Guidance for estimating penetrance of monogenic disease-causing variants in population cohorts

Caroline F Wright[1]*, Luke N Sharp[1]*, Leigh Jackson[1], Anna Murray[1], James S Ware[2,3,4], Daniel G MacArthur[5,6], Heidi L Rehm[4,7], Kashyap A Patel[1]* and Michael N Weedon[1]*

1. Department of Clinical and Biomedical Sciences, Medical School, University of Exeter, Exeter, UK
2. National Heart & Lung Institute & MRC Laboratory of Medical Sciences, Imperial College London, London W12 0HS, UK
3. Hammersmith Hospital, Imperial College Healthcare NHS Trust, London W12 0HS
4. Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
5. Centre for Population Genomics, Garvan Institute of Medical Research and UNSW Sydney, Sydney, NSW, Australia
6. Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, VIC, Australia
7. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

** These authors contributed equally*

Correspondence

caroline.wright@exeter.ac.uk, mnweedon@exeter.ac.uk

ABSTRACT

Penetrance is the probability that an individual with a pathogenic genetic variant develops a specific disease. Knowing the penetrance of variants for monogenic disorders is important for counselling of individuals. Until recently, estimates of penetrance have largely relied on affected individuals and their at-risk family members being clinically referred for genetic testing – a "phenotype-first" approach. This approach substantially overestimates the penetrance of variants because of ascertainment bias. The recent availability of whole genome sequencing data in individuals from very large-scale population-based cohorts now allows "genotype-first" estimates of penetrance for many conditions. Although this type of population-based study can underestimate penetrance due to recruitment biases, it provides more accurate estimates of penetrance for secondary or incidental findings. Here, we provide guidance for the conduct of penetrance studies to ensure that robust genotypes and phenotypes are used to accurately estimate penetrance of variants and groups of similarly annotated variants from population-based studies.

Main

Understanding the cause of disease enables accurate diagnosis and targeted treatment and is the cornerstone of good medical practice. Genome-wide sequencing has hugely increased the number of genetic diagnoses in rare diseases, and there are now thousands of monogenic diseases linked to variants in >4000 individual genes[1]. Within this group of rare monogenic conditions, there is a large subset where just a single rare genotype in a specific gene causes disease. However, despite this apparent simplicity, the reality is often more complex, due to genetic and allelic heterogeneity, incomplete penetrance and variable expressivity, pleiotropy, multi-system phenotypes, and gene-environment interactions. In some cases, a specific genetic variant is both necessary and sufficient to cause the disease, and individuals who carry the variant always have or will develop the disease (complete penetrance). However, in other cases, while a single genotype may be the major cause of disease, other genetic or environmental factors are also required for the disease to manifest, and some individuals who carry the pathogenic genotype may never develop the disease (incomplete penetrance) or may present with a less severe form (variable expressivity)[2]. Penetrance is typically quantified for a particular disease or phenotype with respect to a particular age[3].

Historically, studies of penetrance have relied on large family pedigrees, where the disease (mostly) segregates with the causal genetic variant, and small clinical cohorts where every individual is affected by the disease in question. This “phenotype-first” approach to ascertainment suffers from bias towards individuals and families that have at least one and often multiple affected individuals, often with shared genetic and environmental backgrounds, resulting in inflated estimates of penetrance for many genetic variants[4]. More recently, thanks to next-generation sequencing technologies, it has become possible to study penetrance in

large population cohorts where individuals are not ascertained based on disease state. In comparison to the traditional clinical approach, this "genotype-first" approach to ascertainment may suffer from an opposing bias towards healthy individuals[5], resulting in deflated estimates of penetrance due to depletion of severe or early-onset phenotypes and enrichment of mild or late-onset phenotypes[6–9]. In principle, the two approaches should complement each other, enabling quantitative estimates of the upper and lower bounds of penetrance, and facilitating studies into the molecular mechanisms underlying monogenic disease and incomplete penetrance. However, in practice, both types of studies may be poorly executed, resulting in erroneous gene-disease associations, spurious assertions of variant pathogenicity, and inaccurate estimates of penetrance. Once made, errors in the literature can permeate through citations and genetic databases to diagnostic laboratories, clinicians, and ultimately patients, who may be wrongly diagnosed or inappropriately counselled or treated.

As genomic medicine becomes embedded into healthcare, it is ever more important to be judicious when evaluating disease penetrance. Accurate estimates of penetrance in the population are particularly crucial for reporting of secondary or incidental findings, where there may be no clinical presentation or family history of disease[10]. Since most pathogenic disease-causing variants are extremely rare, population-based estimates of penetrance often group similarly annotated variants together, which can make comparison with penetrance of individual familial variants more challenging. Moreover, consideration of modifying factors beyond the gene of interest is important for understanding penetrance. Clinical cohorts are generally enriched for etiologic co-factors that are either more heterogeneous or depleted in large population cohorts[11], including both genetic factors (e.g. epistasis,

digenic/oligogenic/polygenic contribution, genetic ancestry, chromosomal sex) and environmental exposures.

Here, we highlight a number of key problems that have resulted in substantial errors in published penetrance estimates, and use an exemplar trait to evaluate the effect of taking a more stringent approach to curation.

Avoidable errors that can substantially influence penetrance estimates

Penetrance is generally understood to apply in the context of monogenic disease-causing variants. It is defined simply as the probability of developing a particular disease given the presence of a particular pathogenic genotype. With the availability of increasingly large population-based studies, including individual electronic health records linked to exome or genome sequencing data, assessment of penetrance is becoming increasingly common. These studies are often gene-specific[12–14] or disease-specific[15–17], but there have also been efforts to perform more comprehensive surveys of penetrance[18]. For example, a recent study estimated the median penetrance across 157 autosomal dominant monogenic diseases in two large biobanks to be 0%[19], which is highly surprising.

