The increasing availability of DNA sequence data and access to sophisticated bioinformatic algorithms mean that an unbiased bioinformatics-based assessment of the predicted impact of a genomic variant is rapidly available. The key point of this Viewpoint article is that such bioinformatic assessments are not equivalent to an expert diagnostic interpretation and may be misleading in both research and clinical care. Prediction algorithms in genomic medicine Recently published examples involving monogenic diabetes demonstrate how pathogenicity prediction algorithms can be very inaccurate for predicting which genetic variants are likely causal of dominant monogenic disease (1–4). Here, we highlight the potential pitfalls of variant classification and how they can be avoided. A recent study used a bioinformatic algorithm to identify 88 “likely pathogenic” monogenic diabetes variants in 80 individuals (8.6%) from a cohort of 1019 individuals with type 1 diabetes for 50 or more years (4). Application of the widely used American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) standards and guidelines (5) classifies only nine of these 88 variants as likely pathogenic or pathogenic variants that would be reported by our clinical diagnostic laboratory as likely causative of the patients’ diabetes. This is not an isolated occurrence; other published research studies with an overreliance on in silico prediction tools have reported high levels (~90%) of false […]
The increasing availability of DNA sequence data and access to sophisticated bioinformatic algorithms mean that an unbiased bioinformatics-based assessment of the predicted impact of a genomic variant is rapidly available. The key point of this Viewpoint article is that such bioinformatic assessments are not equivalent to an expert diagnostic interpretation and may be misleading in both research and clinical care.

**Prediction algorithms in genomic medicine**

Recently published examples involving monogenic diabetes demonstrate how pathogenicity prediction algorithms can be very inaccurate for predicting which genetic variants are likely causal of dominant monogenic disease (1–4). Here, we highlight the potential pitfalls of variant classification and how they can be avoided.

A recent study used a bioinformatic algorithm to identify 88 “likely pathogenic” monogenic diabetes variants in 80 individuals (8.6%) from a cohort of 1019 individuals with type 1 diabetes for 50 or more years (4). Application of the widely used American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) standards and guidelines (5) classifies only nine of these 88 variants as likely pathogenic or pathogenic variants. Each of these is discussed in turn below and illustrated in Figure 1.

**How should expert disease-related knowledge guide the interpretation and reporting of genetic variants that might cause autosomal dominant monogenic disease?**

Do not analyze genes without robust evidence to support the gene-disease association (6), for example the BLK, KLF11, and PAX4 genes, where current evidence is limited (7). Reputable clinical diagnostic laboratories do not include these genes in their monogenic diabetes testing.

Do not report a heterozygous variant in autosomal recessive disorders caused by biallelic variants (homozygous or compound heterozygous). For example, biallelic pathogenic WFS1 variants cause Wolfram syndrome (8), a very rare disorder with an estimated prevalence of 1 in 500,000 (heterozygous carrier frequency ~0.3%). WFS1 is an extremely polymorphic gene with rare (allele frequency < 0.1%) missense variants present in more than 2% of the population. Although there are reports of autosomal dominant diabetes caused by heterozygous WFS1 variants, these are extremely rare: one family with dominantly inherited nonsyndromic diabetes (9) and five patients with neonatal- or infancy-onset diabetes, deafness, and cataracts (10). The finding of a rare heterozygous variant is therefore highly unlikely to be causative of monogenic diabetes and should not be reported.

Use appropriate gene-specific allele frequency cutoffs in control population data to exclude variants that are too common to be highly penetrant disease-causing variants. For example, Ming-Qiang et al. found that PAX4 variants were the second most common cause of monogenic diabetes in their Chinese cohort (3), but the missense variants they reported are present in more than 1% of the East Asian population cohort (approximately 15,000 individuals) in the publicly available gnomAD database (11). A tool is available (http://cardiodb.org/alleleFrequencyApp) that allows the user to input inheritance mode, disease prevalence, penetrance, genetic heterogeneity (how many cases can be attributed to the gene), and allelic heterogeneity (how many cases can be attributed to a single variant) and calculate a maximum credible allele frequency (12). Using HNFIA monogenic diabetes as an example with monoallelic inheritance, disease prevalence of 1 in 10,000, allelic heterogeneity 0.16, genetic heterogeneity 0.35, and penetrance 0.95, the maximum tolerated pathogenic allele count is 2 in the gnomAD database (n = 141,456). In a study of diabetic subjects from South India, Mohan et al. reported HNFIA variants as the most common subtype of monogenic diabetes in their study (1). However, six of the 11 patients had variants that are too common in the European ancestry population in gnomAD.
insulinism but does not cause monogenic diabetes. Other examples include the CEL gene, where only variants within the first or fourth repeats of the VNTR region are pathogenic (15); RFX6, where there is only evidence to implicate protein truncating variants (16); and the sole heterozygous PDX1 pathogenic variant, p.P63fs, known to cause monogenic diabetes through a dominant negative effect (17).

