Spargo *et al. Genome Medicine* (2022) 14:141
https://doi.org/10.1186/s13073-022-01142-7

Genome Medicine

METHOD

Open Access

Calculating variant penetrance from family history of disease and average family size in population-scale data

Thomas P. Spargo[1], Sarah Opie-Martin[1], Harry Bowles[1], Cathryn M. Lewis[2,3], Alfredo Iacoangeli[1,4,5*] and Ammar Al-Chalabi[1,6*]

Abstract

Background: Genetic penetrance is the probability of a phenotype when harbouring a particular pathogenic variant. Accurate penetrance estimates are important across biomedical fields including genetic counselling, disease research, and gene therapy. However, existing approaches for penetrance estimation require, for instance, large family pedigrees or availability of large databases of people affected and not affected by a disease.

Methods: We present a method for penetrance estimation in autosomal dominant phenotypes. It examines the distribution of a variant among people affected (cases) and unaffected (controls) by a phenotype within population-scale data and can be operated using cases only by considering family disease history. It is validated through simulation studies and candidate variant-disease case studies.

Results: Our method yields penetrance estimates which align with those obtained via existing approaches in the Parkinson's disease *LRRK2* gene and pulmonary arterial hypertension *BMPR2* gene case studies. In the amyotrophic lateral sclerosis case studies, examining penetrance for variants in the *SOD1* and *C9orf72* genes, we make novel penetrance estimates which correspond closely to understanding of the disease.

Conclusions: The present approach broadens the spectrum of traits for which reliable penetrance estimates can be obtained. It has substantial utility for facilitating the characterisation of disease risks associated with rare variants with an autosomal dominant inheritance pattern. The yielded estimates avoid any kinship-specific effects and can circumvent ascertainment biases common when sampling rare variants among control populations.

Keywords: Penetrance, Variant, Autosomal dominant, Disease, Familial, Sporadic

Background

Penetrance is the probability of developing a specific trait given a genetic variant or set of variants. Some pathogenic variants are fully penetrant, and people harbouring them always develop the associated phenotype. For instance, a trinucleotide CAG repeat expansion within the *HTT* gene [MIM: 613004] is fully penetrant for Huntington's Disease [MIM: 143100] by 80 years of age among people harbouring an expansion variant larger than 41 repeats [1]. For many variants, however, penetrance is incomplete, and those with risk variants can remain unaffected throughout their life. For example, the p.Gly2019Ser (c.6055G > A) variant of the *LRRK2* gene [MIM: 609007] exhibits incomplete penetrance for Parkinson's disease (PD [MIM: 168600]), meaning that it

Alfredo Iacoangeli and Ammar Al-Chalabi are co-senior authors.

*Correspondence: alfredo.iacoangeli@kcl.ac.uk; ammar.al-chalabi@kcl.ac.uk

[1] Department of Basic and Clinical Neuroscience, Maurice Wohl Clinical Neuroscience Institute, King's College London, London SE5 9RX, UK
Full list of author information is available at the end of the article

elevates risk for PD but does not necessarily result in its manifestation [2].

In medical genetics, estimating the penetrance of pathogenic variants is vital for the correct interpretation of genetic test results. This importance will increase as genome sequencing becomes routine, both within and outside clinical practice, alongside advancements in precision medicine and gene therapy [3–6].

Several methods exist for penetrance estimation. The first and most widely used is based on statistical examination of how the variant segregates with the phenotype within pedigrees [7]. However, the generalisability of estimates derived from specific families may be limited. Other approaches examine the incidence of disease in a sample of unrelated people who harbour a variant [8, 9]. Without systematic sampling, these estimates can be affected by ascertainment bias. Where large pedigrees are not available, or if the disease is rare or late onset, these techniques may not be possible [10].

Estimating penetrance for a variant of unknown significance identified, for example, during genome sequencing-based screening can be particularly challenging. The problem is exemplified by the large number of reported *SOD1* gene [MIM: 147450] variants implicated in amyotrophic lateral sclerosis (ALS [MIM: 105400]). ALS is a fatal neurodegenerative disease characterised predominantly by progressive degeneration of motor neurons [11, 12]. *SOD1* variants are an important cause of ALS and over 180 ALS-associated variants in this gene have been reported to date [11, 13–15], however, family pedigrees suitable for establishing penetrance are available for only a minority of these.

We have developed a new method to estimate penetrance for variants with an autosomal dominant inheritance pattern using population-level data from unrelated people who are and are not affected by the associated phenotype. It can be operated using variant information drawn only from affected populations, stratified according to the family history between 'familial' and 'sporadic' disease presentations. This approach is based on our previously published model of disease which explains how variant penetrance and sibship size determine the presence or absence of a disease for families in which the variant occurs [16].

The method is complementary to and fills an important gap left by existing techniques. Using population-scale data, it takes full advantage of the rapidly growing quantity of genetic data that are being generated for a wide range of human disease and therefore it is ideally placed to be a valuable tool in the precision medicine era. Moreover, the capacity to assess penetrance based on the distribution of a variant between samples of unrelated people drawn only from the affected population allows estimates unbiased by kinship-specific effects or ascertainment of unaffected population members.

We have tested the approach in four variant-disease case studies, drawing upon the most common and widely studied autosomal dominant variants implicated in each disease: the *LRRK2* p.Gly2019Ser variant for PD [2]; variants in the *BMPR2* gene [MIM: 600799] for heritable pulmonary arterial hypertension (PAH [MIM: 178600]) [17]; and variants in the *SOD1* and *C9orf72* [MIM: 614260] genes for ALS [11, 12].

Methods

Model

Here, we describe an approach to estimate genetic penetrance for autosomal dominant traits using population-scale data.

Our method builds upon and extends an existing disease model [16] which makes the following assumptions: in a nuclear family, a rare dominant pathogenic variant is necessary but not sufficient for disease to occur, therefore penetrance, denoted f, is not complete and family members who do not harbour the variant are not affected; all variants are inherited from exactly one parent, thus there are no people homozygous for the variant or de novo variants. Our extended model relaxes the assumption that the variant is necessary for disease to occur: it assumes that people not harbouring the variant have a residual risk for developing disease after accounting for the proportion of disease occurrences attributed to the variant, denoted g.

Accordingly, if the probability of an individual being affected by a disease, $P(A)$, is f when harbouring variant M or g if M is absent, denoted $M^{'}$, $P(A)$ can be determined by considering the probability of harbouring M, $P(M)$:

$$P(A) = f \times P(M) + g \times P\left(M^{'}\right), \tag{1}$$

letting $P\left(M'\right) = 1 - P(M)$.

In a family where a single parent harbours, and each child has a 0.5 chance of inheriting, M, the following probabilities of being affected can be determined per family member:

$$P(A)^{M} = f \tag{2}$$

for the variant harbouring parent, where $P(M) = 1$;

$$P(A)^{M^{0.5}} = \frac{f}{2} + \frac{g}{2} \tag{3}$$

for each offspring, each of whom has $P(M) = 0.5$, and thus risk influenced by both f and g; and

$$P(A)^{M'} = g \tag{4}$$

for the parent without M, where $P(M) = 0$, and therefore for whom disease risk is only determined by that which is associated with M'.

Considering these individual disease probabilities, three probabilities can be determined for a nuclear family where one parent harbours a given variant: that no family members are affected, $P(unaffected)$; that exactly one member is affected, $P(sporadic)$; and that more than one member is affected, $P(familial)$. These probabilities are determined by penetrance, f, residual disease risk g if not harbouring the variant, and sibship size, N. In a family with N siblings:

$$P\left(unaffected\right) = \left(1-f\right)\left(1-\frac{f}{2}-\frac{g}{2}\right)^{N}\left(1-g\right), \tag{5}$$

where no family member, with or without the variant, develops the disease, and where each of the sibs has ½ probability of being transmitted the variant.

$$P(sporadic) = f\left(1-\frac{f}{2}-\frac{g}{2}\right)^{N}(1-g) + (1-f)N\left(\frac{f}{2}+\frac{g}{2}\right)\left(1-\frac{f}{2}-\frac{g}{2}\right)^{N-1}(1-g) + (1-f)\left(1-\frac{f}{2}-\frac{g}{2}\right)^{N}g, \tag{6}$$

if one family member develops the disease. This may be either may be the variant-harbouring parent, exactly one of the sibs, or the parent not harbouring the variant (on account of residual risk g). Then,

$$P\left(familial\right) = 1 - P\left(unaffected\right) - P(sporadic) = 1 - \left(\begin{array}{c} \left(1-f\right)\left(1-\frac{f}{2}-\frac{g}{2}\right)^{N}\left(1-g\right)+ \\ f\left(1-\frac{f}{2}-\frac{g}{2}\right)^{N}\left(1-g\right)+ \\ (1-f)N\left(\frac{f}{2}+\frac{g}{2}\right)\left(1-\frac{f}{2}-\frac{g}{2}\right)^{N-1}\left(1-g\right)+ \\ (1-f)\left(1-\frac{f}{2}-\frac{g}{2}\right)^{N}g \end{array}\right), \tag{7}$$

where two or more family members develop the disease, which can be determined from $P(unaffected)$ and $P(sporadic)$ since the total probability of a family being unaffected, sporadic, or familial must sum to 1.

