This electronic thesis or dissertation has been downloaded from the King's Research Portal at https://kclpure.kcl.ac.uk/portal/

The relationship between genotype and phenotype in amyotrophic lateral sclerosis

Spargo, Thomas

Awarding institution:
King's College London

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without proper acknowledgement.

END USER LICENCE AGREEMENT

Unless another licence is stated on the immediately following page this work is licensed

under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

licence. https://creativecommons.org/licenses/by-nc-nd/4.0/

You are free to copy, distribute and transmit the work

Under the following conditions:

- Attribution: You must attribute the work in the manner specified by the author (but not in any way that suggests that they endorse you or your use of the work).
- Non Commercial: You may not use this work for commercial purposes.
- No Derivative Works - You may not alter, transform, or build upon this work.

Any of these conditions can be waived if you receive permission from the author. Your fair dealings and other rights are in no way affected by the above.

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 08. Oct. 2023

The relationship between genotype and phenotype in amyotrophic lateral sclerosis

Thomas Paul Spargo

Thesis submitted for the degree of Ph.D.

Supervisors:

Professor Ammar Al-Chalabi

Dr Alfredo Iacoangeli

Institute of Psychiatry, Psychology and Neuroscience

King's College London

2023

Table of contents

TABLE OF CONTENTS II

TABLE OF FIGURES V

TABLE OF TABLES VII

ACKNOWLEDGEMENTS IX

ABSTRACT 1

Chapter 1. Introduction 2
1.1. Classifying ALS 2
1.2. Patient characteristics and the spectrum of disease 3
1.3. Genetic contributions to ALS 5
1.3.1. The spectrum of ALS genetics 5
1.3.2. Translating genotype to phenotype 7
1.4. Treating ALS 9
1.5. Defining disease subtypes 10
1.6. Future steps 12

Chapter 2. Summary of objectives and thesis overview 13
2.1. Mathematically characterising genotype-phenotype relationships 13
2.1.1. Developing a novel approach to calculate genetic penetrance 13
2.1.2. Modelling population genetic screening in rare neurodegenerative disease 14
2.2. Characterising subtypes of ALS and genetic variation shared across diseases 14
2.2.1. Identifying biological subtypes of ALS with latent class clustering analysis 15
2.2.2. An online utility for comparative phenotype analysis in ALS 15
2.2.3. Examining genetic overlaps between neuropsychiatric disease 16

Chapter 3. Methodology 17
3.1. Probability and Bayesian mathematical principles 17
3.2. Genome-wide association study summary statistic processing 18
3.3. Ethical approval 20
3.4. Data and code availability 20

Chapter 4. Developing a novel approach to calculate genetic penetrance 21
4.1. Publication 21
4.2. Abstract 22
4.3. Background 22
4.4. Material and methods 24
4.4.1. Model 24
4.4.2. Application to penetrance calculation 27
4.4.3. Case examples 33
4.5. Results 38
4.6. Discussion 41
4.7. Conclusions 46

Chapter 5. Modelling population genetic screening in rare neurodegenerative diseases 47
5.1. Abstract 47
5.2. Background 47
5.3. Bayesian framework 49

5.4. Case studies ... 51
5.4.1. Case 1 – Huntington's disease ... 52
5.4.2. Case 2 – amyotrophic lateral sclerosis ... 52
5.4.3. Case 3 – phenylketonuria ... 54
5.5. Post-test disease probability ... 56
5.5.1. Screening versus diagnostic testing ... 56
5.5.2. Relative risk and secondary testing ... 56
5.5.3. Constraints upon post-test disease probability ... 57
5.6. Practical implementation of genetic screening ... 61
5.6.1. Marker selection ... 61
5.6.2. Utility over time and actionability ... 62
5.7. Limitations ... 63
5.8. Conclusion ... 63

Chapter 6. Identifying biological subtypes of ALS with latent class clustering analysis ... 65
6.1. Abstract ... 65
6.2. Background ... 66
6.3. Methods ... 68
6.3.1. Sample ... 68
6.3.2. Study design ... 71
6.3.3. Procedure ... 74
6.4. Results ... 78
6.4.1. Clustering of ALS clinical data ... 78
6.4.2. Clinical characterisation of clusters ... 80
6.4.3. Biological trends across clusters ... 84
6.4.4. Prediction of cluster membership using baseline data ... 87
6.5. Discussion ... 89

Chapter 7. An online utility for comparative phenotype analysis in ALS ... 94
7.1. Abstract ... 94
7.2. Background ... 94
7.3. Materials and methods ... 96
7.3.1. Dataset ... 96
7.3.2. Functionality ... 97
7.3.3. Tool design ... 99
7.3.4. Examples of use ... 99
7.4. Results ... 102
7.4.1. Amino acid hydrophobicity analysis ... 102
7.4.2. p.G94 amino acid residue analysis ... 105
7.5. Discussion ... 105

Chapter 8. Examining genetic overlaps between neuropsychiatric diseases ... 108
8.1. Abstract ... 108
8.2. Background ... 108
8.3. Methods ... 110
8.3.1. Sampled GWAS summary statistics ... 110
8.3.2. Procedure ... 110
8.4. Results ... 115
8.4.1. Genome-wide analyses ... 115
8.4.2. Targeted genetic analyses ... 119
8.5. Discussion ... 122

Chapter 9. Summary and future directions ... 126
9.1. Summary of findings ... 126
9.2. Future directions ... 128

REFERENCES 132

APPENDIX A. CHAPTER 4 SUPPLEMENTARY MATERIALS 160
Appendix A.1. Supplemental methods 160
Appendix A.1.1. Penetrance calculation procedure 160
Appendix A.1.2. Approach validation and testing 165
Appendix A.1.2.1. Lookup table validation: an alternative maximum-likelihood approach 165
Appendix A.1.2.2. Age-dependent penetrance: tolerance to age of sampling 166
Appendix A.1.2.3. Simulation studies 169
Appendix A.1.3. ADPenetrance: a companion web tool 179
Appendix A.2. Supplemental figures 181
Appendix A.3. Supplemental tables 193

APPENDIX B. CHAPTER 5 SUPPLEMENTARY MATERIALS 201
Appendix B.1. Supplemental methods 201
Appendix B.1.1. Estimating analytical validity: sensitivity and specificity 201
Appendix B.1.2. Parameter estimates by case study 202
Appendix B.1.2.1. Huntington's disease 203
Appendix B.1.2.2. Amyotrophic lateral sclerosis 204
Appendix B.1.2.3. Phenylketonuria 209

APPENDIX C. CHAPTER 6 SUPPLEMENTARY MATERIALS 211
Appendix C.1. Supplemental methods 211
Appendix C.1.1. Analysis of gene expression and methylation data 211
Appendix C.1.2. Prediction of cluster membership using baseline data 212
Appendix C.1.2.1. Machine-learning algorithm training procedure 213
Appendix C.1.2.2. Assessment of model performance and feature importance 215
Appendix C.2. Supplemental figures 217
Appendix C.3. Supplemental tables 229

APPENDIX D. CHAPTER 7 SUPPLEMENTARY MATERIALS 240
Appendix D.1. Supplemental figures 240
Appendix D.2. Supplemental tables 240

APPENDIX E. CHAPTER 8 SUPPLEMENTARY MATERIALS 251
Appendix E.1. Supplemental figures 251
Appendix E.2. Supplemental tables 255

Table of Figures

Figure 4-1. Summary of the key steps within this penetrance estimation approach 29
Figure 4-2. Example interface and output of the ADPenetrance web tool 32
Figure 5-1. Probability of a disease given a positive genetic test result for a marker of increased disease risk (P(D|T)) according to the sensitivity (P(T|M)) and the specificity (P(T'|M')) of the testing protocol 58
Figure 5-2. Probability of disease D following a positive genetic test result for marker M (P(D|T)) according to pre-test disease probability (P(D)).....59
Figure 5-3. Change in disease risk following a positive test result for a marker of increased disease risk (P(D|T)) according to penetrance (P(D|M)).....60
Figure 6-1. Summary of data processing and samples available by country for the Project MinE and STRENGH cohorts 70
Figure 6-2. Trends in clinical features used in latent class analysis according to class 82
Figure 6-3. Distribution of people across the first two axes of linear discriminant analysis for all case-complete data 83
Figure 6-4. Odds ratios for association between polygenic risk scores (PRS) for neuropsychiatric diseases and class 86
Figure 6-5. Receiver operator characteristic curves for performance of eXtreme Gradient Boosting algorithms in classifying each class versus all other classes.....89
Figure 7-1. Variant characteristics for the native dataset 97
Figure 7-2. Kaplan-Meier survival curves for age of onset and disease duration analyses ..102
Figure 8-1. Overview of the analysis procedure for this study.....111
Figure 8-2. Genome-wide genetic correlation estimates between all trait pairs 116
Figure 8-3. Local genetic correlation analyses between trait pairs.....118
Figure 8-4. Evidence for colocalisation between amyotrophic lateral sclerosis (ALS) and Alzheimer's disease (AD) in the Chr6:32.63-32.68Mb region 122

Figure A-1. Error in unadjusted penetrance estimates across true penetrance values and according to states modelled for a simulated population where sibship sizes follow a given distribution.....181
Figure A-2. Errors in unadjusted penetrance estimates across true penetrance values and according to states modelled for a simulated population. 182
Figure A-3. Sibship distributions upon which simulated populations were modelled across simulation studies 183
Figure A-4. Cumulative density plots comparing variability in age of ALS onset for people with and without SOD1 or C9orf72 gene variants.....184
Figure A-5. Error in penetrance estimates across true penetrance values when $RXobs$, N and g are specified correctly in the simulated UK (1974) and Next Steps populations 185
Figure A-6. Error in penetrance estimates according to degree of error in estimation of N186
Figure A-7. Error in penetrance estimates according to degree of error in estimation of $RXobs$ 187
Figure A-8. Error in penetrance estimates according to magnitude of disease risk g for people not harbouring the variant. 188
Figure A-9. Penetrance according to age of sampling across only families harbouring a variant of lifetime penetrance f.....189

Figure A-10. Error in lifetime penetrance estimates according to age of sampling based on variant frequency estimates across a wider disease cohort with equal age of onset variability 190
Figure A-11. Error in lifetime penetrance estimates according to age of sampling based on variant frequency estimates across a wider disease cohort when age of onset density is more compressed among people harbouring the tested variant 191
Figure A-12. Error in lifetime penetrance estimates according to age of sampling based on variant frequency estimates across a wider disease cohort when age of onset density is less compressed among people harbouring the tested variant 192
Figure C-1. Upset plots for missingness across clinical features used in LCA for the Project MinE (top) and STRENGTH (bottom) cohorts 217
Figure C-2. Density plot for diagnostic delay across countries for the combined Project MinE and STRENGTH datasets 218
Figure C-3. Accuracy of K-Nearest Neighbours algorithm for predicting class membership in STRENGTH according to number of neighbours 219
Figure C-4. Distribution of people across clusters of the 5-class models fitted to the discovery (Project MinE) and joint (Project MinE and STRENGTH) datasets 220
Figure C-5. Distribution of people and clusters across the first two linear discriminant analysis axes when restricting to people with non-censored disease duration 221
Figure C-6. Distribution of people and clusters across the first two discriminant analysis axes, stratified by country of origin 222
Figure C-7. Receiver operating characteristic curves for random forest and eXtreme Gradient Boosting algorithm predictions of class membership using only clinical data available around the time of diagnosis across all people with complete clinical data 223
Figure C-8. Receiver operating characteristic curves for random forest and eXtreme Gradient Boosting algorithm predictions of class membership using clinical data available around the time of diagnosis and measures of genetic disease liability 224
Figure C-9. Receiver operating characteristic curves random forest and eXtreme Gradient Boosting algorithm predictions of class membership using clinical data available around the time of diagnosis 225
Figure C-10. SHapley Additive exPlanations (SHAP) of feature importance across trained classification algorithms 226
Figure C-11. Comparison of classes from the current latent class model of ALS with the model from a previous study 228
Figure D-1. Kaplan-Meier survival curves for age of onset and disease duration analyses of trends associated with wild type and variant amino acid hydrophobicity 240
Figure E-1. SNP-wise p-value distribution between trait pairs in comparisons where colocalisation analysis suggested a causal variant in both traits 251
Figure E-2. Heatmaps of linkage disequilibrium (LD) in the 1000 Genomes European reference population across variants assigned to any credible set during univariate fine-mapping of trait pairs 252
Figure E-3. Sensitivity of colocalisation analysis to the prior probability of a shared variant between traits 254

Table of Tables

Table 4-1. Valid disease state combinations and corresponding weighting factors for estimating disease state rates 28
Table 4-2. Penetrance estimation across case studies 39
Table 5-1. Input parameters and disease risk estimates following testing for all case study scenarios 55
Table 6-1. Class membership characteristics for a 5-class model in the Project MinE, STRENGTH, and joint datasets 79
Table 6-2. Descriptive statistics for the clinical characteristics of people with ALS across the 5-class solution fitted to the joint dataset 81
Table 6-3. Results of the linear discriminant analysis of all people with case-complete data 84
Table 6-4. Association between class and rare genetic variation in ALS-associated genes 85
Table 6-5. Results of cell composition and omics-based age analysis for BrainBank samples in Class 1 and Class 2 with matching motor cortex expression data 87
Table 6-6. Performance of eXtreme Gradient Boosting classification algorithms for predicting class membership 88
Table 7-1. Data summary for case studies 101
Table 7-2. Inferential statistics for survival analyses across case studies 103
Table 8-1. Genome-wide association studies (GWAS) sampled 115
Table 8-2. Comparison of genome-wide SNP significance against local genetic correlation significance thresholds in all trait pairs and loci analysed 117
Table 8-3. Colocalisation analysis conducted across 95% credible sets identified during univariate fine-mapping of trait pairs 121