Whilst large-scale approaches have the advantage of maximising numbers of individuals in whom penetrance can be assessed, it is easy to include more variants than the evidence supports and to group together variants with different effects. Such a highly permissive approach can overlook important aspects of biology, particularly where numerous different variants/genes/diseases are being evaluated simultaneously, based on database annotations

or bioinformatic predictions. This aggregate approach can thus result in the erroneous inclusion of variants that are benign, caused by a number of avoidable errors outlined below.

1. Inclusion of variants with inappropriately high allele frequency

Pathogenic variants for rare diseases are rare, and it has been shown that excluding common alleles is justified for most rare disease phenotypes (>5% is considered benign in standard variant classification guidelines)[20], with a few notable exceptions[21], such as founder variants in particular populations and hypomorphic variants in specific autosomal recessive diseases[22,23]. Even modestly common variants with appreciable penetrance will be easily observed to have higher frequency in cases than controls[24], and widely used variant classification guidelines state that "in general, an allele frequency in a control population that is greater than expected for the disorder is considered strong support for a benign interpretation for a rare Mendelian disorder"[20]. However, older studies that described novel pathogenic variants in genes previously unlinked with monogenic diseases generally only tested a few hundred controls. Now, with publicly available cohorts of 100,000's of sequenced individuals, many of these previously "pathogenic" variants have turned out to be relatively common variants, either in the same or another ancestral group[14,25–27], and are actually benign or low-risk variants[28,29]. Nonetheless, these benign variants and common risk alleles have polluted the monogenic disease literature and can be difficult to redress. Including common-risk alleles in penetrance estimates leads to artificially low estimates of the true penetrance of pathogenic variants for a monogenic condition, especially as these variants will contribute more to penetrance estimates as common variants will be given most weight in penetrance calculations.

2. Errors in genotype calling

Whilst great technological advancements have been made in genomics, false-positive variant calls still occur at a relatively high rate and some variant types remain particularly difficult to detect. This is a particular problem for penetrance studies of rare disorders, because in general the more deleterious a variant, the more likely it is to be a false-positive caused by a technical artefact[30]. Including these technical false-positive pathogenic variants can substantially reduce estimates of penetrance. For example, very rare variants are poorly assayed using most genotyping arrays[31,32]. Short-read next-generation sequencing technologies are more reliable for genotyping single-nucleotide variants, but they may be unreliable for genotyping complex variants, such as insertions and deletions. For example, a 32bp complex insertion-deletion in *UMOD* was initially called as six different variants from exome sequencing data in UK Biobank[33]; one of these occurred in 11 individuals and was incorrectly annotated as a nonsense variant and found to strongly associate with renal failure through haploinsufficiency[34]. In reality, the variant results in an in-frame change that causes monogenic renal failure through a well-characterised gain-of-function mechanism[33]. As a result, care must be exercised when evaluating the effect of variants without further validation with an orthogonal technique. Although additional laboratory confirmation is often not practicable, applying stringent quality control metrics to the selection of variants, as well as algorithms to address multi-nucleotide variants (MNVs) and other complex variants, coupled with visualisation of the underlying data using tools such as the Integrative Genomics Viewer (IGV)[35], can be helpful for excluding false- positive calls that will otherwise result in a lower estimate of penetrance. It is also essential to distinguish somatic mosaic variation (with low allele balance) from germline variation to enable the penetrance of germline variants to be evaluated[36]. The role of somatic mutations in normal ageing as well as disease is increasingly

being realised[37,38] and may confound genotype-phenotype evaluations, particularly in population biobanks comprising older individuals.

3. Refuted gene-disease assertion

Candidate gene studies and under-powered studies, which pre-date the availability of large population datasets, suffer from incorrect gene-disease associations. There are an increasing number of examples of refuted gene-disease associations, where the availability of larger datasets has proven beyond doubt that an earlier association was incorrect. For example, *SCN9A* was previously included on diagnostic panels for monogenic epilepsy conditions, but has since been shown not to be associated with the condition[39]. Associations between three genes and maturity onset diabetes of the young (MODY)[40], and all but one of 21 genes previously linked with Brugada syndrome[41], have also been refuted. Refuted gene-disease associations can nonetheless persist in case-reports, small case series, genetic testing panels, and variant databases, even following refutation. Evaluating the penetrance of variants in genes whose association with disease has not been proven will thus perpetuate the error and bias penetrance estimates downwards. International expert gene curation efforts, such as ClinGen[42] and the Gene Curation Coalition (GenCC)[43], are curating appraisals and reappraisals of the validity of gene-disease associations across thousands of monogenic conditions, and should be consulted when conducting large-scale studies of penetrance.

4. Inaccurate or outdated variant pathogenicity assertion

Even where a gene-disease association has been robustly proven, the interpretation of individual genetic variants can change as new evidence comes to light[44–46]. Although variant classification guidelines have standardised and improved variant classification[20,47],

unfortunately attempting to dichotomise all variants in genes associated with a monogenic disease into “pathogenic” or “benign” does not reflect the complexities of all classes of functional variant effects that may determine differences in phenotype manifestations. As a result, some variants have been misclassified as pathogenic without robust evidence linking them to a particular phenotype. It has been estimated that >10% of entries in commonly used databases of disease genes are incorrect, with most of the curation errors owing to a benign variant having been erroneously claimed to be pathogenic[48]. Historically, far more variants in the ClinVar database have been reclassified down (i.e. from pathogenic, likely pathogenic or uncertain to likely benign or benign) than up (i.e. from benign, likely benign or uncertain to likely pathogenic or pathogenic)[44], so any errors that persist in the database will likely bias subsequent penetrance estimates downwards. When evaluating penetrance, it is therefore advisable to treat variants of uncertain significance or with conflicting interpretations separately, or exclude them altogether. Variants labelled as likely pathogenic or pathogenic must be assessed according to an evidence framework, particularly when there are limited or old ClinVar submissions or the variants appear multiple times in the cohort, to ensure that variant classifications are as robust as possible. Expertise in the particular gene and disease can help ensure that only genuinely pathogenic variants are included.