If $g = 0$, the original [16] and extended models are equivalent.

Application to penetrance calculation

Conversely, penetrance can be estimated given the observed rates of the unaffected, sporadic, and familial disease states in families where the pathogenic variant occurs, the average sibship size for these families, and an estimate of residual disease risk g. We can also estimate penetrance based on the observed rates of families presenting as unaffected versus *affected*, a fourth disease state whereby

Table 1 Valid disease state combinations and corresponding weighting factors for estimating disease state rates

Variant frequencies provided	Required weighting factors
Familial (M_F), Sporadic (M_S)	$W_F = P(F\|A)$, $W_S = P(S\|A)$
Familial (M_F), Unaffected (M_U)	$W_F = P(F\|A) \times P(A)^{pop}$, $W_U = 1 - P(A)^{pop}$
Sporadic (M_S), Unaffected (M_U)	$W_S = P(S\|A) \times P(A)^{pop}$, $W_U = 1 - P(A)^{pop}$
Familial (M_F), Sporadic (M_S), Unaffected (M_U)	$W_F = P(F\|A) \times P(A)^{pop}$, $W_S = P(S\|A) \times P(A)^{pop}$, $W_U = 1 - P(A)^{pop}$
Affected (M_A), Unaffected (M_U)	$W_A = P(A)^{pop}$, $W_U = 1 - P(A)^{pop}$

The defined weighting factors are used in Step 1 of the penetrance estimation approach, as described in Fig. 1 and Additional File 1: Sect. 1.1$M_{F,S,U,A}$= variant frequencies in the familial, sporadic, unaffected, and affected states;$W_{F,S,U,A}$= weighting factors for the familial, sporadic, unaffected, and affected states;$P(A)^{pop}$= the probability of a member of the sampled population being affected;$P(F|A)$= disease familiality rate;$P(S|A)$= disease sporadic rate

$$P\left(affected\right) = P\left(familial\right) + P(sporadic). \tag{8}$$

The observed rate of the arbitrarily-labelled disease state 'X', $R(X)^{obs}$, is used to indicate the frequency of one of the sampled disease states across all states sampled. $R(X)^{obs}$ can be specified for any valid combination of the four disease states, drawing from any two or three of the familial, sporadic, and unaffected disease states, or from the affected and unaffected states. Data from the affected state cannot be specified alongside that of the familial or sporadic disease states since the former is determined through their combination. $R(X)^{obs}$ may be specified directly if the distribution of disease states across people with the variant is known for the state-combination used or derived as a weighted proportion of estimates of heterozygous variant frequency across people with and without the variant (see Table 1).

Sibship size can be estimated for the sample either directly, based on the average sibship size among sampled families, or indirectly, by designating an estimate representative of the sample (e.g. from global databases).

Under Bayes theorem [18], g can be determined from $P(A)$ and $P(M)$ within the general population, respectively $P(A)^{pop}$ and $P(M)^{pop}$, and the frequency of variant M among people affected by disease, M_A:

$$g = \frac{P(A)^{pop} \times (1 - M_A)}{\left(1 - P(M)^{pop}\right)}. \tag{9}$$

M_A and $P(M)^{pop}$ may each be determined by weighted sums:

$$M_A = M_F \times P(F|A) + M_S \times P(S|A), \tag{10}$$

and

$$P(M)^{pop} = M_A \times P(A)^{pop} + M_U \times \left(1 - P(A)^{pop}\right), \tag{11}$$

where $M_{F,S,U}$ denote the variant frequencies in the familial, sporadic, and unaffected states, $P(F|A)$ is the rate at which people in the affected population, A, are familial, and $P(S|A)$ is the disease sporadic rate ($P(S|A) = 1 - P(F|A)$). If the disease is rare in the population, $g \approx 0$ and has a negligible influence upon penetrance estimates (see the simulation studies in Additional File 1: Sect. 1.2.3).

Our penetrance calculation method involves four steps and includes the option to derive error in the estimate. These processes are summarised in Fig. 1 and comprehensively outlined in Additional File 1: Sect. 1.1.

The method assumes that: one person is sampled per family and disease states are assigned based on the status of the person sampled and first-degree family members only; all variants are inherited from exactly one parent and there are no de novo variants; the value specified for sibship size is representative of sibship size across disease state groups. We recommend providing an estimate of g, however, $g = 0$ by default, which makes the additional assumption that the trait only occurs in members of sampled families owing to the presence of the variant.

A further assumption is made in each of the two scenarios for determining $R(X)^{obs}$. When sampling across only families where the variant occurs, it is assumed that disease state classifications for sampled families will not change at a future time. When estimating variant frequencies within disease states across cohorts of people with and without the variant, it is assumed that family disease states change comparably over time for people with and without the variant. The latter assumption can be partially tested by examining whether age of disease onset is equally variable for people with and without the tested variant; the assumption is further discussed in Additional File 1: Sect. 1.2.2.

Additional File: Sect. 1.2 outlines the steps taken for approach validation, including details of several simulation studies and comparison between using a lookup table or maximum-likelihood approach for Step 3. The included simulation studies test accuracy in penetrance estimation when input parameters are correctly or incorrectly specified, when g is accurately measured or assumed to be 0, and according to age of sampling across several scenarios.

We have made this approach available as the R function *adpenetrance* hosted on GitHub [19]. In the GitHub repository, we additionally provide functions to calculate g; test for equal onset variability across two groups; and simulate how a certain degree of unequal onset variability, as indicated by the previous function, may affect penetrance estimates. To facilitate easy use, the approach is also hosted on a publicly available web-server, developed using the R Shiny package (version 1.7.3) [20, 21]. The web tool is further described in the Additional File: Sect. 1.3, and Fig. 2 presents an example of its usage.

Case studies

Input parameters for included case studies were estimated using publicly available data. Variant frequencies were estimated across people with and without the variant in the familial, M_F, and sporadic, M_S, states in all cases and, in case 1, the unaffected state, M_U. M_U was integrated into penetrance estimation for case study 1 only to demonstrate the application of the method when sampling from various disease state combinations. This was not applied to other case studies as estimation focusses upon rare variants liable to ascertainment bias in control populations. In all cases, we derived the standard error of these values, $\sigma_{\overline{M_X}}$, to allow for assessment of error in the penetrance estimate. Variant frequency estimates were weighted to calculate $R(X)^{obs}$ among variant-harbouring families from those states modelled using the factors presented in Table 1. Accordingly, $P(F|A)$ and $P(S|A)$ were defined as weighting factors in all cases. $P(A)^{pop}$ is used in all case studies to derive g, according to Eqs. (9–11), and is used as a weighting factor in case 1 only.

Sibship size, N, was estimated in each case based on the Total Fertility Rates reported in the World Bank database [22] for the world region(s) best representing the sample.

An R script permitting replication of each case study is provided within our GitHub repository (see the 'Availability of data and materials' [19]).