Table A-1. Sample characteristics and calculation of N for data applied in case study 1 193
Table A-2. Penetrance estimation of the LRRK2 p.Gly2019Ser variant for Parkinson's Disease across populations sampled in case study 1 195
Table A-3. Penetrance estimation for heterozygous inheritance of widely-described SOD1 variants 197
Table A-4. Estimation of the incidence of amyotrophic lateral sclerosis relative to frontotemporal dementia among people of European ancestry who harbour the pathogenic hexanucleotide GGGGCC repeat expansion of the C9orf72 gene ($C9orf72^{RE}$) 198
Table A-5. Comparison of unadjusted penetrance estimates derived for the case studies presented in Table 4-2 between the lookup table and maximum-likelihood approaches 199
Table A-6. Direction of change in $R(X)^{obs}$ and penetrance estimates according to increases in variant frequency and weighting factor inputs 200
Table B-1. Performance benchmarks of next generation sequencing tools specialised for genotyping different types of variants 202
Table B-2. Case study assumptions 203
Table B-3. FUS gene variants recorded in ClinVar (Landrum et al., 2018) as "pathogenic" or "likely pathogenic" for amyotrophic lateral sclerosis (ALS) and their prevalence in databases of people with familial and sporadic disease 205
Table B-4. Estimates of variant frequency among people with ALS, P(M|D), across ALS case study scenarios 206
Table B-5. Estimation of penetrance for FUS variants within the adpenetrance approach based on variant frequencies in people with familial and sporadic ALS 208

Table C-1. Mean and standard deviation of diagnostic delay per country of origin across unique samples from Project MinE and STRENGTH 229
Table C-2. Summary of number of variants identified in ALS-associated genes and assignment of genes to disease pathways according to evidence of role in gene products in pathway 230
Table C-3. Comparison of latent class model solutions for the Project MinE and for the Joint datasets.....233
Table C-4. Five-class latent class model solutions when restricting to samples with recorded diagnostic delay and disease duration 234
Table C-5. Validation of the Project MinE dataset 5-class model solution using independent data from STRENGTH within a K-nearest neighbours (KNN) classification algorithm.....235
Table C-6. Results of linear discriminant analysis after restricting to people with non-censored disease duration.....235
Table C-7. Results of multinomial regression analysis of all people with no missingness across predictors, including people with censored disease duration 236
Table C-8. Results of multinomial regression analysis of all people with no missingness across predictors, restricted to people with non-censored disease duration 237
Table C-9. Summary of cox proportional-hazards model predicting disease duration from onset until death or censoring using Class and all other clinical features from LCA 238
Table C-10. Results of differential expression analysis between BrainBank samples in Class 1 and 2 238
Table C-11. gProfiler gene enrichment results for the top 500 differentially expressed genes between BrainBank samples in Class 1 and 2.....239
Table C-12. Optimum hyperparameter tuning settings and overall AUC for all machine-learning algorithms trained 239
Table D-1. Sample composition by variant including hydrophobicity group assignments ...240
Table D-2. Cox Proportional-Hazards survival analysis for age of onset across all hydrophobicity groups.....245
Table D-3. Cox Proportional-Hazards survival analysis for age of onset across all hydrophobicity groups.....245
Table D-4 SOD1 variants reported in gnomAD v2.1.1 that have a recorded protein consequence and their hydrophobicity group assignments 246
Table E-1. Results of all bivariate local genetic correlation analyses 255
Table E-2. Results of colocalisation analyses performed across all SNPs sampled in region 255
Table E-3. Overview of credible sets identified across fine-mapping analyses in summary statistics harmonised across trait pairs 256

Acknowledgements

This PhD was supported by the National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre and the Motor Neuron Disease Association. I would like to thank my funders for their support and allowing me to contribute research aiming to improve understanding of amyotrophic lateral sclerosis with the ultimate goal of finding a cure.

I would like to acknowledge the contribution of my supervisors Professor Ammar Al-Chalabi and Dr Alfredo Iacoangeli who have provided invaluable guidance throughout this PhD.

I would like to thank Kristina Yordanova who has shown endless patience and graciously supported me throughout the many late nights and weekends. Thank you for picking up the slack where I was unable. This truly would not have been possible without you and words cannot express my gratitude.

Thank you also to my family for your continuous support and encouragement.

I thank the families and people with ALS and other diseases who have contributed to the datasets upon which these investigations were based. Without you this research would not be possible.

The word ‘we’ is used throughout this thesis in recognition that, as is typical for this collaborative field, each investigation is the product of collaboration between me and other researchers/clinicians. All formal analyses are my own work with oversight from these collaborators, with the following exceptions:

- In Chapter 6:
    - I had no role in sample collection or initial processing of the data from Project MinE and STRENGTH data, which were performed prior to this study. This includes the data quality control and principal component analysis in Project MinE. I was, however, responsible for preparing the final datasets for use in these studies, as described within the chapter.

    - All analysis of gene expression and methylation data was performed by Heather Marriott using code developed by herself and Renata Kabiljo.
    - Data on rare variants harboured by people within the Project MinE whole-genome sequencing cohort were provided by Alfredo Iacoangeli.
- In Chapter 7, dataset curation and cleaning were performed by Sarah Opie-Martin as part of her previous work.

I would like to extend particular thanks to:

- Sarah Opie-Martin, whose own research laid the groundwork for Chapter 7, and who has provided guidance wherever needed.
- Heather Marriott and Guy Hunt, whose contributions have greatly enriched the work described in Chapter 6 and who provided monumental support throughout the PhD.
- Oliver Pain, whose oversight has helped extend this work in exciting directions.

I would like to also recognise the contributions of Harry Bowles, Munishikha Kalia, Renata Kabiljo, Ahmad Al Khleifat, Lachlan Gilchrist, Cathryn Lewis, Daniel Stahl, Nicholas Cummins, Mina Ryten, Francesca Forzano, Neil Pearce, and Isabella Fogh who have all graciously shared their ideas and expertise, which have greatly helped across various stages of this work.

Abstract

Amyotrophic lateral sclerosis (ALS) is a devastating neurodegenerative disease characterised by progressive, widespread, degeneration of the upper and lower motor neurons until death from respiratory paralysis. Two major facets of ALS are the heterogenous patient population and complex biological architecture. This heterogeneity may obstruct the discovery of effective therapies, none of which presently exist. Refining knowledge about the genetic basis of the disease could improve understanding of patient phenotypes while improved classification of disease subgroups and biological mechanisms overlapping with other diseases may enable discovery of much-needed treatments.

Accordingly, the focus of this thesis is upon genotype-phenotype relationships in ALS. Included within are investigations of (1) disease risk for people found to harbour certain genetic variants, (2) disease subgroups identified via data-driven approaches or defined by specific genetic variation, and (3) genetic overlaps with Alzheimer's disease, frontotemporal dementia, Parkinson's disease, and schizophrenia. Several tools were developed across these studies, including (1) a novel method for calculating genetic penetrance, implemented within an R function and companion web-tool, (2) a web-server for comparing the ALS phenotype across different groups of people, and (3) a command-line workflow for statistical fine-mapping and colocalisation analysis. These utilities have been made freely available for use in future research.

The first chapter of the thesis overviews current understanding of ALS, focused on its clinical and genetic spectrum, and outlines the current therapeutic landscape and previous attempts to identify homogenous disease subtypes. Subsequent chapters summarise the objective of each study, general methodology, and the investigations performed. The thesis concludes with an overall summary of findings and outlines directions for future research building upon this work.

Chapter 1. Introduction

Amyotrophic lateral sclerosis (ALS), also known as 'motor neuron disease', is a devastating neurodegenerative disease characterised by progressive, widespread, degeneration of the upper and lower motor neurons until death from respiratory paralysis (R. H. Brown & Al-Chalabi, 2017). The lifetime risk of ALS is estimated to be between 1 in 300-400 people, and it is more frequent in men than women (Alonso, Logroscino, Jick, & Hernán, 2009; R. H. Brown & Al-Chalabi, 2017; Johnston et al., 2006). Point prevalence of the disease is estimated at around ~5 per 100,000 persons and incidence at ~1-3 per 100,000 person-years (Alonso et al., 2009; Chiò et al., 2013; P. Mehta et al., 2018).

1.1. Classifying ALS

ALS was first described in the 19th century. The term was coined by Jean-Martin Charcot around 1874, following recognition of the frequent co-occurrence of patterns of degeneration we would now describe as primary lateral sclerosis and progressive muscular atrophy, which are respectively characterised by degeneration restricted to either the upper or lower motor neurons (Rowland, 2001; Turner, Barnwell, Al-Chalabi, & Eisen, 2012). Documented cases of ALS precede the term, dating back to at least 1853. Overlap between ALS and other neurodegenerative syndromes is increasingly acknowledged in contemporary literature. Historic evidence of this dates back to the early 20th century.

The 'Gold Coast' criteria are a recent revision of ALS classification guidelines (Shefner et al., 2020). Three diagnostic criteria are specified:

- Progressive motor impairment
- Upper and lower motor neuron dysfunction in at least one body region, or lower motor neuron dysfunction in at least two regions
- Appropriate investigation to exclude other diseases

These supersede the El Escorial criteria, which were first published in 1994 and had undergone several subsequent refinements (Al-Chalabi et al., 2016). One key improvement over past criteria is simplification into a binary 'ALS' or 'not ALS' decision; diagnoses were previously assigned confidence levels. Another improvement is the acknowledgement of the

heterogeneity of patient clinical presentations, including recognition of overlap with other disorders.

People with ALS are commonly stratified into 'familial' and 'sporadic' disease groups. Familial cases are those occurring alongside family disease history, which is absent for sporadic cases. A 2011 meta-analysis finds the familial rate to be ~5%, but consensus is lacking on what constitutes a positive family history and, accordingly, reported familiality rates are 5-30% (Byrne et al., 2011; Ryan et al., 2018; Vajda et al., 2017). Further issues also exist with these classifications. First, they allude to familial and sporadic ALS forms being distinct, yet clinically they are not (Chiò et al., 2014). Indeed, even an inherited disease origin may appear sporadic, with genetic and kinship characteristics affecting this likelihood (Al-Chalabi & Lewis, 2011). Second, grouping patients in this way can be detrimental, stratifying research samples and providing little-to-no benefit to patients (Al-Chalabi, 2017). Division of patients into meaningful subgroups, linked to specific disease pathways, will likely have greater benefit for both treatment and research.

1.2. Patient characteristics and the spectrum of disease

People with ALS do not display uniform demographic or clinical profiles. Disease may onset across a wide age range, but most commonly between ages 50 and 70 (Johnston et al., 2006; P. Mehta et al., 2018). Around 10% of contemporary cases emerge before age 45; this proportion was historically greater, a change reflecting the proportional relationship between age of onset and life expectancy (Byrne, Jordan, Elamin, & Hardiman, 2013; Turner et al., 2012). From first symptom onset, patients survive a median of three years (Al-Chalabi & Hardiman, 2013). Aggressive manifestations result in death after one year, and others display a longer course; 5-10% of people with ALS survive for over 10 years (Chiò et al., 2009; Juneja, Pericak-Vance, Laing, Dave, & Siddique, 1997a).

Around two thirds of people have onset in the spinal cord, a third start in the bulbar region, and a small subset (~3%) begin in the respiratory region (Beeldman et al., 2016; R. H. Brown & Al-Chalabi, 2017; Masrori & Van Damme, 2020; Shoesmith, Findlater, Rowe, & Strong, 2007; van Es et al., 2017). Whether disease will predominantly affect upper or lower motor neurons is also variable, as is the symmetry of degeneration across the body (Al-Chalabi &

Hardiman, 2013; Al-Chalabi et al., 2016). Flail arm and leg syndromes are symmetrical phenotypes predominantly affecting the upper or lower limbs. Progressive bulbar palsy labels degeneration restricted to bulbar musculature. People with initially restricted degeneration often develop a more generalised phenotype as disease progresses (Al-Chalabi et al., 2016).

Motor-centric aspects only constitute part of the phenotypic disease spectrum, and the occurrence of non-motor symptoms including psychiatric or cognitive abnormalities associated with other disorders is increasingly recognised (Beeldman et al., 2016; Fang, Jozsa, & Al-Chalabi, 2017; Ferentinos et al., 2011; Zarei et al., 2015; Zucchi, Ticozzi, & Mandrioli, 2019). Disorders with reported clinical or biological overlaps with ALS include frontotemporal dementia (FTD), Parkinson's disease, Alzheimer's disease, depression, epilepsy, and schizophrenia.