5. Erroneous variant annotation or effect prediction

Predicting the effect of a genetic variant is important for determining whether it is likely to cause disease. Although many genes have been robustly associated with monogenic diseases, in practice, the evidence will be based on specific variants in specific locations with specific consequences. When evaluating penetrance, it is important to ensure that only variants with functionally similar consequences to those reported to cause disease are included, which may

depend upon transcript selection as well as effect prediction. For example, a missense variant that is predicted to code for an alternative amino acid in one transcript may be non-coding in all biologically relevant transcripts and thus have no effect on the protein, or result in a splicing change in specific tissues. Although nonsense and frameshifting variants are usually predicted to result in truncated or absent protein products, in practice many of these variants evade nonsense-mediated decay and may still produce a functional protein[30,49]. Another complexity is where a variant overlaps more than one transcript, or more than one gene, resulting in different predicted effects in different transcripts or different genes. It is important to ensure that variant effects are selected based on the MANE Select or MANE Plus Clinical transcript[50] (or other biologically relevant alternative transcripts) that may exclude or alter the impact of the variant in question[50,51], as well as confirm that the predicted effect relates to the gene of interest not an incidentally overlapping or nearby gene[19,52]. Including variants in penetrance calculations, whose functional effect in the gene of interest is inconsistent with the mechanism of disease, will necessarily bias the estimates downwards.

6. Inappropriate mode of inheritance

Most high-quality population studies of penetrance focus on autosomal dominant conditions, where the disease manifests in heterozygotes. It is also possible to evaluate penetrance in autosomal recessive conditions, where the disease manifests in homozygotes and compound heterozygotes (where phasing is known)[53,54]. However, although it may be interesting to evaluate whether the phenotype manifests in heterozygous carriers of autosomal recessive conditions, this should not be considered a measure of penetrance of the recessive condition[55]. Thus, it is important to ensure that the zygosity of the variants being evaluated is appropriate to the disease in question[56]. For example, Stargardt disease and other

retinopathies are caused by biallelic variants in *ABCA4*, some of which are relatively common in the population, but heterozygous carriers are unaffected[57,58] and should not be included in penetrance estimates for the disease. Another complication of studying penetrance of autosomal recessive conditions is the alternative clinical manifestations of different variant combinations, and in particular the challenge of evaluating hypomorphic variants with reduced function that are only pathogenic when seen in *trans* with a complete loss-of-function variant[59]. Digenic or oligogenic contributions caused by the combined effects of rare heterozygous variants in two or more disease genes may further complicate the evaluation of penetrance[60].

7. Incorrect mechanism of disease

Linked to variant consequence is the mechanism by which a variant causes disease, often broadly categorised into loss-of-function (LoF), gain-of-function (GoF) or dominant negative. Conditions caused by GoF and dominant-negative mechanisms require even greater caution than LoF, as typically a restricted repertoire of variants cause disease[61,62]. Moreover, GoF encompasses a wide range of functional mechanisms that may be variant and/or tissue-specific, including loss of gene silencing[63], toxic RNA or protein accumulation, and increased or novel protein function[59]. Importantly, predicted LoF variants often do not cause disease where the mechanism is GoF, and vice-versa, so evaluating their penetrance in such conditions is misleading. For example, frameshift variants in the last exon of the gene *PLIN1* result in lipodystrophy through a GoF mechanism[64], whilst haploinsufficiency is not associated with the disease[65]. In contrast, LoF variants in *PLIN1* actually cause a favourable metabolic profile, opposite to the lipodystrophy phenotype[66]. Curation of disease mechanisms and ensuring predicted variant consequences are consistent with that mechanism is important[3], particularly

where the disease is not caused by LoF or where different diseases or disease severities are caused by different variants in the same gene operating through different mechanisms. Hypomorphic variants with differing levels of residual function may also explain attenuated disease phenotypes in some cases and, where possible, constructing allelic series to evaluate genotype-phenotype correlations may be more informative than grouping variants[59,67].

8. Incorrect phenotype and case definition

Defining the phenotype associated with a disease can be extremely challenging, particularly based on Electronic Healthcare Records (EHR) or hospital episode statistics, which may be incomplete and can vary between providers. Ontologies are not always comprehensive for rare disease entities, and defining which individuals have a disease can vary based on self-report or EHR-derived data[68]. More robust ways to define individuals with a condition should be used where possible, such as detailed clinical phenotyping, questionnaires, activity monitoring, imaging, and biomarker data[69]. For example, diseases such as diabetes are particularly well captured in UK Biobank because self-report, EHR and diagnostic biomarkers are measured[7], whereas other conditions, such as intellectual disability and other conditions that lack robust biomarkers and do not routinely require hospitalisation, are not. Failure to correctly identify cases and controls may substantially bias penetrance estimates. Consideration of manifestation at different phenotype levels is also important when defining cases; for example, pathogenic variants in the gene *GCK* cause mild fasting hyperglycemia, elevated hemoglobin A1C (HbA1c) and mean fasting blood glucose, but individuals often do not get a diagnosis of diabetes unless incidentally detected and they do not get diabetes-related complications[7]. Another important factor to consider when appropriately defining phenotypes is pleiotropy[70,71]. Even where a gene-disease association has been robustly proven

and numerous variants have strong evidence robustly linking them to a phenotype, different variants in the same gene may result in very different phenotypes that present at different ages. Therefore, using a variant database such as ClinVar to define a list of pathogenic variants in a particular gene, without due regard for the associated phenotype (which unfortunately may not be well defined or annotated) can result in the penetrance of variants being assessed against the wrong disease.