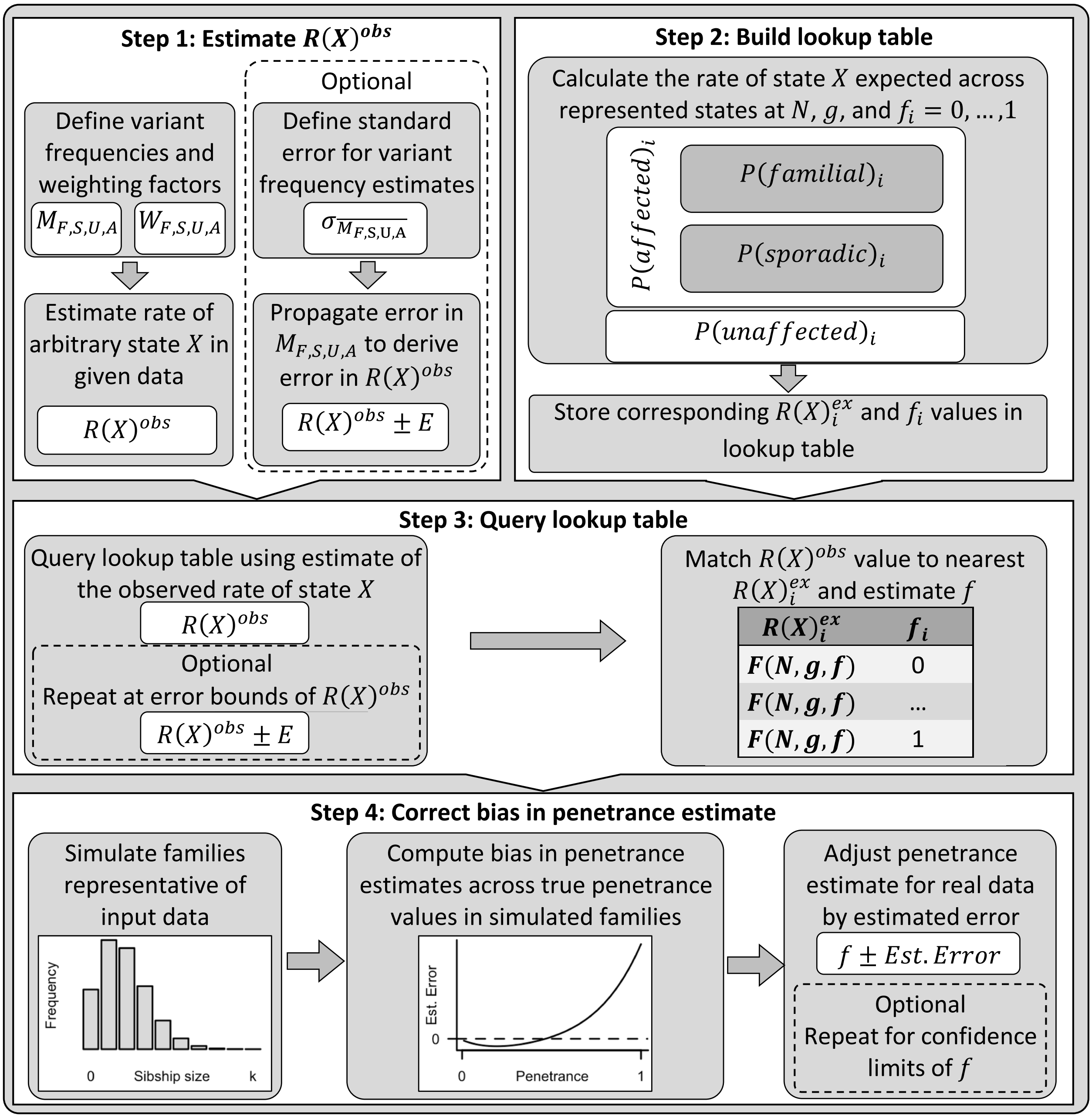
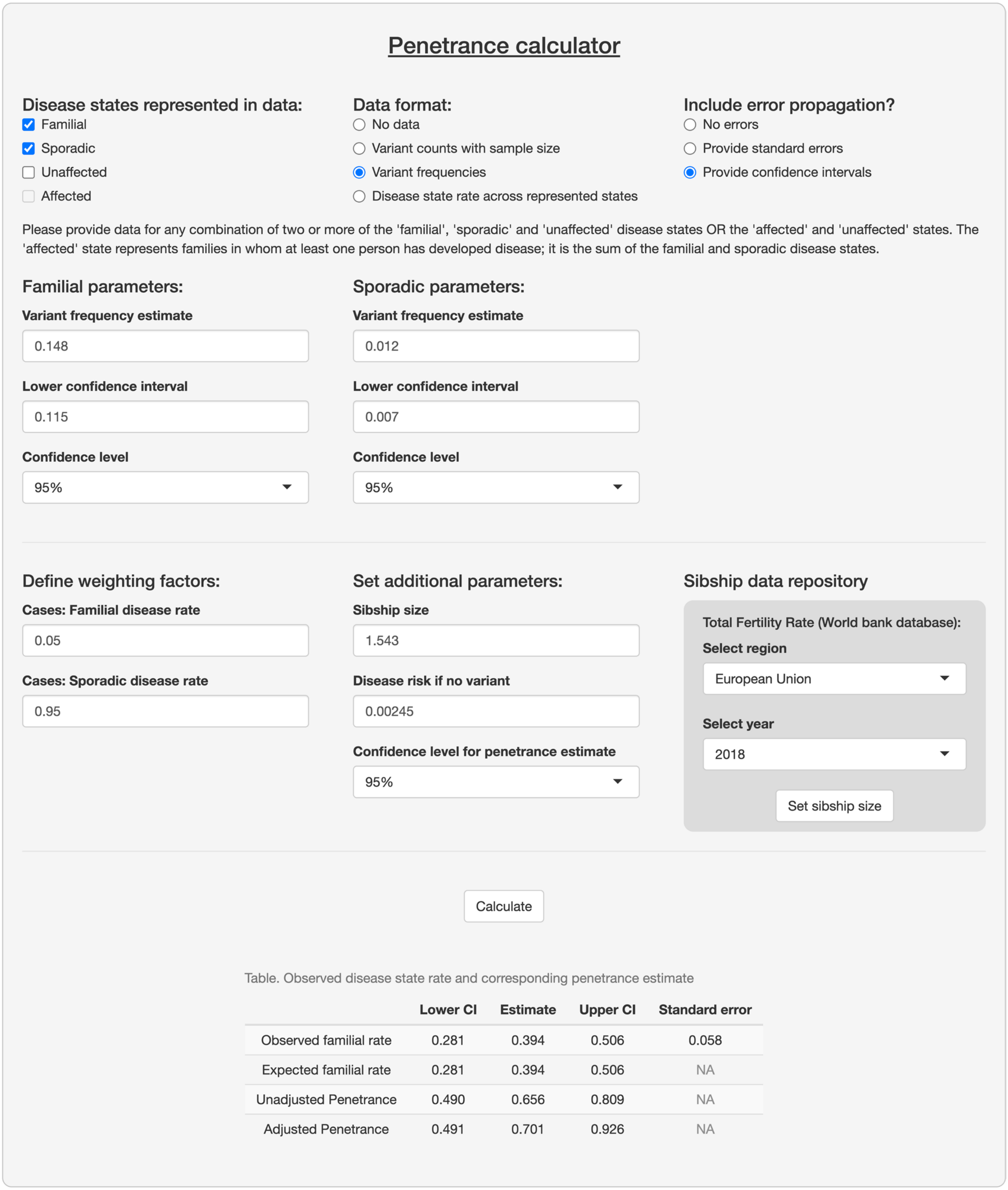

Step 1: Estimate $R(X)^{obs}$
Optional
Define variant frequencies and weighting factors
$M_{F,S,U,A}$
$W_{F,S,U,A}$
Define standard error for variant frequency estimates
$\sigma_{\overline{M_{F,S,U,A}}}$
Estimate rate of arbitrary state X in given data
$R(X)^{obs}$
Propagate error in $M_{F,S,U,A}$ to derive error in $R(X)^{obs}$
$R(X)^{obs} \pm E$
Step 2: Build lookup table
Calculate the rate of state X expected across represented states at N, g, and $f_i = 0, \ldots, 1$
$P(affected)_i$
$P(familial)_i$
$P(sporadic)_i$
$P(unaffected)_i$
Store corresponding $R(X)_i^{ex}$ and f_i values in lookup table
Step 3: Query lookup table
Query lookup table using estimate of the observed rate of state X
$R(X)^{obs}$
Optional
Repeat at error bounds of $R(X)^{obs}$
$R(X)^{obs} \pm E$
Match $R(X)^{obs}$ value to nearest $R(X)_i^{ex}$ and estimate f
$R(X)_i^{ex}$ f_i
$F(N,g,f)$ 0
$F(N,g,f)$ …
$F(N,g,f)$ 1
Step 4: Correct bias in penetrance estimate
Simulate families representative of input data
Frequency
0
Sibship size
k
Compute bias in penetrance estimates across true penetrance values in simulated families
Est. Error
0
0
Penetrance
1
Adjust penetrance estimate for real data by estimated error
$f \pm Est. Error$
Optional
Repeat for confidence limits of f