The ALS-FTD overlap is particularly notable. FTD is the second most common cause of early-onset dementia, and labels a group of heterogenous degenerative disorders characterised by changes in cognition, behaviour, or language proficiency (Ratnavalli, Brayne, Dawson, & Hodges, 2002; Young, Lavakumar, Tampi, Balachandran, & Tampi, 2017). Neurologically, FTD syndromes are characterised by progressive degeneration of either or both of the frontal and temporal lobes of the brain. In some studies, up to 15% of people with ALS have a joint FTD diagnosis and cognitive dysfunction has been found in around 50% (Beeldman et al., 2016; Bora, 2017; Byrne, Heverin, et al., 2013; Crockford et al., 2018; Montuschi et al., 2015; J. Murphy et al., 2016; Phukan et al., 2012; Zucchi et al., 2019). The frequency of FTD increases by ALS clinical stage, which suggests a common underlying pathology (Crockford et al., 2018) we now know to be the finding of TDP43 protein inclusions in affected and unaffected neurons. Other evidence supporting the considerable ALS-FTD overlap is their high co-occurrence within affected kindreds, and a partially shared genetic basis (Balendra & Isaacs, 2018; Boeve et al., 2012; Lattante, Ciura, Rouleau, & Kabashi, 2015). ALS and FTD are therefore thought to represent a spectrum of disease, whose architecture remains to be fully characterised but includes TDP43 proteinopathy.

1.3. Genetic contributions to ALS

Genetic and environmental factors are both important in shaping ALS disease outcomes, which result through their interplay (Al-Chalabi & Visscher, 2014; Cady et al., 2015; Chiò et al., 2018; Morgan et al., 2017; Shatunov & Al-Chalabi, 2021). Both common and rare variants contribute towards the genetic portion of disease variance.

Heritability estimates, which describe the relative contribution of genetic and environmental factors to variance in a phenotype, demonstrate the shared importance of genetic and environmental effects in ALS. Leveraging family-based data from twin studies and parent-offspring dyads, 50-60% of variance in ALS has been attributed to genetic factors (Al-Chalabi et al., 2010; McLaughlin, Vajda, & Hardiman, 2015; Ryan, Heverin, McLaughlin, & Hardiman, 2019). Genomic estimates capturing variance attributable to common single nucleotide variants estimate heritability at ~8-21% (Fogh et al., 2014; Keller et al., 2014; van Rheenen et al., 2016). The disparity between family-based and genomic figures likely reflects the contribution of rare variants not represented within the genomic data; it is expected to fall in genomic studies which apply whole-genome sequencing data (Al-Chalabi & Visscher, 2014; Wainschtein et al., 2022).

1.3.1. The spectrum of ALS genetics

Variants in over 40 genes are implicated as causal for or modifiers of ALS (Chia, Chiò, & Traynor, 2018; P. R. Mehta et al., 2022; Shatunov & Al-Chalabi, 2021). The most recent genome-wide association study (GWAS) for risk of ALS identified 12 genome-wide significant loci in analysis of European (n_{cases} = 27,205; $n_{controls}$ = 110,881) populations, and a further 3 within cross-ancestry analysis of European and Asian (n_{cases} = 2,407; $n_{controls}$ = 11,775) cohorts (van Rheenen et al., 2021).

The most prevalent genetic cause of ALS is a pathogenic hexanucleotide GGGGCC repeat expansion in the *C9orf72* gene, which accounts for 6% of ALS in European, 1% in Asian, and (across 103 individuals) ~7% in African populations (Marogianni et al., 2019; Nel et al., 2022). *C9orf72 is* located at chromosome 9p21.2 and was identified in 2011; its discovery was preceded by years of awareness about the involvement of chromosome 9p in ALS

(DeJesus-Hernandez et al., 2011; Morita et al., 2006; Shatunov et al., 2010; van Es et al., 2009; Vance et al., 2006).

The other prominent gene is *SOD1* which contains over 180 ALS-associated non-synonymous variants, which cumulatively account for 2% of European and 3% of Asian, and (across 103 individuals) ~5% of African ALS (Abel, Powell, Andersen, & Al-Chalabi, 2012; Nel et al., 2022; Opie-Martin et al., 2022; Z.-Y. Zou et al., 2017). *SOD1* is located at chromosome 21q22.11 and was the first known genetic cause of ALS, identified in 1993 (https://www.alsod.ac.uk; Abel et al., 2012; Rosen et al., 1993).

Most implicated genes individually account for a small proportion of cases, and known variants cumulatively explains disease for around 15-20% of people; 70% of familial and 15% of sporadic presentations, when stratified by family history (R. H. Brown & Al-Chalabi, 2017; Chia et al., 2018) Some genes (e.g., *SOD1*, *C9orf72*, *TARDBP,* and *FUS*) have been widely investigated (Marogianni et al., 2019; Z.-Y. Zou et al., 2017), while others (e.g. *DNAJC7* and *NEK1*) are less well documented (Farhan et al., 2019; Kenna et al., 2016). Since many variants putatively pathogenic for ALS are rare, their consequence is rarely characterised at an individual level, instead trends are found by aggregating across variants within a burden analysis framework (Dekker et al., 2019; Farhan et al., 2019). The p.A5V *SOD1* mutation is one variant that has received a particular focus; it is associated with an aggressive disease course and survival of around one year after disease onset (Rosen et al., 1994; Saeed et al., 2009).

Variants associated with ALS have varied characteristics. Some result from small genetic alterations, as is seen across the spectrum of single-nucleotide variants in *SOD1* (https://alsod.ac.uk/; Abel et al., 2012). Others result from larger structural genetic changes, such as protein-truncation variants in *NEK1* and short-tandem repeat expansion variants in *C9orf72* and *ATXN1* (DeJesus-Hernandez et al., 2011; Farhan et al., 2019; Tazelaar et al., 2020).

The burden of ALS-linked genetic variation ranges between a large monogenic effect, akin to Mendelian disease, and a polygenic, or modifying, effect, akin to complex disease (Al-

Chalabi, van den Berg, & Veldink, 2017; Simpson & Al-Chalabi, 2006). Variants with monogenic disease associations most frequently display an autosomal dominant inheritance pattern but autosomal recessive inheritance is also observed (Pensato et al., 2020; Weishaupt, Hyman, & Dikic, 2016). Genes associated with Mendelian ALS presentations include *SOD1, FUS, and TARDBP* (Al-Chalabi et al., 2017; Kenna et al., 2013). Genes containing variants with a modifying effect include *UNC13A* and *TMEM106B,* which reportedly affect the likelihood of FTD features emerging in ALS patients (Placek et al., 2019; van Blitterswijk et al., 2014).

Some people have an oligogenic disease basis, where ALS occurs in the presence of two or more variants (Giannoccaro et al., 2017; Lattante et al., 2015; Nguyen, Van Broeckhoven, & van der Zee, 2018; Pang et al., 2017; van Blitterswijk et al., 2013; van Blitterswijk et al., 2012). Harbouring multiple disease-linked variants likely shapes the phenotype, for example, producing an earlier age of onset or more aggressive disease than in those with fewer disease-linked mutations (Pang et al., 2017; van Blitterswijk et al., 2013). Oligogenic disease is particularly documented among some people harbouring the *C9orf72* repeat expansion, and this variant is thought to confer an effect between the Mendelian and polygenic extremes (Al-Chalabi et al., 2017).

1.3.2. Translating genotype to phenotype

As the spectrum of ALS-associated genetics would suggest, research has shown various pathways to disease. Indeed, the products of genes associated with ALS are implicated in multiple molecular pathways (R. H. Brown & Al-Chalabi, 2017; Masrori & Van Damme, 2020; Nguyen et al., 2018; Weishaupt et al., 2016). Associated processes include proteostasis, autophagy, RNA metabolism, mitochondrial regulation, DNA repair, cytoskeletal dynamics and transport, and inflammation. Unfortunately, the mechanisms through which ALS emerges following dysfunction in these pathways remain unclear.

The multistep model of ALS has been applied to demonstrate how different genetic burdens modify disease susceptibility. The model posits that disease will onset after someone is exposed to a threshold number of disease-causing factors and is shown to fit ALS within several independent populations (Al-Chalabi et al., 2014; Chiò et al., 2018; Vucic et al.,

2020). Typically, around 5-6 molecular steps are required for ALS to onset. However, it is reported that as few as 2-4 steps are required for onset in individuals with vulnerabilities in key genes (*SOD1*, *C9orf72,* and *TARDBP*). An oligogenic disease basis also fits within this model, where genetic burdens can cumulatively contribute towards disease-steps.

Other key determinants of genotype-phenotype relationships must be considered when investigating the architecture of ALS.

The first is pleiotropy, where a genotype is associated with multiple phenotypes. In ALS, pleiotropy is increasingly recognised. The *C9orf72* repeat expansion illustrates this, as carriers are roughly equally likely to develop ALS or FTD, and, less frequently, other neuropsychiatric conditions (Cooper-Knock, Kirby, Highley, & Shaw, 2015; Majounie et al., 2012; N. A. Murphy et al., 2017). Complex genetic influences, from variants with individually small but compounding effects, also appear pleiotropic. Genetic correlations have been shown with other neuropsychiatric traits such as Alzheimer's disease, Parkinson's disease, and schizophrenia (C. Li, Yang, Ou, & Shang, 2021; McLaughlin et al., 2017; van Rheenen et al., 2021). ALS also has positive genetic correlations with smoking status and moderate physical activity, and negative genetic correlations with cognitive performance, educational attainment, and light exercise (Bandres-Ciga et al., 2019; Restuadi et al., 2022; van Rheenen et al., 2021). Pleiotropy is expected within the multistep hypothesis of ALS, because different downstream events may profoundly affect the ultimate phenotype.

The second determinant is penetrance, the probability of an outcome (e.g., disease) given the person harbours a particular variant; this can be considered at an age-dependent level or across the lifespan. Again, this is a prediction of the multistep model, since carrying an implicated variant is not sufficient for disease, and the remaining steps are required for disease onset. Unfortunately, penetrance is poorly characterised in ALS (Chiò et al., 2014; N. A. Murphy et al., 2017). It has been estimated in variants in some genes, such as *SOD1*, *TARDBP*, *FUS* and *C9orf72*, but the accuracy of these is uncertain. This can be attributed to weaknesses of current penetrance estimation approaches. Pedigree-based estimates are difficult to generalise, since they are unique to the studied kinship and shaped by various interacting factors (Chiò et al., 2014). Population based estimates are, in the absence of

systematic sampling of healthy population members, flawed owing to ascertainment bias towards people affected (Chiò et al., 2014; N. A. Murphy et al., 2017). A novel approach to estimate penetrance in ALS would accordingly be beneficial.

1.4. Treating ALS

Discovery of effective treatments for ALS continues to be a pervasive challenge. The benefit of treatment with riluzole was first shown in 1994 (Bensimon, Lacomblez, & Meininger, 1994). However, this drug can extend life expectancy from onset by no more than a few months (Lacomblez, Bensimon, Leigh, Guillet, & Meininger, 1996; van Eijk et al., 2020). In 2017, edaravone became the second FDA approved drug treatment for ALS, slowing motor neuron deterioration but with little evidence to suggest benefits for life expectancy (Gao et al., 2023).

The poor efficacy of existing therapeutic strategies in ALS might reflect the heterogeneity of the disease, and that existing drugs are not suitable for all disease presentations. For instance, mouse models have shown riluzole to be ineffective for treating ALS associated with deleterious variants in genes such as *SOD1*, *TARDBP*, and *FUS* (A. L. Wright et al., 2021). Meanwhile, other studies have suggested that treatment with lithium extends survival trajectories specifically among people homozygous for the 'C' allele of the rs12608932 single nucleotide variant in *UNC13A*, bringing survival in-line with people without the variant (van Eijk et al., 2017; Willemse et al., 2023). Investigation into the therapeutic utility of lithium is ongoing.

Accordingly, evidence to date suggests that successful treatment of ALS may lie within a precision medicine framework, which would tailor therapy to the disease cause relevant for each person. Clinical trials of precision medicine strategies are in progress. To date, these have focussed on gene therapies aiming to offset aberrant function associated with specific genetic variation (Amado & Davidson, 2021). For example, tofersen is an antisense oligonucleotide drug to treat ALS for the 2-3% of people with *SOD1* variants by impeding translation of SOD1 protein (Miller et al., 2022). The drug is not yet approved but reduction in functional decline across several measures after 52 weeks suggests that the approach holds promise for a breakthrough in ALS therapeutics.

1.5. Defining disease subtypes

Beyond diagnostic labels, ALS subtypes are often defined by genetic variation believed to cause the disease (e.g., *SOD1*- or *C9orf72*-associated ALS). These are useful groupings which should be pursued to allow genetically-targeted treatment strategies, as discussed in 1.4. Indeed, this is particularly so in instances like *SOD1*-ALS which appears biologically separate from non-*SOD1* disease (Mackenzie et al., 2007). However, a 'per-gene' approach may not always be effective given the breadth of biological architecture for ALS. This is because, first, disease is often not associated with a monogenic cause and, second, individual variants/genes account for a small proportion of those affected. Accordingly, efforts to identify broader disease subgroups defined by a subgroup-relevant disease mechanism are warranted. This aim seems reasonable since disease associated with certain genetic variation, such as the *C9orf72* repeat expansion, appears biologically indistinguishable from presentations without the variant (Humphrey et al., 2023).

For such subgroups to be identified, we must understand the features which distinguish them. Data-driven approaches to tease apart the heterogenous and interlinked biological and phenotypic components of ALS will likely lie at the centre of these investigations. Comparison between distinct types of ALS may then permit discovery of subgroup particular disease mechanisms (Jones et al., 2015), which themselves can inform future therapeutic approaches.