9. Inappropriate accounting for age of individuals

The concept of penetrance is largely meaningless without defining the age by which a disease or associated phenotype is observed, and many monogenic conditions display age-dependent penetrance[72,73]. The penetrance of late-onset diseases cannot be accurately evaluated in younger individuals who have not yet had the chance to develop a late-onset condition, although related biochemical or functional phenotypes may be observable prior to clinical disease manifestation. Conversely, for some diseases the opposite is true; for example, variants in the gene *MC4R* appear to have higher penetrance for obesity in childhood than adulthood[74], where diet and other interventions can reduce the apparent penetrance. Some variants cause conditions that occur at birth or shortly after and will not occur in population-based studies, either because they are successfully treated or because they are terminal in childhood. For example, neonatal diabetes is a severe form of diabetes diagnosed <6 months that is unlikely to be present in adult population cohorts (e.g. UK Biobank has a lower age at recruitment of 40 years). This limits the ability to perform penetrance studies on this type of variant and condition outside of birth cohorts.

10. Inappropriate calculation of penetrance

The concept of penetrance can be used and calculated in different ways. The simplest definition, based on rare monogenic disease, is the probability of a variant carrier developing the related disease by a specific age, or the proportion of variant carriers who have manifest disease by a specific age. This definition works well for well-defined rare conditions, such as cystic fibrosis. For monogenic forms of common diseases, the background occurrence rate of the disease should also be considered when giving an estimate of risk difference[19]. Risk difference is defined as the difference between penetrance in individuals with and without the genotype of interest[75], and can be estimated as the difference in disease prevalence between individuals with variant alleles and individuals with normal alleles[19]. However, for sufficiently common diseases (or insufficiently penetrant alleles), the disease may be more prevalent in non-carriers, resulting in a negative risk difference and falsely implying that variant alleles are protective. Therefore, an appropriately adjusted Kaplan-Meier curve may be a more appropriate way to present and analyse the data[76,77], as it allows for a range of penetrance values to be calculated to different ages, whilst also estimating the uncertainty around the estimates and showing the background disease risk. Nonetheless, regardless of the methodology, common risk alleles found through genome wide association studies (GWAS) should not be evaluated for penetrance, since its definition does not apply to low-risk alleles. Although there is likely to be a continuous spectrum of effect size in reality, with no clear distinction between these categories, a conceptual debate remains over when a low-penetrance pathogenic variant becomes a risk allele, and a ClinGen Working Group has released recommendations for the reporting of such variants[78]. In many cases, factors beyond the gene of interest are likely to explain instances of incomplete penetrance and variable

expressivity[59], underlining the importance of evaluating pathogenicity in different contexts and explicitly assessing modifiers (such sex and genetic ancestry)[11] where possible.

Insufficient stringency in variant and case selection affects penetrance estimates

To illustrate the effect of these issues on penetrance estimates, we use an example from a recent large-scale attempt to estimate penetrance across a wide range of autosomal dominant monogenic conditions[19]. Amongst numerous diseases, the study included three subtypes of diabetes: neonatal diabetes (NDM), maturity onset diabetes of the young (MODY) and type 2 diabetes (T2D). Using diabetes traits and specific variants in UK Biobank as an example, we illustrate some of the errors that can occur – particularly in grouped-variant penetrance studies through the use of outdated or incorrect annotations – resulting in erroneous conclusions (see **Box 1**).

Including and grouping all variants asserted to be pathogenic or likely pathogenic in ClinVar or annotated as predicted loss-of-function within the disease-associated gene of interest, without due regard for these different issues highlighted above, is likely to lead to erroneous and uninformative penetrance estimates. An estimated median penetrance of zero[19] is likely the result of including numerous benign variants that are not linked with the disease in question, and does not represent the penetrance of true pathogenic variants. Addressing these issues and including only variants that are known to be pathogenic for monogenic disease – with the correct zygosity and mechanism – will increase the accuracy of penetrance estimates.

Starting with 74 variants linked with MODY that were included by Forrest *et al.* in analyses of the first 50,000 exome sequences released by UK Biobank[19], we found just 9 variants that the

national referral centre for monogenic diabetes testing would classify as pathogenic (https://www.diabetesgenes.org/). The number of individuals in the cohort with these MODY variants thus decreases from around 14,000 (28%) to fewer than 50 (<0.1%), and the mean diabetes risk difference between individuals with and without MODY variants increases from ~0% (**Fig. 1A**) to up to 64% (**Fig. 1B**), suggesting that the former is a gross under-estimate caused by the inclusion of large numbers of benign variants that do not cause MODY. A larger list of expertly curated pathogenic MODY variants suggests that the true MODY penetrance in UK Biobank is likely to be around 25-30% (**Fig. 1C**)[7], and that genuine variants of uncertain significance (VUS) also have a non-zero though substantially lower penetrance (**Fig. 1D)**. Whilst our estimates are still lower than penetrance estimates from clinically-ascertained cohorts[7], they are significantly higher than suggested by Forrest *et al.* in this population cohort[19]. It should be noted that for genes such as *KCNJ11,* in which a limited repertoire of activating variants cause NDM[79], it is not possible to determine the population penetrance of specific pathogenic GoF variants as there are none present in UK Biobank, potentially because they are highly penetrant and depleted from the cohort.

Conclusions

Estimates of penetrance of monogenic disease are becoming increasingly common with the availability of whole genome sequencing data in large-scale population-based cohorts. These "genotype-first" estimates provide one important estimate of the true risk of disease for people with pathogenic variants. While there are important issues around ascertainment and survivor bias that can affect penetrance estimates from any study, these nonetheless provide a lower bound to disease risk and will be important for counselling patients, particularly those with secondary or incidental findings. Often population-based estimates are lower than those

based on clinically selected cohorts, but these estimates must be based on robust data for both genotype and phenotype. Important resources, such as automated annotation tools and variant databases, should be applied with care and following critical evaluation of the outputs to ensure that variants included in penetrance estimates are truly pathogenic and relevant to the disease in question. Our guidance provides an attempt to highlight some of the common issues that can occur to ensure more studies meet these requirements. We hope that this guidance serves as a catalyst for advancing the discourse on genetic variant interpretation, and encourages the community towards more precise and scientifically robust practices.