Fig. 1 Summary of the key steps within this penetrance estimation approach. Legend: Step 1: Variant frequencies (M) and weighting factors (W) are defined for a valid subset of the familial (F), sporadic (S), unaffected (U), and affected (A) states (see Table 1) to calculate rate of one of these states, arbitrarily labelled state X, among families harbouring the pathogenic variant across those states with data provided, $R(X)^{obs}$. Step 2: Eqs. (5–8) are applied to calculate $P(familial)$, $P(sporadic)$, $P(unaffected)$, and $P(affected)$, for a series of penetrance values, $f_i = 0, \ldots, 1$, at a defined sibship size, N, and with disease risk g for people not harbouring the variant. The rate of state X expected at each f_i among variant harbouring families from those states represented in Step 1, $R(X)_i^{ex}$, is calculated and stored alongside the corresponding f_i in a lookup table. Step 3: The lookup table is queried using $R(X)^{obs}$ to identify the closest $R(X)_i^{ex}$ value and corresponding f_i. Step 4: Bias in the obtained f_i estimate is corrected by simulating a population of families representative of the sample data, estimating the difference between true and estimated penetrance values in this population between $f = 0, \ldots, 1$and adjusting the estimated f_i by error predicted within a polynomial regression model fitted upon the simulated estimate errors. Optional step: Confidence intervals for $R(X)^{obs}$ can be calculated from error in the estimates of M provided [48]; Penetrance is estimated as in Steps 3 and 4 for the interval bounds. All steps within this approach are comprehensively detailed in Additional File: Sect. 1.1



Penetrance calculator

Disease states represented in data:
☑ Familial
☑ Sporadic
□ Unaffected
□ Affected

Data format:
○ No data
○ Variant counts with sample size
● Variant frequencies
○ Disease state rate across represented states

Include error propagation?
○ No errors
○ Provide standard errors
● Provide confidence intervals

Please provide data for any combination of two or more of the 'familial', 'sporadic' and 'unaffected' disease states OR the 'affected' and 'unaffected' states. The 'affected' state represents families in whom at least one person has developed disease; it is the sum of the familial and sporadic disease states.

Familial parameters:
Variant frequency estimate: 0.148
Lower confidence interval: 0.115
Confidence level: 95%

Sporadic parameters:
Variant frequency estimate: 0.012
Lower confidence interval: 0.007
Confidence level: 95%

Define weighting factors:
Cases: Familial disease rate: 0.05
Cases: Sporadic disease rate: 0.95

Set additional parameters:
Sibship size: 1.543
Disease risk if no variant: 0.00245
Confidence level for penetrance estimate: 95%

Sibship data repository
Total Fertility Rate (World bank database):
Select region: European Union
Select year: 2018

Table. Observed disease state rate and corresponding penetrance estimate

	Lower CI	Estimate	Upper CI	Standard error
Observed familial rate	0.281	0.394	0.506	0.058
Expected familial rate	0.281	0.394	0.506	NA
Unadjusted Penetrance	0.490	0.656	0.809	NA
Adjusted Penetrance	0.491	0.701	0.926	NA

Fig. 2 Example interface and output of the ADPenetrance web tool [20]. Legend: Here we show the example of penetrance of SOD1 variants for amyotrophic lateral sclerosis in a European population, applying variant frequency estimates for familial and sporadic ALS patients of European ancestry, an estimate of ALS disease risk among people not harbouring SOD1 variants, and the average Total Fertility Rate for the European Union in 2018 [22, 35]

Case 1: LRRK2 penetrance for PD

We estimated the penetrance of the *LRRK2* p.Gly2019Ser variant for PD. This case illustrates the flexibility of this method for application using data drawn from several combinations of the defined disease states.

The first-degree familiality rate of PD, about 0.105, was used to estimate $P(F|A)$ and $P(S|A)$ [23, 24]. $P(A)^{pop}$ was estimated as 1 in 37 (0.027), the lifetime risk of developing PD [25].

We estimated $M_{F,S}$ using data aggregated from 18 European ancestry groups within a sample of 24 world populations [26]. Of $3,770$ unrelated people with familial PD manifestations, 126 $(M_F = 0.033, \sigma_{\overline{M_X}} = 2.92 \times 10^{-3})$ harboured the *LRRK2* p.Gly2019Ser variant, compared to 130 of $10,898$ with sporadic PD $(M_S = 0.012, \sigma_{\overline{M_S}} = 1.04 \times 10^{-3})$.

As *LRRK2* p.Gly2019Ser occurred in only 2 members of the unaffected control sample, we estimated M_U using the larger European (non-Finnish) sample of the gnomAD v2.1.1 (controls) database [27], in which 10 of $21,383$ people harboured the variant $(M_U = 4.67 \times 10^{-4}, \sigma_{\overline{M_U}} = 1.47 \times 10^{-4})$.

We estimated that $g = 0.0267$, in accordance with Eqs. (9–11), based on the estimated $M_{F,S,U}$, $P(A)$, and $P(F|A)$.

As no single region is representative of the total sample, we estimated that $N = 1.572$ by aggregating Total Fertility Rate estimates available in the World Bank database [22] across each of the 18 European populations sampled, weighted by the proportional contribution of each population to the sample (see Additional File 1: Table S1) [26].

Additional region-specific and joint population penetrance estimates for this variant are presented in Additional File 1: Table S2.

Case 2: BMPR2 penetrance for heritable PAH

We estimated the penetrance of variants in the *BMPR2* gene for heritable PAH, a gene for which the low penetrance of pathogenic variants is well established [28].

Input parameters were defined based on only people with idiopathic (sporadic) or heritable PAH diagnoses [17]. This captures people with and without family disease history and excludes PAH manifestations associated with comorbidities or drug exposure.

We estimated $P(F|A)$ and $P(S|A)$ using the first-degree familiality rate of heritable PAH, about 0.055 of people affected by either idiopathic or familial PAH [28]. $P(A)^{pop}$ was estimated as 1 in 20 (0.05), according to an estimated 1 in 10 lifetime risk of developing any PAH, and that idiopathic and heritable PAH forms account for approximately 50% of PAH occurrences [28,29].

To minimise any study-specific bias, we applied data from two reports to build independent estimates for each of $M_{F,S}$. The first dataset [17], includes 247 people with familial PAH, of which 202 harboured *BMPR2* variants $(M_F = 0.818, \sigma_{\overline{M_F}} = 0.025)$, compared to 200 of 1174 in the sporadic state $(M_S = 0.170, \sigma_{\overline{M_S}} = 0.011)$. The second dataset [30] identified that 40 of 58 people with familial PAH $(M_F = 0.690, \sigma_{\overline{M_F}} = 0.061)$ harboured *BMPR2* variants, compared to 26 of 126 in the sporadic state $(M_S = 0.206, \sigma_{\overline{M_S}} = 0.036)$. Variant counts were additionally reported separately for small genetic variations (single nucleotide variants and indels) and structural variants in *BMPR2*, allowing penetrance estimation stratified by variant type. Letting $M_U = 0$, we estimated that $g = 0.0401$ for dataset 1, and $g = 0.0388$ for dataset 2, in accordance with Eqs. (9–11).

The first dataset may violate two assumptions of our approach: first, information on familial clustering was reportedly unavailable and so some families may be represented more than once in the familial state; second, it is not specified whether disease familiality is defined only by the disease status of first-degree relatives. The second sample overcomes a limitation of the first as each family is represented only once in variant counts. However, it is not reported whether disease states are defined according to the status of first-degree relatives only. As $R(X)^{obs}$ is calculated after weighting $M_{F,S}$ by the first-degree familial disease rate, the impact of some bias in variant frequency estimates upon penetrance estimates will be minimised.

The first cohort samples people from Asian, European, and North American populations; French, German and Italian cohorts comprise about 60% of the sample [17]. The second cohort samples people exclusively from Western Europe [30]. We therefore estimated that $N = 1.543$ in both instances, the Total Fertility Rate of the European Union in 2018 [22].

Cases 3 and 4: SOD1 and C9orf72 penetrance for ALS

We estimated the penetrance of variants in the *SOD1* and *C9orf72* genes for ALS. For *SOD1*, we examined the aggregated penetrance of *SOD1* variants harboured by people with ALS. For *C9orf72*, we examined the penetrance of a single pathogenic variant, a hexanucleotide GGGGCC repeat expansion (*C9orf72*$^{\mathrm{RE}}$; g.27573529_27573534GGCCCC[30<]). These penetrances have been historically difficult to establish without incurring kinship-specific biases. They represent ideal candidates for application of our method.

The first-degree familiality rate of ALS, about 0.050, was applied to define $P(F|A)$ and $P(S|A)$ in these cases [31, 32]. $P(A)^{pop}$ was estimated as 1 in 400 (0 .0025), the lifetime risk of developing ALS [33].

We drew upon the results of two meta-analyses [34, 35] to estimate $M_{F,S}$ for *SOD1* and *C9orf72*$^{\mathrm{RE}}$. As variant frequencies differed between Asian and European ancestries, we modelled penetrance separately for each group. We derived $\sigma_{\overline{M_{F,S}}}$ using z-score conversion from the 95% confidence intervals (95% CIs) reported: for the arbitrary state X,

$$\sigma_{\overline{M_X}} = \frac{M_X - M_X^{95\%lower}}{z} \tag{12}$$

where $z = 1.96$ and $M_X^{95\%lower}$ is the lower 95% CI bound of the estimate M_X.