Previous work aiming to break-down the heterogeneity of ALS has described five subgroups, identified with latent class cluster analysis, in clinical data sampled from a cohort of 1,467 people from the United Kingdom (Ganesalingam et al., 2009). These subgroups were predictive of disease duration from onset until death or censoring and distinguished first by time from onset to diagnosis (diagnostic delay) and second by site of onset (bulbar or other). Another investigation, applying high-dimensional clinical data and a machine-learning approach across two independent Italian cohorts ($n_{discovery}$ = 2,361; $n_{replication}$ = 989), found clusters conforming to the following clinical subgroups: bulbar, respiratory, flail arm, classical, pyramidal, and flail leg (Chiò, Calvo, Moglia, Mazzini, & Mora, 2011; Faghri et al., 2022). They found that 11 of the included features were important to the model, including

those relating to the site/anatomical level of disease onset and measures of disease progression. These clinically-based studies are useful but do not explore whether the identified subgroups map onto any biological disease trends.

A number of biologically-driven clustering studies do exist and have found subgroups derived from transcriptomic and neuroanatomical data that appear to be reflected across different data modalities (Bede, Murad, Lope, Hardiman, & Chang, 2022; Bede, Murad, Lope, Li Hi Shing, et al., 2022; Dukic et al., 2021; Eshima et al., 2023; Tam et al., 2019). To give examples, one study (N = 208) finds shorter disease duration for people with increased expression of genes associated with glial activation in comparison to other transcriptomic subgroups (Eshima et al., 2023). A second (N = 300) finds that clinical groups of 'ALS', 'primary lateral sclerosis', and 'poliomyelitis survivors' can be predicted using neuroimaging data (Bede, Murad, Lope, Li Hi Shing, et al., 2022). A third (N = 214) distinguishes between two neuroanatomical subgroups, finding the *C9orf72* repeat expansion to be approximately 3.5 times more frequent in one group than the other (Bede, Murad, Lope, Hardiman, et al., 2022).

Taken together, investigations into subtyping of ALS have demonstrated that data-driven disease subgroups can be found and that the signatures of these may translate between different biological modalities or be displayed in the phenotype. Studies to date have not, however, been validated in independent samples from different populations, limiting their generalisability. Moreover, despite ALS being clinically defined, no study has examined whether biological differences exist between clinically-defined subgroups identified through data-driven approaches.

Identification of homogenous ALS subgroups which are robust and consistent across populations would have huge potential benefit for ALS precision medicine discovery. Their utility would be particularly great if these groups can be identified using only data available at or before time of diagnosis, such as genetic variants acting as biomarkers for a particular subgroup.

1.6. Future steps

As outlined within this chapter, the spectrum of ALS is broad, both phenotypically and biologically. Effective treatment strategies are also lacking, and this may be attributable to the heterogeneity of the disease.

It is important that future work aims to better understand the differences between people with or at risk of later developing ALS. Consideration should be given to genetic contributions across the monogenic to polygenic spectrum, including how the phenotype differs between distinct variants. The relevance of variable penetrance and pleiotropy across variants should be better established.

Research should also aim to identify homogenous subtypes of ALS that share a common disease cause. Such investigations are critical for informing both patient care and development of novel therapeutic strategies. Research may only be able to distinguish these groups by sampling across the diverse spectrum of people affected by the disease. Accordingly, continually developing machine-learning and bioinformatics approaches will likely play a substantial role in these investigations, alongside large multi-national datasets (Al-Chalabi et al., 2017).

Since these research questions are wide-spanning and unlikely to be answered within a small number of studies, efforts should be made to develop utilities that facilitate novel investigations as new hypotheses are formed.

Chapter 2. Summary of objectives and thesis overview

The aim of this PhD thesis is to improve understanding of genotype-phenotype relationships in ALS and to define and characterise biologically-relevant disease subgroups. This work describes both empirical investigation of these questions and novel utilities developed to facilitate future work in the area.

Chapter 3 overviews methodology pertaining to multiple chapters. Chapter 4 and Chapter 5 focus on mathematical characterisation of genotype-phenotype relationships. Chapter 6, Chapter 7, and Chapter 8 present work aiming to characterise subtypes of ALS and genetic variation shared with other diseases. Chapter 9 summarises key findings from the investigations presented across the thesis and suggests directions for future studies.

2.1. Mathematically characterising genotype-phenotype relationships

The diversity in genetic architecture that can explain ALS means that risk of disease for people with an implicated genetic variant cannot be readily encapsulated within a single estimate. It is critical therefore to develop understanding of the impact of harbouring, or being suggested to harbour, particular ALS-associated genetic variants upon disease risk.

2.1.1. *Developing a novel approach to calculate genetic penetrance*

In Chapter 4, we aim to address the problem of poorly characterised genetic penetrance in ALS. This goal was approached through development of a novel approach to calculate variant penetrance using population-scale data. The method is built upon the disease model described in Al-Chalabi and Lewis (2011) and follows principals of autosomal dominant inheritance. It aims to circumvent the biases faced by other population-scale approaches by stratifying people according to family disease history, accounting for family size, as opposed to a traditional case-control design.

The chapter describes the approach and its application across several case studies, comparing against existing penetrance estimates of *LRRK2* variants for Parkinson's disease, and *BMPR2* variants for pulmonary arterial hypertension. Novel penetrance estimates were also made for variants in the key ALS genes *C9orf72* and *SOD1.* Implementation of the

approach within an R function and a web-utility aims to facilitate future investigations using the method.

2.1.2. Modelling population genetic screening in rare neurodegenerative disease

In Chapter 5, we mathematically model future disease risks for a person identified to harbour a variant conferring increased liability towards a rare neurodegenerative disease during genetic testing. We contrast, particularly, the difference in risk between a targeted testing scenario (e.g., where a parent is known to harbour a variant which may have been transmitted to the person tested) and a population-wide screening approach which could be applied indiscriminately across a population at any time from birth.

The work draws upon Bayesian mathematical principals and considers disease, variant, and test characteristics affecting the probabilities of subsequent disease following genetic test results. Several candidate case studies are described, including screening for variants implicated in ALS, considering penetrance estimates from Chapter 4, and in Huntington's disease and phenylketonuria.

The chapter aims to address the increasing interest in population genetic screening and provide important context and considerations for the interpretation of genetic test results indicating increased liability for a rare disease (Adhikari et al., 2020; Centers for Disease Control and Prevention, 2021; Dickinson et al., 2018; *Genome UK: The future of healthcare*, 2020; Jansen, Lister, van Kranen, & Cornel, 2017; Moorthie et al., 2021; Murray, Evans, & Khoury, 2019).

2.2. Characterising subtypes of ALS and genetic variation shared across diseases

The limited number and small benefit of existing treatments for ALS may be driven by a 'one-size-fits-all' therapeutic approach. In the advent of the precision and genomic medicine era, new opportunities exist to identify therapeutic strategies targeted to mechanisms relevant to a given person. For this to be possible, we must first define and develop understanding of distinct ALS subgroups, considering both whether these distinctions exist at a 'per-variant' or 'per-gene' level or whether they can be described more broadly by

shared biological mechanisms. Relevant to this issue is genetic pleiotropy, which, if better understood, may elucidate mechanisms in common with overlapping diseases.

2.2.1. Identifying biological subtypes of ALS with latent class clustering analysis

In Chapter 6, we aim to decompose the heterogeneity of ALS with a data-driven classification of homogenous disease subgroups across a multi-national dataset. The subgroups were defined using a latent class clustering analysis machine-learning approach and based on commonly collected clinical features. Their clinical characteristics were statistically examined, alongside differences in their genetic architecture across (1) rare genetic variation in genes previously implicated in ALS and (2) common variants captured within polygenic risk scores measuring genetic liability towards ALS and overlapping neuropsychiatric disorders. We additionally examined the extent to which subgroups can be predicted via machine-learning approaches using only data attainable around the time of diagnosis.

2.2.2. An online utility for comparative phenotype analysis in ALS

Chapter 7 describes design and example implementation of a web tool developed to facilitate analysis of the ALS phenotype across disease subgroups using primarily survival analysis methodology. Reflecting the comparatively high prevalence of *SOD1*-ALS and to allow better characterisation of the diverse phenotypic spectrum associated with variants in the gene, this utility provides access to a large sample of people with various SOD1 variants and a non-SOD1 comparator cohort (Opie-Martin et al., 2022).

This utility is one-step adjacent to the focus of Chapter 4 and Chapter 5, disease susceptibility, and permits analysis of disease trends in people who have developed ALS. It builds upon the clustering work of Chapter 6 in its provision of a facility to compare between different disease subgroups, focussed natively upon a spectrum of SOD1 variants, with full flexibility in how data are stratified and including various options to customise the analysis.

2.2.3. Examining genetic overlaps between neuropsychiatric disease

Chapter 8 describes an investigation of shared genetic variation between ALS and related neuropsychiatric disorders. The goals of these analyses were twofold. First, the study aimed to apply state-of-the-art statistical methods to identify and examine regions associated between traits. This was achieved through a genome-wide local genetic correlation analysis to identify associated loci, which were further analysed with statistical fine-mapping and colocalisation techniques. Second, to benefit future investigations, we aimed to develop a readily implemented workflow for the fine-mapping and colocalisation analysis protocol.

This chapter tackles the increasing recognition of pleiotropy in ALS. It extends the analyses of Chapter 6, which focused on differences in shared genetic architecture across different ALS subgroups, by contributing a broader analysis of the genetic landscape shared between neuropsychiatric diseases.

Chapter 3. Methodology

Most methods applied throughout this thesis are chapter-specific and therefore will be described in full within methods sections respective to each chapter. This methodology chapter will focus on the elements relevant to multiple chapters.

3.1. Probability and Bayesian mathematical principles

Principles of probability and particularly Bayesian theory are widely applied within evidence-based medicine (Hunink et al., 2014). These principles are also applicable to Chapter 4, Chapter 5, and Chapter 8. Accordingly, relevant mathematical concepts are briefly presented here.

Within probability theory, all outcomes (i.e., events) that may occur must have probabilities between 0 and 1. The probabilities for all possible outcomes in the event space of, for instance, 'A' must sum to 1. Therefore, for a binary outcome, if the event A has the probability $P(A)$, then the probability of not-A, A', would be $P(A')$ and can be calculated by subtracting from the total event space (i.e., $P(A') = 1 - P(A)$).

Bayes theorem describes the probability of event A given knowledge about other events upon which A is conditional. Considering the event A which is conditional upon event B,

Equation 3-1

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)} = \frac{P(A \cap B)}{P(B)},$$

letting $P(A)$ and $P(B)$ be the total probability of each of A and B and $P(A \cap B)$ be the probability that the events both occur. The probability of event A given that event B has occurred is denoted as $P(A|B)$, and $P(B|A)$ represents the probability of B given A. With reference to Equation 3-1, $P(A|B)$ is the 'posterior' probability of A given B is true, whereas $P(A)$ is the 'prior' probability of A (i.e., current knowledge about the probability of A).

The total probability of a conditioned event can be derived from the probabilities of all mutually exclusive occurrences of the event within the total event space. Letting A be conditioned upon event B,

Equation 3-2

$$P(A) = P(A|B) \times P(B) + P(A|B') \times P(B').$$

The chain rule is relevant for determining the probabilities of more than 2 events co-occurring. For instance, the probability that all events A, B, and C occur is:

Equation 3-3

$$P(A \cap B \cap C) = P(A) \times P(B|A) \times P(C|B \cap A).$$

Accordingly, the probability of both A and B given that C has occurred can be derived:

Equation 3-4

$$P(A \cap B|C) = \frac{P(A \cap B \cap C)}{P(C)} = P(B|C) \times P(A|B \cap C).$$

Where events A, B, and C are all relevant, the probability of A given that C has occurred can be determined in accordance with the principles of total probability and of Equation 3-4:

Equation 3-5

$$P(A|C) = P(A|B \cap C) \times P(B|C) + P(A|B' \cap C) \times P(B'|C).$$

If A has conditional independence from C when the outcome of B or not-B, B', is known, then $P(A|B \cap C) = P(A|B \cap C') = P(A|B)$ and $P(A|B' \cap C) = P(A|B' \cap C') = P(A|B')$. In this circumstance, Equation 3-5 can be simplified to:

Equation 3-6

$$P(A|C) = P(A|B) \times P(B|C) + P(A|B') \times P(B'|C).$$

3.2. Genome-wide association study summary statistic processing

Chapter 6 and Chapter 8 both draw upon summary statistics from European-ancestry genome-wide association study (GWAS) meta-analyses of risk for ALS (van Rheenen et al., 2021), frontotemporal dementia (Ferrari et al., 2014), Alzheimer's disease (Kunkle et al., 2019), Parkinson's disease (Nalls et al., 2019), and schizophrenia (Trubetskoy et al., 2022).

Different versions of the ALS GWAS were used across the chapters. For Chapter 6, GWAS summary statistics were applied to calculate polygenic risk scores (PRS) in the Project MinE dataset (Project MinE ALS Sequencing Consortium, 2018). Since the ALS GWAS included most of the Project MinE dataset within meta-analysis 'stratum 6', PRS for risk of ALS were derived from summary statistics which exclude the stratum and analyses using the ALS PRS

were performed using only samples from the excluded cohort. Analyses in Chapter 8 are based on the summary statistics only and therefore we used those for the full European-ancestry cohort.

A uniform data cleaning protocol was applied to the GWAS summary statistics (Pain et al., 2021), with small variations between the two chapters that reflect differences in study design.