Exclude variants of uncertain significance. These are variants lacking evidence for classification as pathogenic or likely pathogenic (5). Examples include novel missense variants in constrained genes where the amino acid substitution is predicted to have a deleterious effect upon protein function. Any individual has about 100 variants of this type and should be treated as “uncertain until proven guilty” (18).

Why are errors in the interpretation of genetic variants common in both academic and diagnostic reports?

Misinterpretation of genetic variants is common, both in the published literature (19) and in reports from diagnostic laboratories. Next-generation sequencing technology has facilitated both the ease and scale of genetic testing, but obtaining genotype data is far more straightforward than interpreting it correctly.

Academic studies often use bespoke criteria for defining pathogenicity rather than applying guidelines developed to improve the quality and consistency of variant interpretation (5). Studies may base their variant classifications solely on in silico prediction of pathogenicity using tools such as REVEL or SIFT, PolyPhen, and MutationTaster that provide only supporting evidence within the ACMG/AMP guidelines framework recommended for diagnostic reporting (5).

The availability of large variant data sets such as the gnomAD (11) has shown that many previously reported pathogenic variants are too common to be highly penetrant variants causative of monogenic diabetes. Other examples include the CEL gene, where only variants within the first or fourth repeats of the VNTR region are pathogenic (15); RFX6, where there is only evidence to implicate protein truncating variants (16); and the sole heterozygous PDX1 pathogenic variant, p.P63fs, known to cause monogenic diabetes through a dominant negative effect (17).

Exclude variants of uncertain significance. These are variants lacking evidence for classification as pathogenic or likely pathogenic (5). Examples include novel missense variants in constrained genes where the amino acid substitution is predicted to have a deleterious effect upon protein function. Any individual has about 100 variants of this type and should be treated as “uncertain until proven guilty” (18).

Why are errors in the interpretation of genetic variants common in both academic and diagnostic reports?

Misinterpretation of genetic variants is common, both in the published literature (19) and in reports from diagnostic laboratories. Next-generation sequencing technology has facilitated both the ease and scale of genetic testing, but obtaining genotype data is far more straightforward than interpreting it correctly.

Academic studies often use bespoke criteria for defining pathogenicity rather than applying guidelines developed to improve the quality and consistency of variant interpretation (5). Studies may base their variant classifications solely on in silico prediction of pathogenicity using tools such as REVEL or SIFT, PolyPhen, and MutationTaster that provide only supporting evidence within the ACMG/AMP guidelines framework recommended for diagnostic reporting (5).

The availability of large variant data sets such as the gnomAD (11) has shown that many previously reported pathogenic variants are too common to be highly penetrant disease-causing variants (20). In some cases the evidence supporting gene-disease associations is no longer valid, but publications refuting these genes are rare and may consider only a single putative mutation (21, 22). The utility of population variant databases will increase as more exome and
How can the accuracy of variant interpretation and reporting be improved?

A number of initiatives are addressing various aspects of variant interpretation. For example, the NIH-funded ClinGen resource (https://clinicalgenome.org/) includes curating of gene-disease validity evidence, theClinVar variant repository, and expert groups developing gene- or disease-specific criteria for variant classification. Sharing genetic variant data on a global scale is an essential requirement (25).

The ACMG/AMP variant classification guidelines have been adopted in many countries, and we recommend that all academic studies use these guidelines for variant classification (5). Genetic testing for clinical diagnosis should be performed in an accredited laboratory that participates in external quality assessment schemes that include variant classification.

Conclusion

We have entered a new era in which the generation of massive quantities of accurate genetic data from an individual is no longer difficult, but the new challenge is how to correctly interpret this data. This Viewpoint emphasizes how disease-specific expertise is required when interpreting genetic data and how failure to use this information will result in errors. Misdiagnosis not only affects the individual patients for whom testing is being performed but also can be amplified through predictive testing of relatives and use of incorrect variant classifications in databases and publications for interpretation of the same variant in other patients. Accurate genetic diagnosis is needed to predict disease prognosis and guide clinical management.

Acknowledgments

SE and ATH are Wellcome Senior Investigators (098395/Z/12/Z). ATH is a Senior Investigator at the National Institute for Health Research and is supported by the National Institute for Health Research Exeter Clinical Research Facility. KAP has a postdoctoral fellowship funded by the Wellcome Trust (110082/Z/15/Z).

Address correspondence to: Sian Ellard, Exeter Genomics Laboratory, RILD Building Level 3, Royal Devon & Exeter Hospital, Barrack Road, Exeter, EX2 5DU, United Kingdom. Phone: 44.1392.408259; Email: sian.ellard@nhs.net.