Acknowledgements

The authors thank Prof Andrew Hattersley and numerous other colleagues and reviewers for insightful conversations and guidance. This research has been conducted using the UK Biobank Resource under Application Numbers 49847 and 9072. The current work was supported by Diabetes UK [19/0005994], MRC [MR/T00200X/1] and Wellcome [226083/Z/22/Z]. K.A.P. is supported by a Wellcome Clinical Fellowship [219606/Z/19/Z]. J.S.W is supported by the Medical Research Council (UK), Sir Jules Thorn Charitable Trust [21JTA], British Heart Foundation [RE/18/4/34215], and the NIHR Imperial College Biomedical Research Centre. The authors would like to acknowledge the use of the University of Exeter High-Performance Computing (HPC) facility in carrying out this work. This study was supported by the National Institute for Health and Care Research Exeter Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Author Contributions

C.F.W. and M.N.W. conceived the study; L.N.S. and K.A.P. performed the diabetes analysis outlined in Figure 1; C.F.W., M.N.W., L.J., A.M. and K.A.P. curated variants/genes/conditions to identify potential errors; C.F.W. wrote the first draft of the manuscript; J.S.W., D.G.M. and H.L.R. provided expert input into the manuscript; all authors contributed to revisions and the final manuscript.

Conflict of interest statement

M.N.W is a co-investigator on a Randox Laboratories R&D research grant and received translational industry academic funding from Randox Laboratories R&D relating to autoimmune GRS for prediction and classification of disease. M.N.W and K.A.P have received royalties from Randox as co-inventors of a Type 1 diabetes genetic risk score product. D.G.M. is a paid advisor to GlaxoSmithKline (GSK), Insitro, Variant Bio, and Overtone Therapeutics and has received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Google, Merck, Microsoft, Pfizer, and Sanofi-Genzyme. J.S.W. has received research support from Bristol Myers Squibb, and has acted as a consultant for MyoKardia, Pfizer, Foresite Labs, Health Lumen, and Tenaya Therapeutics. The other authors have no conflict of interest to declare.

Box 1. Examples of curation errors leading to erroneous penetrance estimates

Some classes of errors can be minimised through a judicious bioinformatics approach (**1, 2, 5**) and using regularly updated or curated databases (**3, 4, 6, 7**), whilst other classes may require detailed evaluation of individual variants (**2, 7**) and disease phenotypes (**7, 8, 9, 10**). Example variants taken from Forrest *et al.* JAMA 2022[19].

1. **Inclusion of high frequency variants**

   e.g. rs2074308 (NM_000352.6:c.1672-74G>A) in *ABCC8* has a population allele frequency of ~12% but was linked with Neonatal Diabetes Mellitus (NDM), a condition that affects fewer than 1 in 2 million people. Variants with an allele frequency far in excess of the disease in question should generally be excluded unless there is robust evidence of a strong association with disease[14].

2. **Inclusion of technical false positives**

   e.g. the commonest cause of maturity onset diabetes of the young (MODY) is a frameshifting C-insertion in a poly-C tract of *HNF1A*. This variant rs766191969 (NM_000545.8:c.863_864insC HNF1A:p.(Pro289AlafsTer28)) is challenging to sequence accurately, but was apparently present in 330 individuals and included with no suggestion that genotyping accuracy was evaluated[7].

3. **Inclusion of variants from refuted genes**

   e.g. variants in *PAX4* were included, presumably due to the presence of a pathogenic variant in ClinVar linked with MODY dating from 2007, despite the fact that the gene has since been refuted as a cause of MODY[40]. Notably, variants in other genes robustly linked with MODY, such as *RFX6,* were not included[22].

4. **Inclusion of variants with conflicting or uncertain evidence**

e.g. a missense variant rs1131691259 in *WFS1* (NM_006005.3:c.2074C>T p.(His692Tyr)) formerly classified as likely pathogenic that is now classified as uncertain significance or uncertain risk allele in ClinVar. Other included variants, such as rs397518475 (NM_006208.3:c.530G>A p.(Cys177Tyr)) in *ENPP1* have "no assertion criteria" in ClinVar but were still labelled pathogenic despite the gene *ENPP1* only being associated with diabetes through a refuted common GWAS variant[80].

5. **Inclusion of variants with incorrect annotations**

   e.g. four predicted LoF variants in *MYL7* were included as causes of NDM, despite the gene never having been linked to the condition. Since *MYL7* is a nearby neighbour of *GCK*, in which numerous variants have been shown to cause NDM[81,82], these variants were presumably included due to a bioinformatics error based on their predicted consequences across both genes (e.g. stop-gain variant in *MYL7* but downstream gene variant in *GCK*).

6. **Inclusion of variants with incorrect zygosity for the mode of inheritance of the disease**

   e.g. heterozygous LoF variants in *SLC2A2* have no phenotype although homozygous pathogenic variants in *SLC2A2* cause NDM; despite this, numerous heterozygous LoF variants of *SLC2A2* were included in the penetrance evaluation.

7. **Inclusion of pathogenic variants with the wrong mechanism**

   e.g. heterozygous GoF missense variants in *ABCC8* and *KCNJ11* cause NDM <6 months, whilst homozygous LoF variants in these genes cause the opposite phenotype of familial hyperinsulinism[79]. Importantly, heterozygous LoF variants in *ABCC8* and *KCNJ11* have no phenotype, but were nonetheless included in diabetes penetrance estimates.

8. **Sub-optimal phenotype and case definition**

e.g. the same ICD10 code was used for T2D and MODY. The definition of diabetes could be substantially improved, for example by the inclusion of HbA1c, glucose, the self-reported questionnaire, and primary care data – all of which are available in UK Biobank. Failure to identify these individuals as cases rather than controls would bias penetrance estimates downwards[83].

9. **Inappropriate age of individuals**

   e.g. NDM occurs <6 months and, because of its severity and rarity, there is likely a substantial bias away from pathogenic variants for these genes in population based studies with recruitment >40 years.