We retained only single nucleotide polymorphisms (SNPs), excluding any non-SNP or strand-ambiguous variants. Sampled SNPs were filtered to those present within and harmonised to the allele order of the 1000 Genomes phase 3 (1KG) European ancestry population (n = 503) reference dataset (Auton et al., 2015), For Chapter 6, under the *GenoPredPipe* analysis protocol (Pain et al., 2021), variants within the 1KG data were first restricted to those within the HapMap3 reference panel (Altshuler et al., 2010). If chromosomal positions were provided, SNPs were matched between summary statistics and the reference population by GRCh37 chromosomal position using *bigsnpr* (v1.11.6) (Privé, Aschard, Ziyatdinov, & Blum, 2018). Chromosomal positions were not available for the ALS GWAS subset used in Chapter 6 and these summary statistics were therefore harmonised to the reference population after matching variants by rsID.

If not already reported, and where possible, effective sample size (N_{eff}) was calculated from per-SNP case and control sample sizes (Grotzinger, Fuente, Privé, Nivard, & Tucker-Drob, 2023). When this could not be determined per-SNP, all variants were assigned a single N_{eff}, calculated as a sum of N_{eff} values for each cohort of the GWAS meta-analysis.

Further processing was performed where possible, excluding SNPs with imputation INFO<0.9, p-values ≤0 or >1, N_{eff} >3 standard deviations from the median N_{eff}, or an absolute minor allele frequency (MAF) difference of >0.2 between the GWAS and reference dataset.

For Chapter 6 we restricted to SNPs with a MAF >0.01 in both the GWAS and reference samples, whereas variants with MAF >0.005 were retained in Chapter 8. Different thresholds were selected to reflect differences in study aims. The PRS derived for Chapter 6

indicate genetic liabilities for disease across common variation across the genome and the MAF >0.01 threshold is typical. Meanwhile, the objective of Chapter 8 was to examine genetic variation shared between ALS and overlapping neuropsychiatric traits at associated regions of the genome. A lower MAF threshold was selected to retain additional variants in these comparisons, recognising the relevance of rare variation to traits studied.

3.3. Ethical approval

Ethical approval for analysis of genetic and molecular studies is established as part of Project MinE. Ethical approval was granted by the Trent Research Ethics Committee within the Integrated Research Application System (IRAS), reference number 08/H0405/60.

3.4. Data and code availability

Analyses presented across this thesis were conducted using the code/resources at the following locations:

- Chapter 4: https://github.com/ThomasPSpargo/adpenetrance
- Chapter 5: https://github.com/ThomasPSpargo/neuroGeneScreening
- Chapter 6: https://github.com/ThomasPSpargo/LatentClusterALS
- Chapter 7: https://sod1-als-browser.rosalind.kcl.ac.uk
- Chapter 8: https://github.com/ThomasPSpargo/COLOC-reporter

Chapter 4. Developing a novel approach to calculate genetic penetrance

4.1. Publication

The work described in this chapter was published in *Genome Medicine* (Spargo et al., 2022):

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,**,#].

[1]Maurice Wohl Clinical Neuroscience Institute, King's College London, Department of Basic and Clinical Neuroscience, 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.

[#]co-senior author

Correspondence should be addressed to alfredo.iacoangeli@kcl.ac.uk* and ammar.al-chalabi@kcl.ac.uk**

Author 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.

4.2. 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.

4.3. 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 (Langbehn, Brinkman, Falush, Paulsen, & Hayden, 2004). 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 elevates risk but does not necessarily result in its manifestation (Goldwurm et al., 2011).

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 (Dewey et al., 2014; S. S. Kalia et al., 2017; Saelaert, Mertes, Moerenhout, De Baere, & Devisch, 2019; Senol-Cosar et al., 2019).

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 (Otto & Horimoto, 2012). 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 (Minikel et al., 2016; C. F. Wright et al., 2019). Without systematic sampling, these estimates can be affected by ascertainment bias. Where large pedigrees are not available, or if disease is rare or late onset, these techniques may not be possible (Chiò et al., 2014).

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 (Iacoangeli, Al Khleifat, Jones, et al., 2019; Shatunov & Al-Chalabi, 2021). *SOD1* variants are an important cause of ALS and over 180 ALS-associated variants in this gene have been reported to date (Abel et al., 2012; Iacoangeli, Al Khleifat, Sproviero, Shatunov, Jones, Opie-

Martin, et al., 2019; Shatunov & Al-Chalabi, 2021), 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 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 (Al-Chalabi & Lewis, 2011).

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 examples, drawing upon the most common and widely studied autosomal dominant variants implicated in each disease: the p.Gly2019Ser variant of the *LRRK2* gene for PD (Goldwurm et al., 2011); variants in the *BMPR2* gene [MIM: 600799] for heritable pulmonary arterial hypertension (PAH [MIM: 178600]) (Evans et al., 2016); and variants in the *SOD1* and *C9orf72* [MIM: 614260] genes for ALS (Iacoangeli, Al Khleifat, Jones, et al., 2019; Shatunov & Al-Chalabi, 2021).

4.4. Material and methods

4.4.1. 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 (Al-Chalabi & Lewis, 2011) 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, $P(M)$:

Equation 4-1

$$P(A) = f \times P(M) + g \times P(M'),$$

letting $P(M') = 1 - P(M)$.

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:

Equation 4-2

$$P(A)^{M} = f$$

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

Equation 4-3

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

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

Equation 4-4

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

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:

Equation 4-5

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

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

Equation 4-6

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

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

Equation 4-7

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

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 (Al-Chalabi & Lewis, 2011) and extended models are equivalent.

4.4.2. 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

Equation 4-8

$$P(affected) = P(familial) + P(sporadic)\ .$$

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 4-1). 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).

Table 4-1. Valid disease state combinations and corresponding weighting factors for estimating disease state rates

The defined weighting factors are used in Step 1 of the penetrance estimation approach, as described in Figure 4-1 and Appendix A.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.

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

Under Bayes theorem (Hunink et al., 2014), 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:

Equation 4-9

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

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

Equation 4-10

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

and

Equation 4-11

$$P(M)^{pop} = M_A \times P(A)^{pop} + M_U \times (1 - P(A)^{pop})$$

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 negligible influence upon penetrance estimates (see the simulation studies in Appendix A.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 Figure 4-1 and comprehensively outlined in Appendix A.1.1.

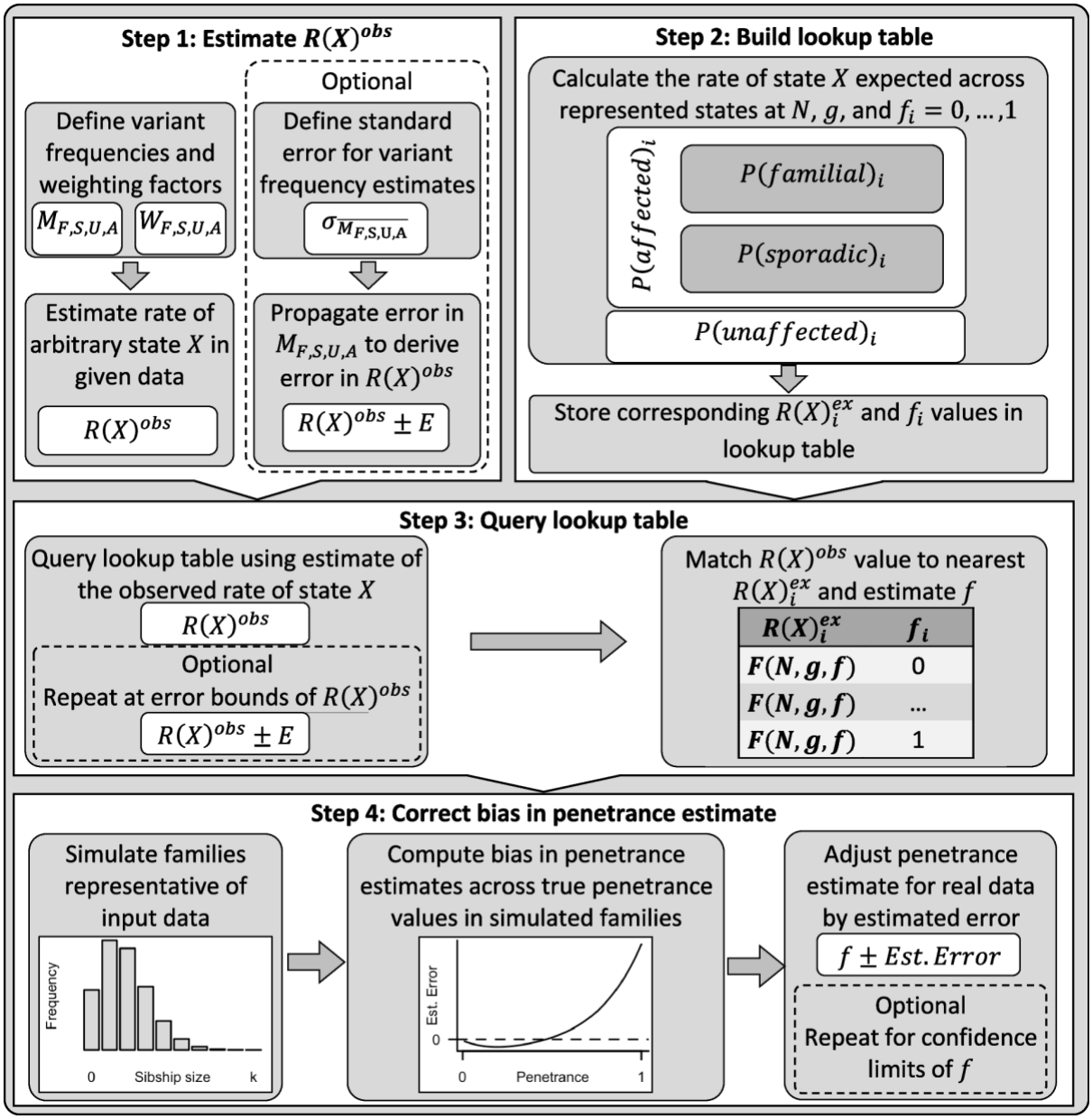


Figure 4-1. Summary of the key steps within this penetrance estimation approach

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 4-1) 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: Equation 4-5, Equation 4-6, Equation 4-7, and Equation 4-8 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 $f = 0, \dots, 1$ 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 (Hughes & Hase, 2010); Penetrance is estimated as in Steps 3 and 4 for the interval bounds. All steps within this approach are comprehensively detailed in Appendix A.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 Appendix A.1.2.2.

Appendix A.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: https://github.com/ThomasPSpargo/adpenetrance. In the GitHub repository, we additionally provide functions to: calculate g; test for equal onset variability across two groups; 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 (v1.7.3): https://adpenetrance.rosalind.kcl.ac.uk/. The web tool is further described in Appendix A.1.3 and Figure 4-2 presents an example of its usage.

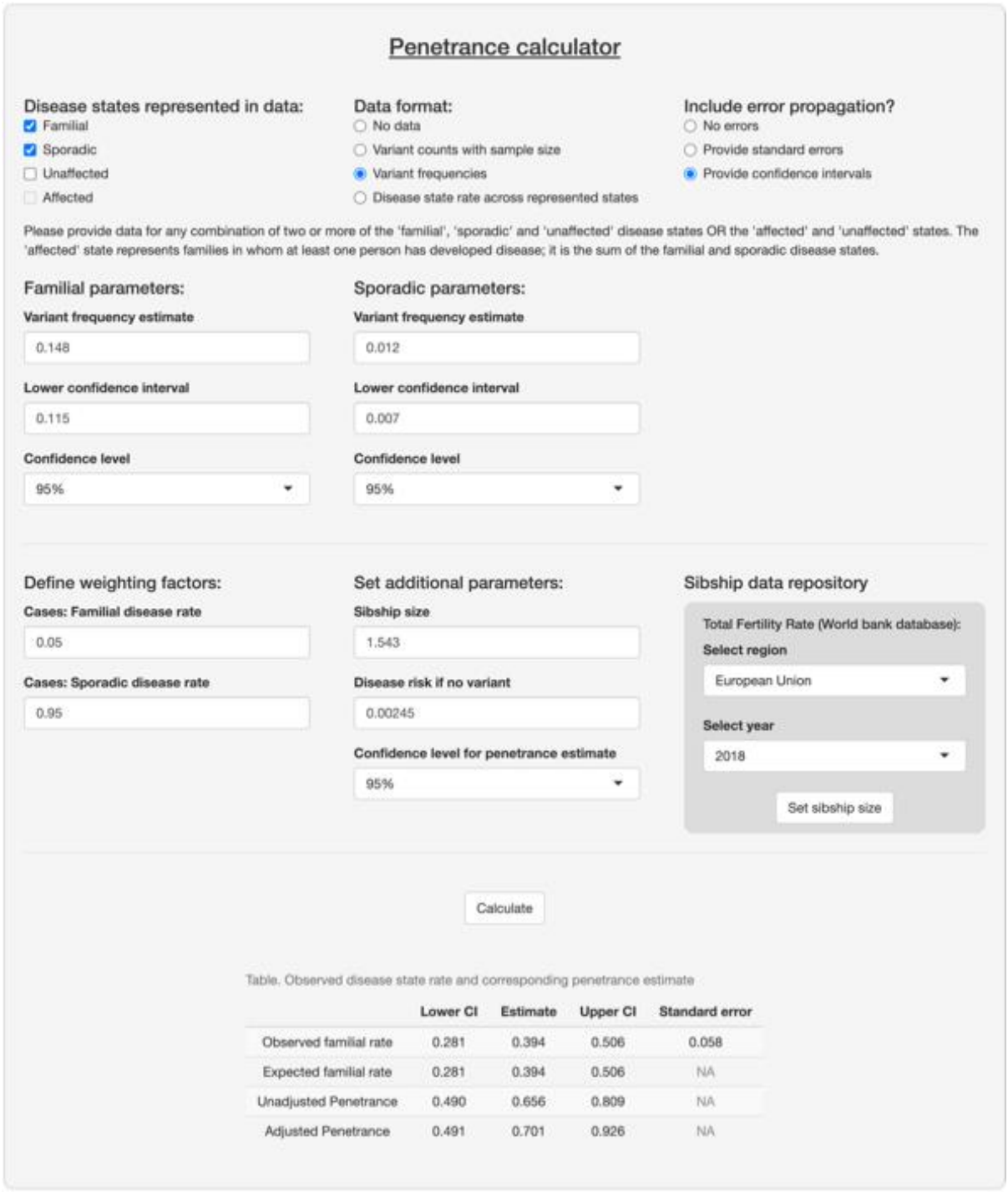


Figure 4-2. Example interface and output of the ADPenetrance web tool

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 (World Bank, 2020; Z.-Y. Zou et al., 2017).

