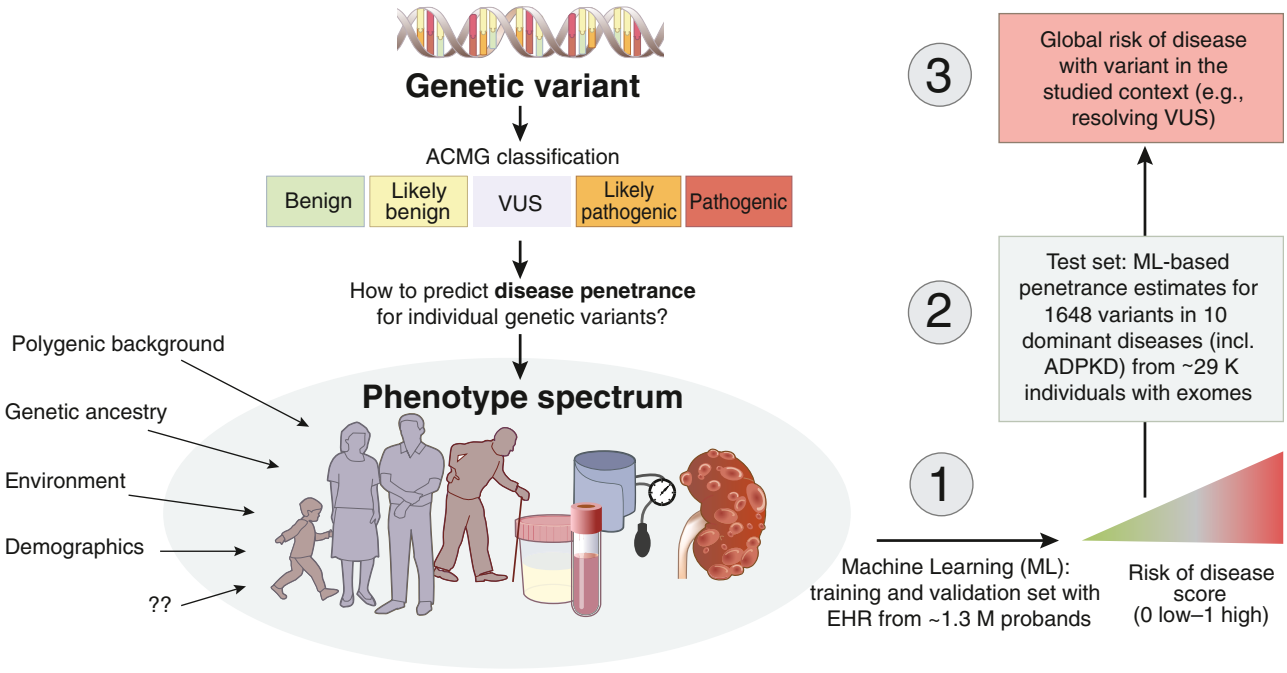

Figure 1 | Machine learning (ML)–based estimation of penetrance for genetic variants. ML models trained and validated using electronic health records (EHR) in >1,300,000 individuals from the Mount Sinai Data Warehouse and the Bio*Me* ML cohort (approximately 22,000 individuals with broad racial, ethnic, and socioeconomic diversity) to generate continuous disease scores for polycystic kidney disease (PKD) with positive and negative predictive values (PPV and NPV) of approximately 0.7. These ML disease scores are then integrated with exome data in approximately 29,000 individuals to generate ML-based penetrance estimates for 1648 variants in 10 dominant diseases (including variants in *PKD1* and *PKD2*). ML penetrance correlated with ClinVar and AlphaMissense predictions and was associated with clinical outcomes. Compared with traditional case versus control assignments, ML-based penetrance yielded more precise quantitative estimates and facilitated the interpretation of variants of uncertain significance (VUS) by mapping clinical trajectories over time. ADPKD, autosomal dominant polycystic kidney disease.

models analyzed continuous clinical data (routine laboratory and vital sign measurements) and demographics to estimate individual disease probability (ML-based disease scores). Notably, the PKD model's most important features were serum or plasma creatinine, urea nitrogen, and estimated glomerular filtration rate. (It should be noted that the model did not include the use of imaging data.) Models were tested in the independent Bio*Me* ML cohort and finally applied to the Bio*Me* exome cohort, intersecting ML-derived disease probabilities with exome variants to generate ML-based penetrance estimates for each genetic variant adjusted for age, sex, body mass index, and genetic ancestry (Figure 1). ML penetrance was computed for 1648 rare variants across 31 autosomal dominant disease-predisposition genes (including *PKD1* and *PKD2* for ADPKD), spanning different classes of variants and previously uncharacterized loss-of-function variants. When compared across ClinVar classes, ML-based penetrance behaved as expected: it was highest for pathogenic and loss-of-function variants, intermediate for variants of uncertain significance, and lowest for benign variants. Penetrance estimates decreased as variant population frequency increased and were supported by functional data for several genes (*BRCA1*, *KCNQ1*, and *LDLR*). AlphaMissense scores (building on the protein structure prediction tool AlphaFold) showed only a weak correlation with ML penetrance estimates, highlighting that computational predictions alone cannot reliably predict clinical phenotype. ML penetrance was further associated with disease-relevant clinical outcomes in heterozygous individuals. For ADPKD, the top significantly associated outcomes were cystic kidney disease, CKD, kidney failure, and secondary hypertension, with odds ratios per 0.1 increase in ML penetrance ranging between 1.05 (for hypertension) and 1.11 (for CKD). Eight carriers of highly ML penetrant *PKD1*/*PKD2* loss of function variants (splice-site, nonsense, or frameshift variants) showed markedly reduced estimated glomerular filtration rate (lowered by 40 ml/min) and reduced hemoglobin over time and were diagnosed