In Asian ALS populations: *SOD1* variants were harboured by 0.300 ($\sigma_{\overline{M_F}} = 0.025$) of people with familial and 0.015 ($\sigma_{\overline{M_S}} = 2.55 \times 10^{-3}$) with sporadic disease; *C9orf72*$^{\mathrm{RE}}$ was harboured by 0.04 ($\sigma_{\overline{M_F}} = 0.010$) of people with familial and 0.01 ($\sigma_{\overline{M_S}} = 5.10 \times 10^{-3}$) with sporadic disease. In accordance with Eqs. (9–11), and letting $M_U = 0$, we estimated that $g = 0.00243$ for *SOD1*, and $g = 0.00247$ for *C9orf72*.

In Europeans, *SOD1* variants were harboured by 0.148 ($\sigma_{\overline{M_F}} = 0.017$) of people with familial and 0.012 ($\sigma_{\overline{M_S}} = 2.55 \times 10^{-3}$) with sporadic disease; *C9orf72*$^{\mathrm{RE}}$ was harboured by 0.32 ($\sigma_{\overline{M_F}} = 0.020$) of people with familial and 0.05 ($\sigma_{\overline{M_S}} = 5.10 \times 10^{-3}$) with sporadic disease. In accordance with Eqs. (9–11), and letting $M_U = 0$, we estimated that $g = 0.00245$ for *SOD1*, and $g = 0.00234$ for *C9orf72*.

The *SOD1* meta-analysis allowed consideration of the extended kinship when defining familial ALS. The familiality definition used in the *C9orf72* analysis is not stated. As before, the weighting of $M_{f,s}$ by the first-degree familial disease rate when calculating $R(X)^{obs}$ will minimise any impact of some bias in variant frequencies upon penetrance estimates.

We tested for equal onset variability (see Additional File 1: Sect. 1.2.2) with the *checkOnsetVariability* R function provided in the associated GitHub repository [19], comparing variability in age of ALS onset for people with *SOD1* or *C9orf72* variants to that of people without variants in these genes. The results (see Additional File 1: Fig. S4) suggested approximately equal onset variability between the *SOD1* and no (*SOD1* or *C9orf72*) variant groups, indicated by visual inspection of the cumulative density plot provided and by an approximately equal time spanned between the first and third quartiles of age of onset across the groups. Onset variability appears more unequal in the *C9orf72* case study, with a ~ 1.36 times shorter interquartile interval for people harbouring *C9orf72*^{RE} than the no variant cohort. One of the simulation studies presented in Additional File 1: Sect. 1.2.3; Fig. S11 models a comparable departure from the equal onset variability and demonstrates that a small but tolerable inflation of penetrance estimates may occur if sampling a younger cohort. Since the present penetrance estimates are based on pooled variant frequency estimates from large meta-analyses of variant frequencies in these genes, the present degree of unequal onset variability is unlikely to have impacted penetrance estimation.

In these datasets, the Asian ancestry cohorts were predominantly individuals from East Asia, with small proportion from South Asia. The European ancestry cohorts primarily comprise people from European countries, with some from North America and Australasia. Accordingly, N was estimated for the Asian population samples as 1.823, the Total Fertility Rate for East Asia and Pacific in 2018, and for the European population as 1.543, the Total Fertility Rate for the European Union in 2018 [22].

Results

Here we summarise the input data and results of the case studies modelled (see Table 2). Penetrance estimates are presented both when accounting for residual disease risk g among people with no variant and when assuming that $g = 0$; those accounting for g are preferred.

Estimated penetrance of the *LRRK2* p.Gly2019Ser variant for PD was roughly consistent across the modelled disease state combinations. Additional penetrance estimates across various populations within the dataset from which this European sample was drawn are presented in Additional File 1: Table S2.

The penetrance of *BMPR2* variants for PAH was estimated comparably across the two sample sets, although slightly higher within dataset one [17] than two [30]. Penetrance was also comparable between the defined *BMPR2* variant subtypes of the second sample. Differences in these estimates reflect variation in $M_{F,S}$ between the cohorts and may result from a different admix of variants between samples, or unspecified family clustering within the first sample set. It is not known for either dataset whether family history classifications were restricted to first-degree relatives only and so the estimates obtained may be slightly inflated. We note, however, that impact of any inflation was minimised because variant frequency weighting factors were correctly defined. With the available data, these possibilities cannot be explored further.

Penetrance estimates of *SOD1* and *C9orf72* variants for ALS demonstrate consistency within genes across populations and indicate that the penetrance for ALS is greater in people harbouring *SOD1* variants than in those harbouring *C9orf72*$^{\mathrm{RE}}$. Additional File 1: Table S3 presents additional penetrance estimates for several widely-described *SOD1* variants: penetrance was estimated as 1 for p.Ala5Val (c.14C > T), 0.644 for p.Ile114Thr (c.341 T > C), and 0 for p.Asp91Ala (c.272A > C). Each estimate made in these case studies may be slightly inflated owing to inclusion of extended kinship within familiality definition. However, as before, accurately

Table 2 Penetrance estimation for the present case studies

Case study [a]	Data subset	Variant frequency in state (standard error)			Average sibship size	Residual disease risk [a]	States modelled [c]	Familial disease rate for people with variant across states modelled (95% CI)	Penetrance (95% CI)[e]	
		Familial	Sporadic	Unaffected					Assuming no residual disease risk	Accounting for residual disease risk
-	-	$M_F(\sigma_{\overline{M_F}})$	$M_S(\sigma_{\overline{M_S}})$	$M_U(\sigma_{\overline{M_U}})$	N	g	-	$R(X)$	$f(g=0)$	$f(g=g)$
LRRK2 p.G2019S for PD [26, 27]	European ancestry	0.033 (2.92×10^{-3})	0.012 (1.04×10^{-3})	4.67×10^{-4} (1.47×10^{-4})	$1.572^{b.1}$	0.0267	F, S, U	0.113 (0.071, 0.155)	0.37 (0.285, 0.443)	0.334 (0.249, 0.408)
							F, S	0.247 (0.202, 0.292)	0.429 (0.348, 0.509)	0.379 (0.299, 0.461)
							F, U	0.172 (0.081, 0.264)	0.35 (0.247, 0.428)	0.32 (0.215, 0.399)
							S, U	0.388 (0.235, 0.541)[d]	0.293 (0.161, 0.45)	0.275 (0.138, 0.438)
BMPR2 variants for PAH	All variants [17]	0.818 (0.025)	0.170 (0.011)	-	$1.543^{b.2}$	0.0401	F, S	0.218 (0.195, 0.242)	0.382 (0.339, 0.426)	0.308 (0.266, 0.351)
	All variants [30]	0.690 (0.061)	0.206 (0.036)	-	$1.543^{b.2}$	0.0388	F, S	0.163 (0.111, 0.215)	0.281 (0.186, 0.376)	0.212 (0.12, 0.305)
	SNVs and indels [30]	0.569 (0.065)	0.159 (0.033)	-	$1.543^{b.2}$	0.0413	F, S	0.173 (0.107, 0.238)	0.299 (0.179, 0.419)	0.225 (0.11, 0.342)
	Structural variants [30]	0.121 (0.043)	0.048 (0.019)	-	$1.543^{b.2}$	0.0475	F, S	0.129 (0.011, 0.246)	0.218 (0.014, 0.432)	0.138 (0, 0.345)
SOD1 variants for ALS [35]	Asian	0.300 (0.025)	0.015 (2.55×10^{-3})	-	$1.823^{b.3}$	0.00243	F, S	0.513 (0.420, 0.606)	0.829 (0.665, 1)	0.826 (0.661, 1)
	European	0.148 (0.017)	0.012 (2.55×10^{-3})	-	$1.543^{b.2}$	0.00245	F, S	0.394 (0.281, 0.506)	0.705 (0.496, 0.933)	0.701 (0.491, 0.926)
C9orf72^{RE} for ALS [34]	Asian	0.04 (0.010)	0.01 (5.10×10^{-3})	-	$1.823^{b.3}$	0.00247	F, S	0.174 (0.013, 0.335)	0.263 (0.0156, 0.522)	0.258 (0.0108, 0.518)
	European	0.32 (0.020)	0.05 (5.10×10^{-3})	-	$1.543^{b.2}$	0.00234	F, S	0.252 (0.208, 0.296)	0.443 (0.363, 0.524)	0.439 (0.358, 0.52)

[a] Disease characteristics of lifetime disease risk ($P(A)_{pop}$) and proportion familial ($P(F|A)$(F|A)) are used as weighting factors (per Table 1) and for calculating g (per Eqs. (9–11), letting $M_U = 0$ in the ALS and PAH case studies), and are defined as follows: in PD, $P(A)^{pop} = 0.027, P(F|A) = 0.105$; in PAH, $P(A)^{pop} = 0.05, P(F|A) = 0.055$; in ALS, $P(A)^{pop} = 0.0025, P(F|A) = 0.050$. [b]Estimated using Total Fertility Rates reported for the: populations sampled to calculate variant frequencies (see Additional File 1: Table S1) [b.1], European Union[b.2], or East Asia and Pacific[b.3] regions in 2018 [22]; [c]F = familial, S = sporadic, U = unaffected (controls); [d]Rate of sporadic disease has been calculated here because the familial state is not represented; [e]Step 4 penetrance estimates are presented, see Additional File 1: Table S5 for unadjusted penetrance estimates derived in Step 3. *PD*, Parkinson's disease; *PAH*, pulmonary arterial hypertension; *ALS*, amyotrophic lateral sclerosis; *C9orf72*^{RE}, the pathogenic C9orf72 GGGGCC hexanucleotide repeat expansion; *SNV*, single nucleotide variant; *indel*, small insertions or deletions; *CI*, confidence interval

defined weighting factors will have minimised this impact.