4.4.3. Case examples

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 4-1. 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 Equation 4-9 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 (World Bank, 2020) for the world region(s) best representing the sample.

An R script permitting replication of each case study is provided within our GitHub repository: https://github.com/ThomasPSpargo/adpenetrance.

4.4.3.1. 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 $P(S|A)$ (Elbaz et al., 1999; Shino et al., 2010). $P(A)^{pop}$ was estimated as 1 in 37 (0.027), the lifetime risk of developing PD (Parkinson's UK, 2017).

We estimated $M_{F.S}$ using data aggregated from 18 European ancestry groups within a sample of 24 world populations (Healy et al., 2008). 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 (Karczewski et al., 2020), 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 Equation 4-9, 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 (World Bank, 2020) across each of the 18 European populations sampled, weighted by the proportional contribution of each population to the sample (Table A-1) (Healy et al., 2008).

Additional region-specific and joint population penetrance estimates for this variant are presented in Table A-2.

4.4.3.2. 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 (Thenappan, Ryan, & Archer, 2012).

Input parameters were defined based on only people with idiopathic (sporadic) or heritable PAH diagnoses (Evans et al., 2016). 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 (Thenappan et al., 2012). $P(A)^{pop}$ 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 (Corris & Seeger, 2020; Thenappan et al., 2012).

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 (Evans et al., 2016), 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 (Aldred et al., 2006) 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 Equation 4-9.

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 (Evans et al., 2016). The second cohort samples people exclusively from Western Europe (Aldred et al., 2006).

We therefore estimated that $N = 1.543$ in both instances, the Total Fertility Rate of the European Union in 2018 (World Bank, 2020).

4.4.3.3. 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*^{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 (Byrne, Heverin, et al., 2013; Byrne et al., 2011). $P(A)^{pop}$ was estimated as 1 in 400 (0.0025), the lifetime risk of developing ALS (Alonso et al., 2009).

We drew upon the results of two meta-analyses (Marogianni et al., 2019; Z.-Y. Zou et al., 2017) to estimate $M_{F,S}$ for *SOD1* and *C9orf72*^{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,

Equation 4-12

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

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*^{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 Equation 4-9, and letting $M_U = 0$, we estimated that $g = 0.00243$ for *SOD1*, and $g = 0.00247$ for *C9orf72*^{RE}.

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*^{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 Equation 4-9, and letting $M_U = 0$, we estimated that $g = 0.00245$ for *SOD1*, and $g = 0.00234$ for *C9orf72*^{RE}.

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 Appendix A.1.2.2) with the *checkOnsetVariability* R function provided in the associated GitHub repository (https://github.com/ThomasPSpargo/adpenetrance), 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 Figure A-4) 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 Appendix A (see Figure A-11) 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 (World Bank, 2020).

4.5. Results

Here we summarise the input data and results of the case studies modelled (see Table 4-2). Penetrance estimates are presented both when accounting for residual disease risk g among people with no variant and when g is assumed to equal 0; those accounting for g are preferred.

Table 4-2. Penetrance estimation across case studies

[a]Disease characteristics of lifetime disease risk ($P(A)_{pop}$) and proportion familial ($P(F|A)$) are used as weighting factors (per Table 4-1) and for calculating g (per Equation 4-9, 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 Table A-1) [b.1], European Union[b.2], or East Asia and Pacific[b.3] regions in 2018. (World Bank, 2020); [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 Table A-5 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.

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 (Healy et al., 2008; Karczewski et al., 2020)	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 (Evans et al., 2016)	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 (Aldred et al., 2006)	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 (Aldred et al., 2006)	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 (Aldred et al., 2006)	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 (Z.-Y. Zou et al., 2017)	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 (Marogianni et al., 2019)	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 (0.016, 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)

Estimated penetrance of the *LRRK2* gene 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 Table A-2.

The penetrance of *BMPR2* variants for PAH was estimated comparably across the two sample sets, although slightly higher within dataset one (Evans et al., 2016) than two (Aldred et al., 2006). 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*^{RE}. Table A-3 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.341T>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 defined weighting factors will have minimised this impact.

4.6. 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 Appendix A.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 (Goldwurm et al., 2011). 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 (Larkin et al., 2012). 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 Appendix A.1.2.3; Figure A-8).

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 (Andersen, 2006; Chiò et al., 2014). The widely-described p.Ala5Val (formally p.Ala4Val) variant has a recorded penetrance of 0.91 by age 70 (Cudkowicz et al., 1997). Among other variants, penetrance is typically lower (Andersen, 2006; Chiò et al., 2014). 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 (Chiò et al., 2014; Cudkowicz et al., 1997; Iacoangeli, Al Khleifat, Sproviero, Shatunov, Jones, Opie-Martin, et al., 2019). Our estimates for heterozygous inheritance of these individual variants aligned with these observations (see Table A-3) 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 (van Blitterswijk et al., 2012). 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 (Beck et al., 2013). 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 (Beck et al., 2013; Iacoangeli, Al Khleifat, Jones, et al., 2019; N. A. Murphy et al., 2017). 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 (N. A. Murphy et al., 2017). Accounting for lifetime risk of each phenotype and their respective familiality rates, people of European ancestry harbouring *C9orf72*^{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 Table A-4) (Alonso et al., 2009; Coyle-Gilchrist et al., 2016; Majounie et al., 2012; Marogianni et al., 2019; Turner et al., 2017). 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*^{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 Appendix A.1.2). 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 (Chiò et al., 2018), where harbouring *SOD1* variants is associated with a 2-step disease process, converse to the 3-step process associated with *C9orf72*^{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 4-2, Table A-2).

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 (C. F. Wright et al., 2019). 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 (C. F. Wright et al., 2019).

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 (Byrne, Heverin, et al., 2013; Vajda et al., 2017). For example, ALS and frontotemporal dementia 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 Appendix A.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 Appendix A.1.2.3; Figure A-10, Figure A-11, Figure A-12).

Point penetrance estimates have several applications, for instance, improving 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.

4.7. Conclusions

Our novel method for penetrance estimation fills an important gap in medical genetics because, making use of the available amounts of 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 examples. 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: https://github.com/ThomasPSpargo/adpenetrance and the method is available and free to use via a public webserver: https://adpenetrance.rosalind.kcl.ac.uk/.

Chapter 5. Modelling population genetic screening in rare neurodegenerative diseases

5.1. Abstract

Genomic sequencing enables rapid identification of a breadth of genetic variants. With costs decreasing and increasing genotyping accuracy, interest has risen in developing genetic screening protocols to identify pathogenic variants indicating increased disease risk in sequencing data. These have potential utility for identifying disease liabilities. However, usefulness is constrained when a person's probability of being affected by a rare disease remains strikingly low despite a positive genetic test result. The problem is recognised among statisticians and statistical geneticists but less well understood by clinicians and researchers who will act on these results. Here, in a Bayesian framework, we explore the probability of subsequent disease following genetic screening for two contrasting neurological conditions, Huntington's disease and amyotrophic lateral sclerosis, comparing with screening for phenylketonuria which is well established. We discuss characteristics affecting post-test disease probabilities and highlight considerations for organising and interpreting genomic results from population screening for rare neurological diseases.

5.2. Background

Bioinformatic tools for identifying a wide range of genetic variants in next-generation sequencing data have become increasingly reliable and fast (Iacoangeli, Al Khleifat, Sproviero, Shatunov, Jones, Morgan, et al., 2019). Single nucleotide variants (SNVs) and small insertions and deletions (indels) are now genotyped with over 99% accuracy (Illumina, 2019; Z. Li, Wang, & Wang, 2018; Wenger et al., 2019), while genotyping of larger, structural, variants is often less reliable (Kosugi et al., 2019). For clinical purposes, the analytical validity of sequencing for small genetic variations is a solved problem. Issues of analytical validity of sequencing for structural variants and the downstream challenges of clinical validity and utility remain.

Widened availability of high-quality genomic sequencing data has enabled rapid development in understanding of the role of genetic factors in various phenotypes (Bycroft et al., 2018; Davey et al., 2011; J. J. Lee et al., 2018; Rhoades, Jackson, & Teng, 2019).

Genetic testing has therefore become valuable for healthcare and gained popularity among consumers who can access genetic testing directly, effectively testing variation across the genome without advice from clinicians trained to guide interpretation of results (Hörster et al., 2017; Majumder, Guerrini, & McGuire, 2021; Perrone et al., 2018).

Interest has recently risen among governing bodies in developing protocols for population-wide genetic screening; such initiatives are being rolled out in the United Kingdom and considered in the United States of America (Adhikari et al., 2020; Centers for Disease Control and Prevention, 2021; Dickinson et al., 2018; *Genome UK: The future of healthcare*, 2020; Jansen et al., 2017; Moorthie et al., 2021; Murray et al., 2019). Genetic screening involves testing a population for genetic variants indicative of risk for specific diseases to identify people with either higher predisposition of developing that disease or the potential to pass it on to their offspring. Such an approach utilises the efficiency and breadth of modern sequencing techniques to evaluate multiple genes across the genome with known associations to selected traits. Screening can be contrasted with 'targeted' tests, which are those performed because of some suggestion that a person may harbour a certain genetic variant (e.g., symptoms of an associated disease). Although screening is relevant to liabilities ranging between monogenic and polygenic (cf. (Moorthie et al., 2021; Pain, Gillett, Austin, Folkersen, & Lewis, 2022)), we focus here on screening for pathogenic variants with monogenic associations to rare diseases, particularly as applied to neurodegenerative disorders, since that is the most usual clinical context.

Although no widespread implementation of a genetic screening protocol currently exists internationally, comparable metabolic screening approaches, testing neonates for metabolites that are markers of various metabolic diseases, are routine in many countries (Adhikari et al., 2020; Loeber et al., 2012; Tarini, Christakis, & Welch, 2006). Positive metabolic test results are typically validated via confirmatory secondary testing, including targeted genetic tests (Hörster et al., 2017; Rinaldo, Zafari, Tortorelli, & Matern, 2006; Southern et al., 2007).

When applied to outcomes uncommon in the general population, such as rare diseases, genetic screening may have limited utility. Utility can be assessed by the extent to which

action can be taken following a positive test result indicating the presence of a disease marker: its actionability (Hunter et al., 2016; Rehm et al., 2015). One key tenet of actionability is the probability of having or later developing a disease following a positive test genetic result. Yet post-test disease probability can be strikingly low where disease risk prior to testing is low, as would be inherent to population screening for rare outcomes (Biesecker, 2019).

Bayesian inference, which is routine within clinical decision making (Hunink et al., 2014), can be used to understand post-test (also known as 'posterior') disease risk. Under this logic, disease probability following a test can be inferred given existing knowledge of the probability of other relevant events. Key considerations to understanding post-test disease risk, beyond pre-test (also known as 'prior') disease risk, include the penetrance of the genetic marker, the frequency of the marker among people displaying disease symptoms, and the sensitivity and specificity of the testing protocol (analytical validity). This reasoning is therefore highly relevant to screening for rare neurodegenerative diseases, for which genetic causes are typically rare variants of variable penetrance (Bertram & Tanzi, 2005).

Accordingly, this chapter overviews important considerations for genetic screening of rare neurodegenerative disorders in the context of several relevant case studies. Considering conditional disease probabilities is not novel but the relevance of these principles must be emphasised in the genomic medicine era, since many results that could be obtained within large-scale indiscriminate testing of genetic variation across a population will not be actionable but may be readily misinterpreted. We modelled genetic screening for two rare neurodegenerative diseases (Huntington's disease (Langbehn et al., 2004), HD, and amyotrophic lateral sclerosis (R. H. Brown & Al-Chalabi, 2017), ALS) by applying Bayesian logic to examine the probability of disease following a positive test result for a dichotomous genetic marker of liability towards that disease. We additionally modelled screening for phenylketonuria (PKU) (Hillert et al., 2020), to compare genetic and metabolic screening.