10. **Inappropriate calculation of penetrance**

    e.g. as a result of calculating the risk difference versus the background rate of diabetes in the UK Biobank, some variants appear artifactually to be protective (i.e. lower disease prevalence than background), which further biases the penetrance estimate downwards by including negative point estimates in calculations of averages.

Figure legend

Fig. 1. Comparison of penetrance estimates in diabetes, calculated using risk difference of groups of variants purportedly linked with MODY in 50,000 individuals from UK Biobank with exome sequencing data.

The complete penetrance dataset from Forrest *et al.* JAMA 2022 was downloaded from Mendeley Data (accessed 4 July 2022, as outlined in Data Sharing Statement in Supplement)[19]. Variants from that dataset in UK Biobank linked with the disease "Maturity-onset diabetes of the young" and annotated as pathogenic or likely pathogenic in ClinVar or predicted to result in loss-of-function (frameshift, stop-gain, splice acceptor, splice donor) were retained. Variants were then re-annotated using Ensembl VEP[84] and their presence confirmed in the UK Biobank 50K exome release and curated. Additional pathogenic MODY variants were obtained from the Exeter clinical database (https://www.diabetesgenes.org/). Diabetes was defined either narrowly using ICD10 code E11x only **(red)**, or broadly using ICD10 coding with self-reported questionnaire and HbA1C ≥48 mmol/mol[7] **(blue)**. Risk difference was calculated as the difference in disease prevalence between individuals with variant alleles and individuals with normal alleles for each variant and analysed for each of the individual variants included in each group **(left)**, where the height of the points reflects the number of variants at a specific risk difference, as well as for all pathogenic MODY variants grouped together with 95% confidence intervals **(right)**. Results are presented in four groups: the first group (**A**) includes 74 variants previously annotated as causative[19]; the second group (**B**) contains 9 variants from the initial dataset[19] that were judged to be pathogenic for MODY following expert curation and is an extremely conservative estimate of penetrance; the third group (**C**) contains an additional 27 clinically-reported pathogenic variants (not included in the previous set), judged to be pathogenic for MODY following expert curation and likely representing the most accurate

estimate of MODY penetrance in this cohort; the fourth group **(D)** contains variants of uncertain significance (VUS), defined as rare (gnomADv2.1 minor allele count <2)[85] missense variants predicted to be damaging (REVEL >0.7)[86] with ClinVar classifications of "uncertain", "conflicting" or "unknown"[87]. Numbers in brackets = number of variants included in each group, followed by the number of heterozygotes in 50,000 individuals in UK Biobank.

Bibliography

1. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).

2. Kingdom, R. & Wright, C. F. Incomplete Penetrance and Variable Expressivity: From Clinical Studies to Population Cohorts. *Frontiers in Genetics* (2022).

3. Roberts, A. M. *et al.* Towards robust clinical genome interpretation: developing a consistent terminology to characterize disease-gene relationships - allelic requirement, inheritance modes and disease mechanisms. *medRxiv* (2023) doi:10.1101/2023.03.30.23287948.

4. Otto, P. A. & Horimoto, A. R. V. R. Penetrance rate estimation in autosomal dominant conditions. *Genet. Mol. Biol.* **35**, 583–588 (2012).

5. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).

6. Wright, C. F. *et al.* Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting. *Am. J. Hum. Genet.* **104**, 275–286 (2019).

7. Mirshahi, U. L. *et al.* Reduced penetrance of MODY-associated HNF1A/HNF4A variants but not GCK variants in clinically unselected cohorts. *Am. J. Hum. Genet.* **109**, 2018–2028 (2022).

8. Pizzo, L. *et al.* Rare variants in the genetic background modulate cognitive and developmental phenotypes in individuals carrying disease-associated variants. *Genet. Med.* **21**, 816–825 (2019).

9. Crawford, K. *et al.* Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. *J. Med. Genet.* **56**, 131–138 (2019).

10. McGurk, K. A. *et al.* The penetrance of rare variants in cardiomyopathy-associated genes: A cross-sectional approach to estimating penetrance for secondary findings. *Am. J. Hum. Genet.* **110**, 1482–1495 (2023).

11. Ciesielski, T. H., Sirugo, G., Iyengar, S. K. & Williams, S. M. Characterizing the pathogenicity of genetic variants: the consequences of context. *NPJ Genom. Med.* **9**, 3 (2024).

12. Kassabian, B. *et al.* Intrafamilial variability in SLC6A1-related neurodevelopmental disorders. *Front. Neurosci.* **17**, 1219262 (2023).

13. Martins Custodio, H. *et al.* Widespread genomic influences on phenotype in Dravet syndrome, a “monogenic” condition. *Brain* **146**, 3885–3897 (2023).

14. Minikel, E. V. *et al.* Quantifying prion disease penetrance using large population control cohorts. *Sci. Transl. Med.* **8**, 322ra9 (2016).

15. Fan, X. *et al.* Penetrance of Breast Cancer Susceptibility Genes From the eMERGE III Network. *JNCI Cancer Spectr* **5**, (2021).

16. Shekari, S. *et al.* Penetrance of pathogenic genetic variants associated with premature ovarian insufficiency. *Nat. Med.* **29**, 1692–1699 (2023).

17. de Marvao, A. *et al.* Phenotypic expression and outcomes in individuals with rare genetic variants of hypertrophic cardiomyopathy. *J. Am. Coll. Cardiol.* **78**, 1097–1110 (2021).

18. Goodrich, J. K. *et al.* Determinants of penetrance and variable expressivity in monogenic metabolic conditions across 77,184 exomes. *Nat. Commun.* **12**, 3505 (2021).

19. Forrest, I. S. *et al.* Population-Based Penetrance of Deleterious Clinical Variants. *JAMA* **327**, 350–359 (2022).

20. Richards, 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.* **17**, 405–424 (2015).

21. Ghosh, R. *et al.* Updated recommendation for the benign stand-alone ACMG/AMP criterion. *Hum. Mutat.* **39**, 1525–1530 (2018).