Discussion

We have developed a novel approach to estimate the penetrance of genetic variants pathogenic for autosomal dominant traits. The method was tested via simulation studies (see Additional File: Sect. 1.2.3) and application to several case studies.

Our penetrance estimates of the *LRRK2* p.Gly2019Ser variant for PD and of the aggregate penetrance of *BMPR2* variants for PAH closely matched those obtained using other approaches. Previous research on *LRRK2* p.Gly2019Ser estimates its lifetime penetrance between 0.24 (95% CI: 0.135, 0.437) and 0.45 (no CI reported) when analysing data that is not liable to inflation owing to biased selection of familial cases [2]. Longitudinal analysis of disease trends among 53 families harbouring *BMPR2* variants finds penetrance as 0.27 overall, 0.42 for women and 0.14 for men [36]. These case studies additionally demonstrated the importance of considering residual disease risk g for family members not harbouring the variant when estimating penetrance in more common traits. This importance is explored further within simulation studies (see Additional File 1: Sect. 1.2.3; Fig. S8).

The estimates in the *SOD1* and *C9orf72* case studies align with current understanding of penetrance in these ALS genes.

For *SOD1* variants, penetrance for ALS is incomplete and differs between variants [10, 37]. The widely-described p.Ala5Val (formally p.Ala4Val) variant has a recorded penetrance of 0.91 by 70 years of age [38]. Among other variants, penetrance is typically lower [10, 37]. Of those best characterised, p.Ile114Thr approaches complete penetrance in some pedigrees and p.Asp91Ala reaches polymorphic frequency in some populations, with linked ALS presentations typically displaying an autosomal recessive pattern [10, 13, 38]. Our estimates for heterozygous inheritance of these individual variants aligned with these observations (see Additional File 1: Table S3) and highlight the spectrum of penetrance across variants in *SOD1*. Our estimate for the p.Asp91Ala variant in particular supports the hypothesis that it is associated with ALS via a recessive or oligogenic inheritance pattern [39]. The absence of p.Asp91Ala within the familial ALS database sampled further corroborates the finding. Accordingly, our penetrance estimates in Asian and European populations can be taken to suitably represent an aggregated penetrance of risk variants in *SOD1* for ALS; some variation between populations can be expected, reflecting differences in the admix of variants between them.

For *C9orf72*, we modelled the penetrance of its pathogenic hexanucleotide repeat expansion for ALS. Pleiotropy of this variant is widely reported, additionally conferring risk for frontotemporal dementia and, to a lesser degree, other neuropsychiatric conditions [40]. Past penetrance estimates made for this variant are vulnerable to inflation from biased ascertainment of affected people, and the variant is more common among unaffected people than would be expected if these estimates were accurate [12, 40, 41]. A previous study tentatively reports the penetrance of *C9orf72*^{RE} for either ALS or frontotemporal dementia as 0.90 by age 83 after attempting to adjust for ascertainment bias within their sample [41]. Accounting for lifetime risk of each phenotype and their respective familiality rates, people of European ancestry harbouring *C9orf72*$^{\mathrm{RE}}$ appear to develop ALS or frontotemporal dementia with comparable frequency, we calculated that 1.012 cases of ALS emerge per case of frontotemporal dementia (see Additional File 1: Table S4) [33, 34, 42–44]. It is therefore reasonable to predict that, if the variant has 0.90 penetrance for the joint condition of ALS and frontotemporal dementia, its penetrance of for ALS alone would be around 0.45. The 0.45 estimate is comparable to the upper bound of our findings. However, we note that our calculation does not account for the common co-occurrence of ALS and frontotemporal dementia and that the true penetrance of *C9orf72*$^{\mathrm{RE}}$ for the joint ALS-frontotemporal dementia condition is likely lower than the tentative 0.90 estimate.

The method presented has high validity. Internal validity is demonstrated within simulation studies (see Additional File: Sect. 1.2.3). Criterion and face validity are shown across the present case studies, aligning with estimates made using other techniques and current understanding of the assessed cases. Construct validity is also demonstrated: in the ALS case studies, disease risk was greater for those harbouring a pathogenic *SOD1* variant than for those with the *C9orf72* repeat expansion. This aligns with the multi-step model of ALS [45], where harbouring *SOD1* variants is associated with a 2-step disease process, converse to the 3-step process associated with *C9orf72*$^{\mathrm{RE}}$.

The data necessary to operate our approach is distinct from other techniques which examine patterns of disease among affected people, allowing assessment of penetrance in unrelated populations rather than families. The estimates are therefore unaffected by kinship-specific modifiers and are instead applicable to the sampled population. Since penetrance may vary according to genetic background, ancestry-specific penetrance estimates are best obtained by stratifying input data according to ancestral groups; this approach is demonstrated in the PD and ALS case studies (see Table 2, Additional File 1: Table S2).

Where analysis is confined to people affected by disease, across the familial and sporadic states, we circumvent the ascertainment biases affecting designs which examine the distribution of a variant between affected and unaffected populations [9]. Where analysis includes data for unaffected samples (i.e. controls harbouring the variant) these would not be avoided; ascertainment of controls compared to cases has equivalent challenges irrespective of the penetrance estimation approach. However, as our method does not require this information if data of familial and sporadic cases are available, this does not majorly limit the approach.

Furthermore, limitations of ascertainment will diminish as huge datasets of genetic and phenotypic information available within public databases become increasingly available. Therefore, the usefulness of penetrance estimates generated through population data will grow alongside the increasing size and scope of genetic data within such datasets [9].

A limitation of this approach is the definition of familiality, which is the occurrence of the studied trait in a first-degree relative under this model. In practice, familial disease may be defined using various criteria, for example considering the disease status of second- or third-degree relatives, or including related diseases that may share a genetic basis [32, 46]. For example, ALS and frontotemporal dementia each share a genetic basis, and considering a family history of frontotemporal dementia is reasonable when assessing familiality in a person with ALS. If the extended kinship is incorporated within familial disease state definitions, then the familial rate will trend upwards and inflate penetrance estimates. Using a wider definition of being affected is acceptable, although it will yield penetrance estimates for the joint condition.

A further caveat is that the model equations assume a particular family structure. It is not feasible to include all possible family configurations for large quantities of summary data however and approximations made are sufficiently close to provide an estimate of penetrance.

This method is suitable for calculating the point, rather than age-dependent, penetrance of pathogenic variants and can be applied to any germline genetic variation associated with a disease via an autosomal dominant inheritance pattern. Penetrance can be derived for an individual variant or for an aggregated set of variants, with the latter indicating an averaged burden of variants meeting the given criteria. We suggest that confidence intervals should be included when using this approach; the size of the interval returned will provide a useful indication of whether the data provided are sufficient for precise penetrance estimation.

When samples include only people harbouring the variant, the method assumes the stability of disease states among sampled families over time. This assumption is typical in case–control research designs, which expect that members of the control sample will not later become cases. However, in traits with age-dependent penetrance, estimates would be influenced by the age at time of sampling. Younger samples would yield reduced estimates if fewer than two family members are affected when sampled and others will only later become affected. A lifetime penetrance estimate would therefore be best obtained within this sampling scenario if sampling people beyond the typical age of onset for the studied disease.