5.3. Bayesian framework

We use Bayesian logic to calculate the probability of having or subsequently manifesting disease *D* following a test result indicating presence or absence of marker *M*. *M* is

associated with increased liability of *D*, and positive test result *T* indicates presence of *M*, while negative test result *T'* indicates its absence, denoted *M'*. Disease risk following a positive result is denoted *P(D|T)*, using *P(D|T')* for a negative result. In genetic testing, *M* represents harbouring a genetic variant associated with increased liability of *D*.

The mathematical principles followed are widely applied within evidence-based medicine (Hunink et al., 2014); Chapter 3.1 summarises the underlying logic. We assume that all model parameters represent binary events. This was a necessary simplification of reality as, for example, disease severity is not considered.

We estimate *P(D|T)* and *P(D|T')* using the following input parameters:

- *P(D)*, probability of a person having or later manifesting disease *D* prior to testing
- *P(M|D)*, frequency of marker *M* among those affected by *D*
- *P(D|M)*, penetrance, probability of having or later manifesting *D* for people harbouring *M*
- *P(T|M)*, sensitivity (true positive rate) of the testing procedure for detecting *M*
- *P(T'|M')*, specificity (true negative rate) of the testing procedure for identifying the absence of *M*

Bayes theorem is applied to derive the total probability of harbouring disease marker *M*,

Equation 5-1

$$P(M) = \frac{P(D) \times P(M|D)}{P(D|M)},$$

and of disease *D* manifesting given the absence of *M*,

Equation 5-2

$$P(D|M') = \frac{P(D) \times \big(1 - P(M|D)\big)}{\big(1 - P(M)\big)}\,.$$

We next calculate the total probability of positive test result *T*, *P(T)*, according to the sensitivity and specificity of the test and the probabilities of *M* being present versus absent:

Equation 5-3

$$P(T) = P(T|M) \times P(M) + \big(1 - P(T'|M')\big) \times \big(1 - P(M)\big)\,.$$

Bayes theorem is then used to derive the probabilities of *M* being present after positive,

Equation 5-4

$$P(M|T) = \frac{P(M) \times P(T|M)}{P(T)} ,$$

and negative,

Equation 5-5

$$P(M|T') = \frac{P(M) \times \left(1 - P(T|M)\right)}{(1 - P(T))} ,$$

results.

The probabilities of manifesting disease *D* (which has conditional independence from *T* when considering *M*) after receiving positive test result *T*,

Equation 5-6

$$P(D|T) = P(D|M) \times P(M|T) + P(D|M') \times \left(1 - P(M|T)\right) ,$$

and negative test result *T'*,

Equation 5-7

$$P(D|T') = P(D|M) \times P(M|T') + P(D|M') \times \left(1 - P(M|T')\right) ,$$

can then be determined.

5.4. Case studies

Most input parameters were defined using data from published literature and online databases. Sensitivity (*P(T|M)*) and specificity (*P(T'|M')*) were defined by performance benchmarks for variant calling with state-of-the-art genomic sequencing techniques specialised for genotyping particular variant types (see Appendix B.1.1; Table B-1). Although analytical accuracy will vary across the genome and other sources of error exist, these heuristics are sufficient for our purposes.

Table 5-1 presents the parameter estimates and post-test disease risks calculated across various scenarios. A comprehensive description of parameter ascertainment, including penetrance estimation, is given in Appendix B.1.2; Table B-2 summarises the assumptions and corresponding reality.

5.4.1. Case 1 – Huntington's disease

HD is a late-onset Mendelian disease with autosomal dominant inheritance caused by a trinucleotide, CAG, short tandem repeat expansion (STRE) in the *HTT* gene (OMIM: 613004). We let *M* be a CAG expansion of >40 repeat units, which would have complete penetrance in a normal lifespan (Langbehn et al., 2004).

We modelled two scenarios for this example, (1) as in genetic screening, defining pre-test disease probability by baseline risk of HD in a general population and (2) as a targeted test, considering pre-test disease probability for an individual whose parent harbours the fully penetrant *HTT* STRE and who has a 0.5 probability of inheriting *M* (we have not modelled genetic anticipation (Paulson, 2018)).

5.4.2. Case 2 – amyotrophic lateral sclerosis

ALS is a late-onset disease with locus and allelic heterogeneity; genetic associations with risk and phenotype modification range between monogenic and polygenic. Variants in at least 40 genes are associated with ALS (R. H. Brown & Al-Chalabi, 2017; Iacoangeli, Al Khleifat, Sproviero, Shatunov, Jones, Opie-Martin, et al., 2019; Marogianni et al., 2019; Shatunov & Al-Chalabi, 2021; Z.-Y. Zou et al., 2017). *SOD1* (OMIM: 147450) and *C9orf72* (OMIM: 614260) are the most frequently implicated genes, where variants of each account for fewer than 10% of cases. Autosomal dominant inheritance is typical for most people with a known genetic disease cause.

We modelled several definitions of markers for liability towards ALS, drawing from three of the most frequently implicated and well-researched ALS genes: *SOD1, C9orf72,* and *FUS* (OMIM: 137070). *SOD1*- and *FUS*-linked ALS is typically attributed to SNVs and many variants supposed to cause ALS have been reported in these two genes, with varying strength of supporting evidence (Abel et al., 2012). The pathogenic form of *C9orf72* is a hexanucleotide, GGGGCC, STRE associated principally with the onset of either one or both of ALS and frontotemporal dementia (DeJesus-Hernandez et al., 2011; Renton et al., 2011). For known variants in these genes, penetrance is typically incomplete; examples include

~90-100% penetrance for ALS associated with *SOD1* p.A5V and ~45% for the *C9orf72* STRE (Andersen, 2006; Chiò et al., 2014; Cudkowicz et al., 1997); see also Chapter 4.

The definitions of *M* modelled in this case study were:

- *SOD1* (all) – *M* includes any rare variant reported in people with ALS of European ancestry contained within the meta-analysis sample set from which the variant frequencies were derived (see Appendix B.1.2.2) (Z.-Y. Zou et al., 2017).
- *SOD1* (A5V) – *M* represents the pathogenic *SOD1* p.A5V variant, one of the most common *SOD1* variants among North American ALS populations, characterised by high penetrance (Cudkowicz et al., 1997; Saeed et al., 2009).
- *FUS* (all) – *M* includes any rare variant reported in people with ALS of European ancestry contained within the meta-analysis sample set from which the variant frequencies were derived (see Appendix B.1.2.2) (Z.-Y. Zou et al., 2017).
- *FUS* (ClinVar) – *M* includes any of 21 *FUS* variants reported as pathogenic or likely pathogenic for ALS within ClinVar and present within databases of familial and sporadic ALS (see Table B-3) (ALS Variant Server; Landrum et al., 2018; van der Spek et al., 2019).
- *C9orf72* – *M* represents a pathogenic *C9orf72* STRE of 30≤ repeat GGGGCC units within the first intron of the *C9orf72* gene.

In the *SOD1* (all) and *FUS* (all) scenarios, *M* encompasses variants classed as pathogenic, likely pathogenic, and variants of unknown significance (Richards et al., 2015). It is appropriate to model these scenarios since many variants of unknown significance in ALS-implicated genes have a high probability of being deleterious and should not necessarily be ignored because they lack a formal pathogenic or likely pathogenic classification (P. R. Mehta et al., 2022).

For the *C9orf72* marker, we modelled two testing scenarios: (1) genetic screening with sensitivity and specificity defined by performance of existing tools for genotyping STREs in sequencing data; (2) using repeat-primed polymerase chain reaction with amplicon-length analysis (Akimoto et al., 2014) as a secondary test to validate a positive result from screening via sequencing in scenario 1.

5.4.3. Case 3 – phenylketonuria

PKU is an autosomal recessive disease with infantile onset, caused by variants in the *PAH* gene (OMIM: 612349) (Hillert et al., 2020); most variants pathogenic for PKU are SNVs. *M* represents being homozygous or compound heterozygous for any of the three most common *PAH* variants recorded in European populations of people with PKU: p.Arg408Trp, c.1066-11G>A, p.Arg261Gln (Hillert et al., 2020).

We modelled two testing scenarios for PKU: (1) genetic screening, with pre-test disease probability defined per the baseline population risk of PKU, and (2) secondary testing to confirm positive results obtained using tandem mass spectrometry (Schulze et al., 2003) as in established metabolic screening protocols (see Appendix B.1.2.3).

Table 5-1. Input parameters and disease risk estimates following testing for all case study scenarios

*Parameter ascertainment is comprehensively described within Appendix B.1.2. HD = Huntington's disease; ALS = amyotrophic lateral sclerosis; PKU = phenylketonuria. SNV = single nucleotide variant; STRE = short tandem repeat expansion. §Estimates are based on populations of predominantly European ancestry – 95% confidence intervals shown for newly derived estimates in the ALS case study; *includes FUS variants classified as pathogenic or likely pathogenic for ALS in the ClinVar database and recorded within ALS population databases (see Table B-3, Table B-4); ¶defined by variant calling performance benchmarks of tools for genotyping in sequencing data by variant type (see Table B-1) and, where marked †, by aggregate laboratory accuracy for genotyping C9orf72 STRE with repeat-primed polymerase chain reaction and amplicon-length analysis (Akimoto et al., 2014). ΩRisk following positive results in primary screening and confirmatory tests relative to a negative screening result (probability of PKU given a negative metabolic screening result is approximated as 1×10^{-6}*

Case study	Gene containing marker (case study scenario)	Variant type	Pre-test disease probability§	Marker frequency in people affected§	Penetrance§	Test sensitivity ¶	Test specificity ¶	Disease risk after positive test	Disease risk after negative test	Relative disease risk after positive rather than negative test
-	-	-	P(D)	P(M|D)	P(D|M)	P(T|M)	P(T'|M')	P(D|T)	P(D|T')	-
1: HD	*HTT* (screening)	STRE	0.000410	1.000	1.000	0.990	0.900	0.00404	0.00000456	887
	HTT (targeted)	STRE	0.500	1.000	1.000	0.990	0.900	0.908	0.011	82.7
2: ALS	*SOD1* (all)	SNV	0.00333	0.0188 (0.0138, 0.0238)	0.701 (0.491, 0.926)	0.9996	0.9995	0.109	0.00327	33.3
	SOD1 (A5V)	SNV	0.00333	0.000529 (4.43×10^{-5}, 0.00101)	0.91	0.9996	0.9995	0.00683	0.00333	2.05
	FUS (all)	SNV	0.00333	0.00425 (0.0023, 0.0062)	0.579 (0.291, 0.884)	0.9996	0.9995	0.0302	0.00332	9.09
	FUS (ClinVar*)	SNV	0.00333	0.00251 (0.00125, 0.00377)	0.536 (0.211, 0.877)	0.9996	0.9995	0.0194	0.00333	5.84
	C9orf72	STRE	0.00333	0.0635 (0.0538, 0.0732)	0.439 (0.358, 0.520)	0.990	0.900	0.00519	0.00313	1.66
	C9orf72 (positive sequencing screening confirmation)	STRE	0.0052	0.0635 (0.0538, 0.0732)	0.439 (0.358, 0.520)	0.95†	0.98†	0.0198	0.00489	4.06 (6.35 Ω)
3: PKU	*PAH* (screening)	SNV	0.000100	0.743	0.892	0.9996	0.9995	0.127	0.0000257	4,961
	PAH (positive metabolic screening confirmation)	SNV	0.167	0.743	0.892	0.9996	0.9995	0.889	0.0497	17.9 (889,000 Ω)

5.5. Post-test disease probability

5.5.1. Screening versus diagnostic testing

Across case studies, we showed low probability of disease following positive results within contextually-blind genetic screening scenarios; risk ranged between 12.7% and 0.4% (see Table 5-1). Disease risk was always negligible following a negative test result, indicating absence of the tested variant, as would be seen for any rare trait.

The HD case study illustrates the distinction between contextually-blind screening and diagnostic testing for rare diseases: following a positive test result, lifetime HD risk was high (90.8%) using targeted testing but low (0.4%) in screening. This difference reflects that, unlike screening, targeted testing is performed based on some indication of a person's elevated disease risk (e.g., existing disease symptoms or family history) compared to other population members. Accordingly, pre-test disease probability is greater. Inherently low pre-test disease probability will be a pervasive issue in screening for rare diseases.

5.5.2. Relative risk and secondary testing

The utility of a test for identifying at-risk individuals can be examined based on relative disease risk following positive rather than negative test results: utility is limited when risk is only minimally greater for people testing positive rather than negative. This is observed in the ALS *C9orf72* case study, where risk was only 1.7 times greater (Table 5-1) following a positive screening result from sequencing (within the modelled protocol) alone, despite this variant being the most common genetic cause of ALS (Z.-Y. Zou et al., 2017).

We additionally observed 6.35 times greater ALS risk for a person testing positive on both screening for the *C9orf72* marker and a secondary test performed because of the initial result than for a person testing negative on the initial screening. This increased relative risk reflects that a person testing positive on two independent measures of disease risk has greater absolute probability of disease than after the initial screening result alone.

We emphasise that secondary testing is important to increase certainty in positive test results. The PKU case study demonstrates its potentially sizable impact. A positive screening

result using the established metabolic approach alone indicated 16.7% PKU risk, versus 12.7% within genetic screening (as presently modelled). The metabolic marker, which is universal across people with PKU and indicates existing disease manifestation, eclipses need for genetic screening for PKU, marked by variants with incomplete penetrance that are not present for all people with PKU. However, the genetic test remains useful for validating the positive metabolic screening result (Adhikari et al., 2020): probability of PKU following a confirmatory genetic test conducted on the basis of a positive metabolic screening result was 88.9%.