with PKD, CKD, and/or kidney failure, validating ML-based penetrance estimates for these canonical variant carriers. Interestingly, ML-based penetrance also helped to stratify clinical significance for variants of uncertain significance. For instance, 2 carriers of a variant of uncertain significance in *PKD2* that was estimated as highly ML-penetrant (penetrance = 0.89; 95% confidence interval, 0.72–0.97, *PKD2* NM_000297.4: c.1927G>A, p.(Asp643Asn)) had decreased estimated glomerular filtration rate and hemoglobin over time. Both carriers had multiple diagnoses of renal calculi, acute renal failure, and CKD. By contrast, 2 carriers of a variant of uncertain significance in *PKD2* that was weakly ML-penetrant (penetrance = 0.17; 95% confidence interval, 0.10–0.17, *PKD2* NM_000297.4: c.154C>A, p.(Leu52Met)) had no changes in these parameters and no kidney disease diagnoses.

Why is the study important?

Ultimately, both clinicians and patients seek to understand the risk that carriers of a given genetic variant will develop a clinical phenotype (penetrance). Traditionally, this has been addressed from a molecular perspective, using metrics such as evolutionary conservation, predicted effects on protein structure and function, and variant prevalence. This framework implicitly assumes that rare variants with a high predicted molecular impact are also highly penetrant. However, analyses of large-scale biobank data demonstrate that this view is overly simplistic.[9] Assigning fixed pathogenicity categories to variants and reducing individuals to dichotomous case-control labels fails to capture the inherently continuous and multifactorial biological reality. Phenotypic expression is shaped not only by genetic variation, but also by demographics, environmental exposures, and oligogenic or polygenic background.

Forrest *et al.* address this challenge from the opposite direction by looking at ranges of phenotype, that is, the quantifiable "end product" of the integration of all the nuances of known and unknown genetic, demographic, and environmental factors. They profile phenotype spectra using large electronic health record datasets and ML-derived disease scores that span continuously from 0 to 1, more faithfully reflecting clinical reality. The authors then assign ML-based penetrance scores to individual variants, capturing disease risk within specific clinical contexts while adjusting for relevant covariates (Figure 1). This strategy overcomes key limitations of traditional penetrance estimates and positions ML-derived penetrance as a possible tool for dissecting complex variant-phenotype relationships. Particularly for ADPKD, however, the absence of imaging, family history, and segregation data in the models limits accurate assessment of variant penetrance and classification. Although genotype and phenotype—most notably kidney volume—are prognostic markers for ADPKD with clear-cut pathogenic variants, ML-based disease and penetrance models are likely to be most valuable for nontruncating uncharacterized *PKD1/PKD2* variants. These approaches may also prove useful for other conditions, such as monoallelic *COL4A3/A4/A5* variants, for which genotype-phenotype correlations and predictive frameworks remain limited. An inherent constraint of ML-based penetrance, as presented here, is its limited transposability across populations (as many notably non-European ethnicities are still underrepresented in population databases). The relatively weak correlation observed for variants shared between the UK Biobank and Bio*Me*—restricted to a subset of disorders—highlights the context-dependent nature of penetrance and constrains variant score generalizability. Nonetheless, this study provides a valuable framework for refining genetic risk assessment, guiding additional screening, and informing clinical decision-making in precision medicine.

DISCLOSURE

All the authors declared no competing interests.

REFERENCES

1. Franceschini N, Feldman DL, Berg JS, et al. Advancing genetic testing in kidney diseases: report from a National Kidney Foundation Working Group. *Am J Kidney Dis*. 2024;84:751–766.
2. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17:405–424.
3. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet*. 2016;99:877–885.
4. Cheng J, Novati G, Pan J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*. 2023;381:eadg7492.

5. Khan A, Shang N, Nestor JG, et al. Polygenic risk alters the penetrance of monogenic kidney disease. *Nat Commun*. 2023;14:8318.
6. Sadeghi-Alavijeh O, Chan MM, Doctor GT, et al. Quantifying variant contributions in cystic kidney disease using national-scale whole-genome sequencing. *J Clin Invest*. 2024;134:e181467.
7. Schmidt RJ, Steeves M, Bayrak-Toydemir P, et al. Recommendations for risk allele evidence curation, classification, and reporting from the ClinGen Low Penetrance/Risk Allele Working Group. *Genet Med*. 2024;26:101036.
8. Forrest IS, Vy HMT, Rocheleau G, et al. Machine learning-based penetrance of genetic variants. *Science*. 2025;389:eadm7066.
9. Forrest IS, Chaudhary K, Vy HMT, et al. Population-based penetrance of deleterious clinical variants. *JAMA*. 2022;327:350–359.