Reasonable lifetime penetrance estimates can however be obtained at earlier sampling times even in circumstances of age-dependent onset when disease state rates are calculated via weighted proportions of variant frequency estimates within those states sampled. This sampling approach was applied in each of the 4 case studies, and follows an assumption that family disease states change comparably over time for people with and without the variant (see Additional File 1: Sect. 1.2.2). Simulation studies demonstrated this relative stability across ages and a relative tolerance to an unequal variability in age of onset profiles between variant and non-variant groups (see Additional File 1: Sect. 1.2.3; Fig. S10-S12).

Point penetrance estimates have several applications, for instance, improving the characterisation of pathogenic variants at a population level, facilitating research involving tested variants and, in particular, aiding clinical trial design by supporting the curation of homogenous study populations. They would have additional utility once gene therapies move towards preventative treatment, giving justification for or against such treatment after accounting for possible side effects and risks.

In a scenario where penetrance can be estimated via multiple approaches, we recommend applying each applicable method, given the complimentary nature of these techniques. If the results of multiple approaches conflict, we would suggest inspection of the suitability of the input data given for each method and to prioritise the result which these best fit.

Conclusions

Our novel method for penetrance estimation fills an important gap in medical genetics because, utilising population-scale data, it enables the unbiased and valid calculation of penetrance in genetic disease instances that would be otherwise difficult or impossible using existing methods. It serves to expand the range of genetic diseases and variants for which high-quality penetrance estimates can be obtained, as we illustrate in the ALS case studies. Estimates drawn via this approach have clear utility and will be useful for characterisation of pathogenic variants, with benefits for both clinical practice and research. They have wider relevance to the population than those obtained by studying particular kinships and will be more interpretable for clinical professionals.

The tool code is available as an R function on GitHub [19] and the method is available and free to use via a public webserver [20].

Abbreviations

ALS: Amyotrophic lateral sclerosis; CI: Confidence interval; $M_{F,S,U,A}$: Variant frequencies in the familial, sporadic, unaffected, and affected states; PAH: Pulmonary arterial hypertension; PD: Parkinson's disease; P(A): Probability of being affected by a given disease; P(F|A): Disease familiality rate; P(S|A): Disease sporadic rate; P(M): Probability of harbouring the variant M; $W_{F,S,U,A}$: Weighting factors for the familial, sporadic, unaffected, and affected states; $R(X)^{obs}$: Observed rate of arbitrary state x.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-022-01142-7.

Additional file 1. This file contains all supplementary methods, figures, and tables referred to throughout this manuscript.

Acknowledgements

We acknowledge the use of the research computing facility at King's College London, Rosalind (https://rosalind.kcl.ac.uk), which is delivered in partnership with the National Institute for Health Research (NIHR) Biomedical Research Centres at South London & Maudsley and Guy's & St. Thomas' NHS Foundation Trusts and part-funded by capital equipment grants from the Maudsley Charity (award 980) and Guy's and St Thomas' Charity (TR130505). The authors also acknowledge the use of the CREATE research computing facility at King's College London [47].

Authors' contributions

TPS: conceptualisation, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft, writing – review and editing, visualisation. SOM: methodology, software, writing – review and editing, supervision. HB: methodology, writing – review and editing. CML: methodology, writing – review and editing, supervision. AI: conceptualisation, methodology, writing – review and editing, supervision. AAC: conceptualisation, methodology, writing – review and editing, supervision.
All authors read and approved the final manuscript.

Funding

This is an EU Joint Programme-Neurodegenerative Disease Research (JPND) project. The project is supported through the following funding organisations under the aegis of JPND-http://www.neurodegenerationresearch.eu/ (United Kingdom, Medical Research Council MR/L501529/1 to A.A.-C., principal investigator [PI] and MR/R024804/1 to A.A.-C., PI]; Economic and Social Research Council ES/L008238/1 to A.A.-C. [co-PI]) and through the Motor Neurone Disease Association. This study represents independent research part funded by the NIHR Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The work leading up to this publication was funded by the European Community's Horizon 2020 Programme (H2020-PHC-2014-two-stage; grant 633413). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, King's College London, or the Department of Health and Social Care. A.I. is funded by South London and Maudsley NHS Foundation Trust, MND Scotland, Motor Neurone Disease Association, National Institute for Health and Care Research, Spastic Paraplegia Foundation, Rosetrees Trust, Darby Rimmer MND Foundation, the Medical Research Council (UKRI) and Alzheimer's Research UK.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in our GitHub repository [19]: https://github.com/ThomasPSpargo/adpenetrance. This includes R scripts used for approach validation as described in this manuscript and Additional File: Sect. 1.2. The datasets supporting the conclusions of this article are included within the article (and its additional file).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

CML sits on the SAB for Myriad Neuroscience. AAC reports consultancies or advisory boards for Amylyx, Apellis, Biogen Idec, Brainstorm, Cytokinetics, GSK, Lilly, Mitsubishi Tanabe Pharma, Novartis, OrionPharma, Quralis, and Wave Pharmaceuticals.
The remaining authors declare that they have no competing interests.

Author details

[1]Department of Basic and Clinical Neuroscience, Maurice Wohl Clinical Neuroscience Institute, King's College London, London SE5 9RX, UK. [2]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, de Crespigny Park, London SE5 8AF, UK. [3]Department of Medical and Molecular Genetics, Faculty of Life Sciences and Medicine, King's College London, London, UK. [4]Department of Biostatistics and Health Informatics, King's College London, London, UK. [5]NIHR Maudsley Biomedical Research Centre (BRC) at South London and Maudsley NHS Foundation Trust and King's College London, London, UK. [6]King's College Hospital, Bessemer Road, London SE5 9RS, UK.

Received: 13 June 2022 Accepted: 18 November 2022
Published online: 15 December 2022

References

1. Langbehn DR, Brinkman RR, Falush D, Paulsen JS, Hayden MR. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. Clin Genet. 2004;65(4):267–77. https://doi.org/10.1111/j.1399-0004.2004.00241.x.
2. Goldwurm S, Tunesi S, Tesei S, et al. Kin-cohort analysis of *LRRK2*-G2019S penetrance in Parkinson's disease. Mov Disord. 2011;26(11):2144–5. https://doi.org/10.1002/mds.23807.
3. Dewey FE, Grove ME, Pan C, et al. Clinical Interpretation and Implications of Whole-Genome Sequencing. JAMA. 2014;311(10):1035–45. https://doi.org/10.1001/jama.2014.1717.
4. Kalia SS, Adelman K, Bale SJ, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet Med. 2017;19(2):249–55. https://doi.org/10.1038/gim.2016.190.
5. Saelaert M, Mertes H, Moerenhout T, De Baere E, Devisch I. Criteria for reporting incidental findings in clinical exome sequencing – a focus group study on professional practices and perspectives in Belgian genetic centres. BMC Med Genomics. 2019;12(1):123. https://doi.org/10.1186/s12920-019-0561-0.
6. Senol-Cosar O, Schmidt RJ, Qian E, et al. Considerations for clinical curation, classification, and reporting of low-penetrance and low effect size variants associated with disease risk. Genet Med. 2019;21(12):2765–73. https://doi.org/10.1038/s41436-019-0560-8.