22. Patel, K. A. *et al.* Heterozygous RFX6 protein truncating variants are associated with MODY with reduced penetrance. *Nat. Commun.* **8**, 888 (2017).

23. Wiltshire, K. M., Hegele, R. A., Innes, A. M. & Brownell, A. K. W. Homozygous lamin A/C familial lipodystrophy R482Q mutation in autosomal recessive Emery Dreifuss muscular dystrophy. *Neuromuscul. Disord.* **23**, 265–268 (2013).

24. Minikel, E. V. & MacArthur, D. G. Publicly Available Data Provide Evidence against NR1H3 R415Q Causing Multiple Sclerosis. *Neuron* **92**, 336–338 (2016).

25. Hanany, M. & Sharon, D. Allele frequency analysis of variants reported to cause autosomal dominant inherited retinal diseases question the involvement of 19% of genes and 10% of reported pathogenic variants. *J. Med. Genet.* **56**, 536–542 (2019).

26. Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from gnomAD. *Hum. Mutat.* **43**, 1012–1030 (2022).

27. Whiffin, N. *et al.* CardioClassifier: disease- and gene-specific computational decision support for clinical genome interpretation. *Genet. Med.* **20**, 1246–1254 (2018).

28. Laver, T. W. *et al.* The Common p.R114W HNF4A Mutation Causes a Distinct Clinical Subtype of Monogenic Diabetes. *Diabetes* **65**, 3212–3217 (2016).

29. Loveday, C. *et al.* p.Val804Met, the Most Frequent Pathogenic Mutation in RET, Confers a Very Low Lifetime Risk of Medullary Thyroid Cancer. *J. Clin. Endocrinol. Metab.* **103**, 4275–4282 (2018).

30. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).

31. Weedon, M. N. *et al.* Very rare pathogenic genetic variants detected by SNP-chips are usually false positives: implications for direct-to-consumer genetic testing. *BioRxiv* (2019) doi:10.1101/696799.

32. Weedon, M. N., Wright, C. F., Patel, K. A. & Frayling, T. M. Unreliability of genotyping arrays for detecting very rare variants in human genetic studies: Example from a recent study of MC4R. *Cell* **184**, 1651 (2021).

33. Valluru, M. K. *et al.* A founder UMOD variant is a common cause of hereditary nephropathy in the British population. *J. Med. Genet.* **60**, 397–405 (2023).

34. Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).

35. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics* **14**, 178–192 (2013).

36. Carlston, C. M. *et al.* Pathogenic ASXL1 somatic variants in reference databases complicate germline variant interpretation for Bohring-Opitz Syndrome. *Hum. Mutat.* **38**, 517–523 (2017).

37. Steensma, D. P. Clinical implications of clonal hematopoiesis. *Mayo Clin. Proc.* **93**, 1122–1130 (2018).

38. Ariste, O., de la Grange, P. & Veitia, R. A. Recurrent missense variants in clonal hematopoiesis-related genes present in the general population. *Clin. Genet.* (2022) doi:10.1111/cge.14259.

39. Fasham, J. *et al.* No association between SCN9A and monogenic human epilepsy disorders. *PLoS Genet.* **16**, e1009161 (2020).

40. Laver, T. W. *et al.* Evaluation of evidence for pathogenicity demonstrates that BLK, KLF11, and PAX4 should not be included in diagnostic testing for MODY. *Diabetes* **71**, 1128–1136 (2022).

41. Hosseini, S. M. *et al.* Reappraisal of Reported Genes for Sudden Arrhythmic Death: Evidence-Based Evaluation of Gene Validity for Brugada Syndrome. *Circulation* **138**, 1195–1205 (2018).

42. Strande, N. T. *et al.* Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. *Am. J. Hum. Genet.* **100**, 895–906 (2017).

43. DiStefano, M. T. *et al.* The Gene Curation Coalition: A global effort to harmonize gene-disease evidence resources. *Genet. Med.* **24**, 1732–1742 (2022).

44. Harrison, S. M. & Rehm, H. L. Is “likely pathogenic” really 90% likely? Reclassification data in ClinVar. *Genome Med.* **11**, 72 (2019).

45. Mighton, C. *et al.* Variant classification changes over time in BRCA1 and BRCA2. *Genet. Med.* (2019) doi:10.1038/s41436-019-0493-2.

46. Shah, N. *et al.* Identification of misclassified clinvar variants via disease population prevalence. *Am. J. Hum. Genet.* **102**, 609–619 (2018).

47. Ellard, S. *et al.* ACGS Best Practice Guidelines for Variant Classification in Rare Disease 2020. (2020).

48. Biesecker, L. G. Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: lessons from the ClinSeq project. *Genet. Med.* **14**, 393–398 (2012).

49. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

50. Wright, C. F., FitzPatrick, D. R., Ware, J. S., Rehm, H. L. & Firth, H. V. Importance of adopting standardized MANE transcripts in clinical reporting. *Genet. Med.* **25**, 100331 (2023).

51. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).

52. Karlin, S., Chen, C., Gentles, A. J. & Cleary, M. Associations between human disease genes and overlapping gene groups and multiple amino acid runs. *Proc Natl Acad Sci USA* **99**, 17008–17013 (2002).

53. Barton, A. R., Hujoel, M. L. A., Mukamel, R. E., Sherman, M. A. & Loh, P.-R. A spectrum of recessiveness among Mendelian disease variants in UK Biobank. *Am. J. Hum. Genet.* **109**, 1298–1307 (2022).

54. Lipov, A. *et al.* Exploring the complex spectrum of dominance and recessiveness in genetic cardiomyopathies. *Nat. Cardiovasc. Res.* (2023) doi:10.1038/s44161-023-00346-3.

55. Heyne, H. O. *et al.* Mono- and biallelic variant effects on disease at biobank scale. *Nature* **613**, 519–525 (2023).

56. Ellard, S., Colclough, K., Patel, K. A. & Hattersley, A. T. Prediction algorithms: pitfalls in interpreting genetic variants of autosomal dominant monogenic diabetes. *The Journal of Clinical Investigation* (2020).