7. Otto PA, Horimoto ARVR. Penetrance rate estimation in autosomal dominant conditions. Genet Mol Biol. 2012;35(3):583–8. https://doi.org/10.1590/S1415-47572012005000051.
8. Minikel EV, Vallabh SM, Lek M, et al. Quantifying prion disease penetrance using large population control cohorts. Sci Transl Med. 2016;8(322):322ra9. https://doi.org/10.1126/scitranslmed.aad5169.
9. Wright CF, West B, Tuke M, et al. Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting. Am J Hum Genet. 2019;104(2):275–86. https://doi.org/10.1016/j.ajhg.2018.12.015.
10. Chiò A, Battistini S, Calvo A, et al. Genetic counselling in ALS: facts, uncertainties and clinical suggestions. J Neurol Neurosurg Psychiatry. 2014;85(5):478. https://doi.org/10.1136/jnnp-2013-305546.
11. Shatunov A, Al-Chalabi A. The genetic architecture of ALS. Neurobiol Dis. 2021;147:105156. https://doi.org/10.1016/j.nbd.2020.105156.
12. Iacoangeli A, Al Khleifat A, Jones AR, et al. *C9orf72* intermediate expansions of 24–30 repeats are associated with ALS. Acta Neuropathol Commun. 2019;7:115. https://doi.org/10.1186/s40478-019-0724-4.
13. Iacoangeli A, Al Khleifat A, Sproviero W, et al. ALSgeneScanner: a pipeline for the analysis and interpretation of DNA sequencing data of ALS patients. Amyotroph Lateral Scler Frontotemporal Degener. 2019;20:207–15. https://doi.org/10.1080/21678421.2018.1562553.
14. Abel O, Powell JF, Andersen PM, Al-Chalabi A. ALSoD: A user-friendly online bioinformatics tool for amyotrophic lateral sclerosis genetics. Hum Mutat. 2012;33(9):1345–51. https://doi.org/10.1002/humu.22157.
15. Opie-Martin S, Iacoangeli A, Topp SD, et al. The *SOD1*-mediated ALS phenotype shows a decoupling between age of symptom onset and disease duration. Nat Commun. 2022;13(1):6901. https://doi.org/10.1038/s41467-022-34620-y.
16. Al-Chalabi A, Lewis CM. Modelling the Effects of Penetrance and Family Size on Rates of Sporadic and Familial Disease. Hum Hered. 2011;71(4):281–8. https://doi.org/10.1159/000330167.
17. Evans JDW, Girerd B, Montani D, et al. *BMPR2* mutations and survival in pulmonary arterial hypertension: an individual participant data meta-analysis. Lancet Respir Med. 2016;4(2):129–37. https://doi.org/10.1016/S2213-2600(15)00544-5.
18. Hunink MGM, Weinstein MC, Wittenberg E, et al. Decision Making in Health and Medicine: Integrating Evidence and Values. 2nd ed. Cambridge: Cambridge University Press; 2014.
19. Spargo TP, Opie-Martin S, Bowles H, Lewis CM, Iacoangeli A, Al-Chalabi A. ADPenetrance. 2022. GitHub. Available from: https://github.com/ThomasPSpargo/adpenetrance.
20. Spargo TP, Opie-Martin S, Bowles H, Lewis CM, Iacoangeli A, Al-Chalabi A. ADPenetrance: penetrance estimation for autosomal dominant traits. 2022. Available from: https://adpenetrance.rosalind.kcl.ac.uk/.
21. Chang W, Cheng J, Allaire J, et al. shiny: Web Application Framework for R. 2022. R package version 1.7.3. Available from: https://CRAN.R-project.org/package=shiny.
22. Fertility rate, total (births per woman) [Internet]. 2020. Available from: https://databank.worldbank.org/reports.aspx?source=2&series=SP.DYN.TFRT.IN. Accessed 12 Jan 2021.
23. Shino MY, McGuire V, Van Den Eeden SK, et al. Familial aggregation of Parkinson's disease in a multiethnic community-based case-control study. Mov Disord. 2010;25(15):2587–94. https://doi.org/10.1002/mds.23361.
24. Elbaz A, Grigoletto F, Baldereschi M, et al. Familial aggregation of Parkinson's disease. Neurology. 1999;52(9):1876. https://doi.org/10.1212/WNL.52.9.1876.
25. Parkinson's UK. The Incidence and Prevalence of Parkinson's in the UK: Results from the Clinical Practice Research Datalink Reference Report. 2017. Available from: https://www.parkinsons.org.uk/professionals/resources/incidence-and-prevalence-parkinsons-uk-report.
26. Healy DG, Falchi M, O'Sullivan SS, et al. Phenotype, genotype, and worldwide genetic penetrance of *LRRK2*-associated Parkinson's disease: a case-control study. Lancet Neurol. 2008;7(7):583–90. https://doi.org/10.1016/S1474-4422(08)70117-0.
27. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43. https://doi.org/10.1038/s41586-020-2308-7.
28. Thenappan T, Ryan JJ, Archer SL. Evolving epidemiology of pulmonary arterial hypertension. Am J Respir Crit Care Med. 2012;186(8):707–9. https://doi.org/10.1164/rccm.201207-1266ED.
29. Corris PA, Seeger W. Call it by the correct name—pulmonary hypertension not pulmonary arterial hypertension: growing recognition of the global health impact for a well-recognized condition and the role of the Pulmonary Vascular Research Institute. Am J Physiol Lung Cell Mol. 2020;318(5):L992–4. https://doi.org/10.1152/ajplung.00098.2020.
30. Aldred MA, Vijayakrishnan J, James V, et al. *BMPR2* gene rearrangements account for a significant proportion of mutations in familial and idiopathic pulmonary arterial hypertension. Hum Mutat. 2006;27(2):212–3. https://doi.org/10.1002/humu.9398.
31. Byrne S, Walsh C, Lynch C, et al. Rate of familial amyotrophic lateral sclerosis: a systematic review and meta-analysis. J Neurol Neurosurg Psychiatry. 2011;82(6):623–7. https://doi.org/10.1136/jnnp.2010.224501.
32. Byrne S, Heverin M, Elamin M, et al. Aggregation of neurologic and neuropsychiatric disease in amyotrophic lateral sclerosis kindreds: A population-based case–control cohort study of familial and sporadic amyotrophic lateral sclerosis. Ann Neurol. 2013;74(5):699–708. https://doi.org/10.1002/ana.23969.
33. Alonso A, Logroscino G, Jick SS, Hernán MA. Incidence and lifetime risk of motor neuron disease in the United Kingdom: a population-based study. Eur J Neurol. 2009;16(6):745–51. https://doi.org/10.1111/j.1468-1331.2009.02586.x.
34. Marogianni C, Rikos D, Provatas A, et al. The role of C9orf72 in neurodegenerative disorders: a systematic review, an updated meta-analysis, and the creation of an online database. Neurobiol Aging. 2019;84:1.e-e10. https://doi.org/10.1016/j.neurobiolaging.2019.04.012.
35. Zou Z-Y, Zhou Z-R, Che C-H, Liu C-Y, He R-L, Huang H-P. Genetic epidemiology of amyotrophic lateral sclerosis: a systematic review and meta-analysis. J Neurol Neurosurg Psychiatry. 2017;88:540–9. https://doi.org/10.1136/jnnp-2016-315018.
36. Larkin EK, Newman JH, Austin ED, et al. Longitudinal Analysis Casts Doubt on the Presence of Genetic Anticipation in Heritable Pulmonary Arterial Hypertension. Am J Respir Crit Care Med. 2012;186(9):892–6. https://doi.org/10.1164/rccm.201205-0886OC.
37. Andersen PM. Amyotrophic lateral sclerosis associated with mutations in the CuZn superoxide dismutase gene. Curr Neurol Neurosci Rep. 2006;6(1):37–46. https://doi.org/10.1007/s11910-996-0008-9.
38. Cudkowicz ME, McKenna-Yasek D, Sapp PE, et al. Epidemiology of mutations in superoxide dismutase in amyotrophic lateral sclerosis. Ann Neurol. 1997;41(2):210–21. https://doi.org/10.1002/ana.410410212.
39. van Blitterswijk M, van Es MA, Hennekam EAM, et al. Evidence for an oligogenic basis of amyotrophic lateral sclerosis. Hum Mol Genet. 2012;21(17):3776–84. https://doi.org/10.1093/hmg/dds199.
40. Beck J, Poulter M, Hensman D, et al. Large *C9orf72* hexanucleotide repeat expansions are seen in multiple neurodegenerative syndromes and are more frequent than expected in the UK population. Am J Hum Genet. 2013;92(3):345–53. https://doi.org/10.1016/j.ajhg.2013.01.011.
41. Murphy NA, Arthur KC, Tienari PJ, Houlden H, Chio A, Traynor BJ. Age-related penetrance of the *C9orf72* repeat expansion. Sci Rep. 2017;7(1):2116. https://doi.org/10.1038/s41598-017-02364-1.
42. Coyle-Gilchrist ITS, Dick KM, Patterson K, et al. Prevalence, characteristics, and survival of frontotemporal lobar degeneration syndromes. Neurology. 2016;86(18):1736. https://doi.org/10.1212/WNL.0000000000002638.
43. Turner MR, Al-Chalabi A, Chio A, et al. Genetic screening in sporadic ALS and FTD. J Neurol Neurosurg Psychiatry. 2017;88(12):1042–4. https://doi.org/10.1136/jnnp-2017-315995.
44. Majounie E, Renton AE, Mok K, et al. Frequency of the *C9orf72* hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. Lancet Neurol. 2012;11(4):323–30. https://doi.org/10.1016/s1474-4422(12)70043-1.
45. Chiò A, Mazzini L, Alfonso S, et al. The multistep hypothesis of ALS revisited. Neurology. 2018;91(7):e635. https://doi.org/10.1212/WNL.0000000000005996.
46. Vajda A, McLaughlin RL, Heverin M, et al. Genetic testing in ALS: A survey of current practices. Neurol. 2017;88(10):991–9. https://doi.org/10.1212/wnl.0000000000003686.
47. King's College London. King's Computational Research, Engineering and Technology Environment (CREATE) 2022 [October 12 2022]. Available from: https://doi.org/10.18742/rnvf-m076.
48. Hughes I, Hase T. Measurements and their uncertainties: a practical guide to modern error analysis. Oxford: Oxford University Press; 2010.

Publisher's note