57. Cremers, F. P. M., Lee, W., Collin, R. W. J. & Allikmets, R. Clinical spectrum, genetic complexity and therapeutic approaches for retinal disease caused by ABCA4 mutations. *Prog. Retin. Eye Res.* **79**, 100861 (2020).

58. Runhart, E. H. *et al.* The Common ABCA4 Variant p.Asn1868Ile Shows Nonpenetrance and Variable Expression of Stargardt Disease When Present in trans With Severe Variants. *Invest. Ophthalmol. Vis. Sci.* **59**, 3220–3231 (2018).

59. Zschocke, J., Byers, P. H. & Wilkie, A. O. M. Mendelian inheritance revisited: dominance and recessiveness in medical genetics. *Nat. Rev. Genet.* **24**, 442–463 (2023).

60. Cicerone, A. P. *et al.* A Survey of Multigenic Protein-Altering Variant Frequency in Familial Exudative Vitreo-Retinopathy (FEVR) Patients by Targeted Sequencing of Seven FEVR-Linked Genes. *Genes* **13**, (2022).

61. Backwell, L. & Marsh, J. A. Diverse Molecular Mechanisms Underlying Pathogenic Protein Mutations: Beyond the Loss-of-Function Paradigm. *Annu. Rev. Genomics Hum. Genet.* **23**, 475–498 (2022).

62. Gerasimavicius, L., Livesey, B. J. & Marsh, J. A. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nat. Commun.* **13**, 3895 (2022).

63. Wakeling, M. N. *et al.* Non-coding variants disrupting a tissue-specific regulatory element in HK1 cause congenital hyperinsulinism. *Nat. Genet.* **54**, 1615–1620 (2022).

64. Gandotra, S. *et al.* Perilipin deficiency and autosomal dominant partial lipodystrophy. *N. Engl. J. Med.* **364**, 740–748 (2011).

65. Laver, T. W. *et al.* PLIN1 haploinsufficiency is not associated with lipodystrophy. *J. Clin. Endocrinol. Metab.* **103**, 3225–3230 (2018).

66. Patel, K. A. *et al.* PLIN1 haploinsufficiency causes a favorable metabolic profile. *J. Clin. Endocrinol. Metab.* **107**, e2318–e2323 (2022).

67. Magge, S. N. *et al.* Familial leucine-sensitive hypoglycemia of infancy due to a dominant mutation of the beta-cell sulfonylurea receptor. *J. Clin. Endocrinol. Metab.* **89**, 4450–4456 (2004).

68. DeBoever, C. *et al.* Assessing digital phenotyping to enhance genetic studies of human diseases. *Am. J. Hum. Genet.* **106**, 611–622 (2020).

69. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

70. Jacob, K. N. & Garg, A. Laminopathies: multisystem dystrophy syndromes. *Mol. Genet. Metab.* **87**, 289–302 (2006).

71. Magrinelli, F., Balint, B. & Bhatia, K. P. Challenges in Clinicogenetic Correlations: One Gene - Many Phenotypes. *Mov Disord Clin Pract (Hoboken)* **8**, 299–310 (2021).

72. Pilling, L. C. *et al.* Common conditions associated with hereditary haemochromatosis genetic variants: cohort study in UK Biobank. *BMJ* **364**, k5222 (2019).

73. Murphy, N. A. *et al.* Age-related penetrance of the C9orf72 repeat expansion. *Sci. Rep.* **7**, 2116 (2017).

74. Wade, K. H. *et al.* Loss-of-function mutations in the melanocortin 4 receptor in a UK birth cohort. *Nat. Med.* **27**, 1088–1096 (2021).

75. Khoury, M. J. & Flanders, W. D. On the measurement of susceptibility to genetic factors. *Genet. Epidemiol.* **6**, 699–711 (1989).

76. Bland, J. M. & Altman, D. G. Survival probabilities (the Kaplan-Meier method). *BMJ* **317**, 1572 (1998).

77. Jonker, M. A., Rijken, J. A., Hes, F. J., Putter, H. & Hensen, E. F. Estimating the penetrance of pathogenic gene variants in families with missing pedigree information. *Stat. Methods Med. Res.* **28**, 2924–2936 (2019).

78. Lebo, M. *et al.* O31: Risk allele evidence curation, classification, and reporting: Recommendations from the ClinGen Low Penetrance/Risk Allele Working Group*. *Genetics in Medicine Open* **1**, 100457 (2023).

79. De Franco, E. *et al.* Update of variants identified in the pancreatic β-cell KATP channel genes KCNJ11 and ABCC8 in individuals with congenital hyperinsulinism and diabetes. *Hum. Mutat.* **41**, 884–905 (2020).

80. Weedon, M. N. *et al.* No evidence of association of ENPP1 variants with type 2 diabetes or obesity in a study of 8,089 U.K. Caucasians. *Diabetes* **55**, 3175–3179 (2006).

81. Hughes, A. E. *et al.* Identification of GCK-maturity-onset diabetes of the young in cases of neonatal hyperglycemia: A case series and review of clinical features. *Pediatr Diabetes* **22**, 876–881 (2021).

82. Raimondo, A. *et al.* Phenotypic severity of homozygous GCK mutations causing neonatal or childhood-onset diabetes is primarily mediated through effects on protein stability. *Hum. Mol. Genet.* **23**, 6432–6440 (2014).

83. Bastarache, L. & Peterson, J. F. Penetrance of deleterious clinical variants. *JAMA* **327**, 1926–1927 (2022).

84. McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

85. Wang, Q. *et al.* Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nat. Commun.* **11**, 2539 (2020).

86. Ioannidis, N. M. *et al.* REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).

87. Harrison, S. M. *et al.* Using clinvar as a resource to support variant interpretation. *Curr. Protoc. Hum. Genet.* **89**, 8.16.1-8.16.23 (2016).