The overall benefit of secondary testing will however differ by scenario. In the case of ALS, the risk remained moderate (~2%) despite two positive test results for the *C9orf72* marker.

5.5.3. Constraints upon post-test disease probability

Figure 5-1, Figure 5-2, and Figure 5-3 illustrate how the post-test disease probability reduces as the probability of any test, disease, or marker characteristic decreases. Sensitivity and specificity critically constrain certainty about post-test disease risk, and this role is amplified as the other parameter probabilities decrease. Figure 5-1 particularly demonstrates the increased role of specificity in rarer diseases, where disease risk following a positive test result will be moderated only by penetrance in a protocol of perfect specificity.

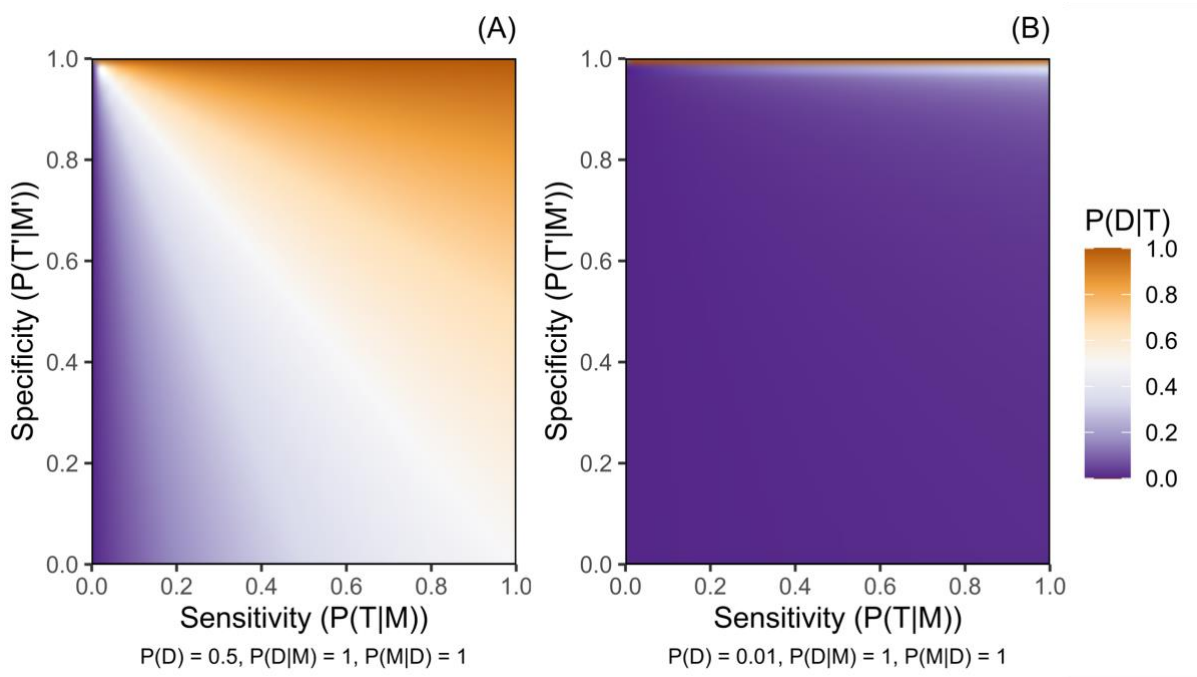


Figure 5-1. Probability of a disease given a positive genetic test result for a marker of increased disease risk (P(D|T)) according to the sensitivity (P(T|M)) and the specificity (P(T'|M')) of the testing protocol

***Panel A** presents this for a disease with pre-test probability (P(D)) of 0.5, while **panel B** presents a disease with P(D) of 0.01. Penetrance is complete (P(D|M) = 1) and variant M is harboured by all people with disease D (P(M|D) = 1) in both panels.*

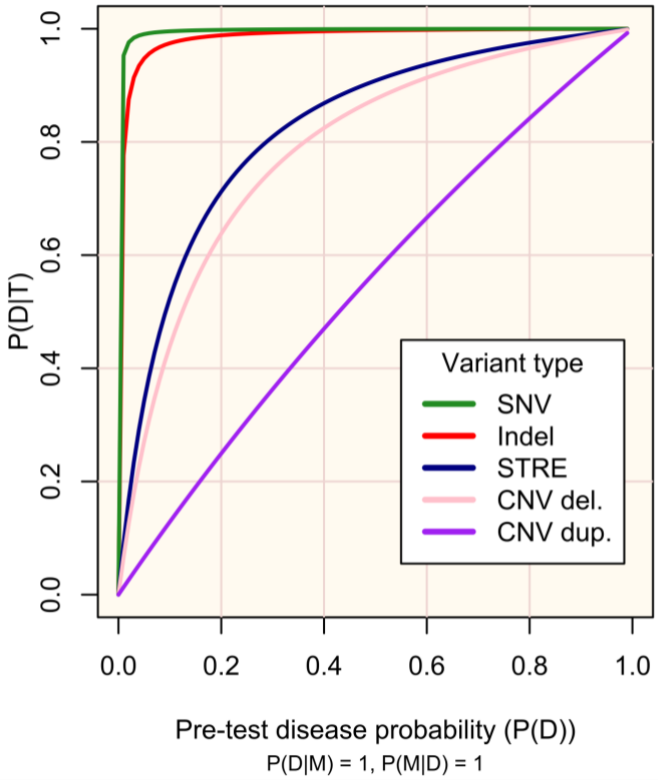


Figure 5-2. Probability of disease D following a positive genetic test result for marker M (P(D|T)) according to pre-test disease probability (P(D))

M occurs in all people with D (P(M|D) = 1) and has complete penetrance (P(D|M) = 1)). Plot lines are defined according to sensitivity (P(T|M)) and specificity (P(T'|M')) of existing protocols for genotyping variant types (see Supplementary Materials 2): single nucleotide variant (SNV), P(T|M) = 0.9996, P(T'|M') = 0.9995; small insertion or deletion (Indel), P(T|M) = 0.9962, P(T'|M') = 0.9971; short tandem repeat expansion (STRE), P(T|M) = 0.99, P(T'|M') = 0.90; copy number variant (CNV) – del. (deletion), P(T|M) = 0.289, P(T'|M') = 0.959; CNV – dup. (duplication), P(T|M) = 0.1020, P(T'|M') = 0.9233.

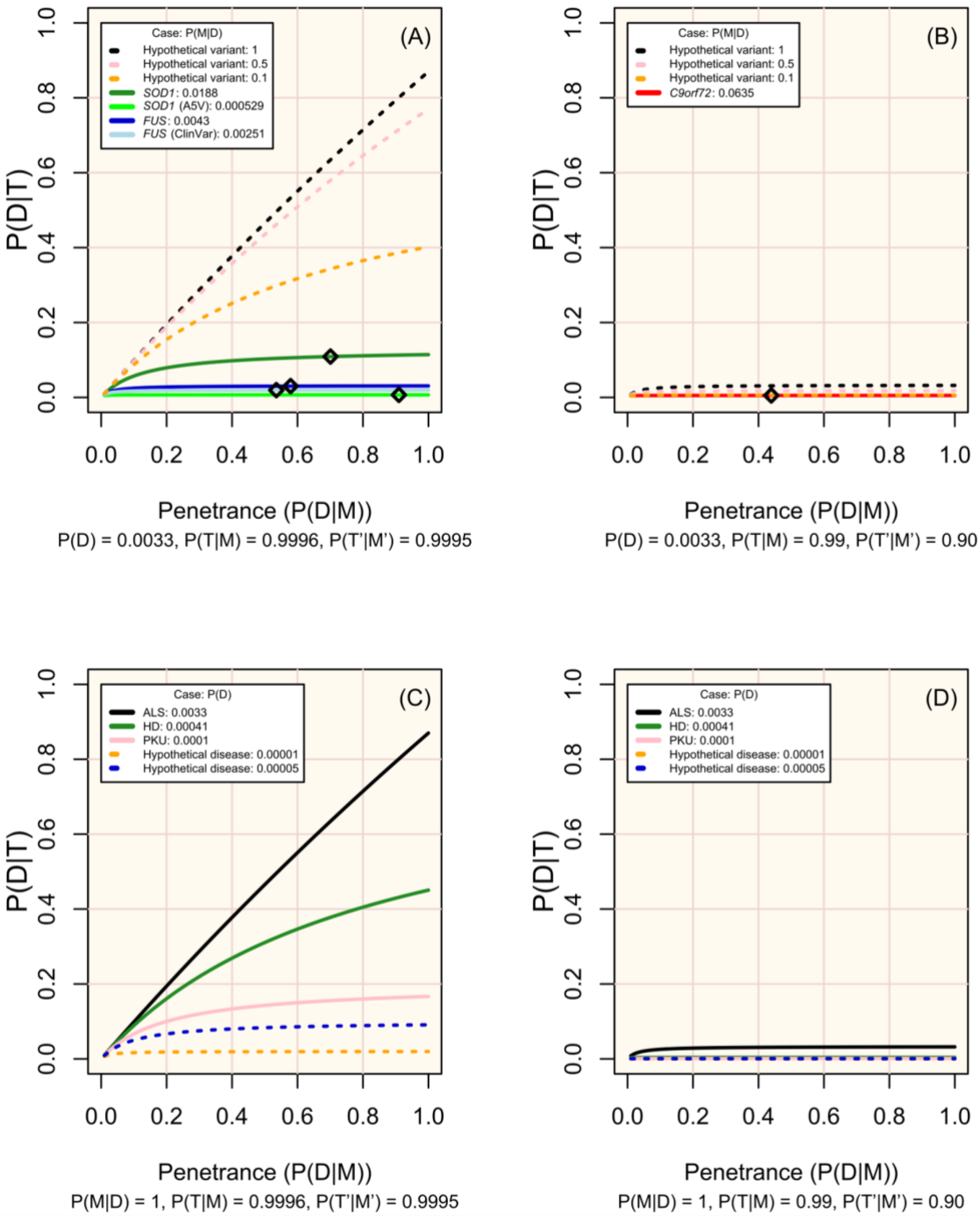


Figure 5-3. Change in disease risk following a positive test result for a marker of increased disease risk (P(D|T)) according to penetrance (P(D|M))

***Panels A and B**: modelled and hypothetical markers of ALS which differ in frequency across people affected by ALS (P(M|D)), where pre-test disease probability (P(D)) is 0.0033 and diamonds mark the penetrance estimated for non-hypothetical variants (see Table 5-1). **Panels C and D**: diseases in which P(M|D) = 1 and with P(D) set in line with the modelled case studies or a hypothetical rare disease. Sensitivity and specificity are defined according to the performance of tools for genotyping single nucleotide variants in sequence data in A and C, and of short tandem repeat expansions in B and D (see Table B-1; Figure 5-2).*

As is well-recognised, high sensitivity and specificity are essential to maximise utility of testing. The respective trade-offs between prioritising each of these must be regarded: high sensitivity is required to detect the tested marker, while high specificity increases confidence in positive results. Established screening protocols prioritise high sensitivity to maximise detection of at-risk individuals, with confirmatory secondary testing being vital to minimise false-positive results (Hörster et al., 2017; Rinaldo et al., 2006; Southern et al., 2007).

Since the characteristics of diseases and associated variants are all pre-determined within a population, disease markers (i.e., variants) screened for should be chosen carefully. The most useful variants will be those more prevalent among people affected, of high penetrance, and which can be genotyped with high sensitivity and specificity (see Table 5-1, ALS case study; Figure 5-3, panels A and B).

5.6. Practical implementation of genetic screening

5.6.1. Marker selection

Before a marker is used in screening, its relevance across people must be evaluated, recognising particularly that this may vary by ancestry. For instance, particular variants may be less common or only present in certain populations, and penetrance can also vary between them (cf. (Iacoangeli, Al Khleifat, Sproviero, Shatunov, Jones, Opie-Martin, et al., 2019; A. J. Lee et al., 2017; Saeed et al., 2009)). Screening protocols must therefore account for these differences to prevent systemic inequalities, especially for minority populations which are often under-studied and therefore have limited genetic information available.

Regard must be given to the clinical interpretability of selected markers. We illustrated several approaches to defining markers in the ALS case study. Within the *SOD1* (all) scenario, disease risk is marked by an aggregation of putatively deleterious *SOD1* variants. Without curation, the relationship to disease will likely vary across these. For instance, and as observed in Chapter 4, the p.I114T *SOD1* variant has substantially lower penetrance than p.A5V, and many potentially relevant variants have unknown significance (Abel et al., 2012; Landrum et al., 2018; P. R. Mehta et al., 2022). Curation could involve defining a positive

result as presence of one of a range of variants with sufficient evidence to be designated pathogenic (Richards et al., 2015), as in the *FUS* (ClinVar) scenario, or as harbouring a specific variant, as in the *SOD1* (A5V) scenario.

De novo variants and variants of unknown significance present a substantial challenge for screening since they will frequently be identified, yet must be set aside until variant interpretation is possible despite potentially being deleterious (Murray et al., 2018; Murray et al., 2019). PKU demonstrates the scale of this issue for rare diseases with multiple implicated variants, as 55% of deleterious *PAH* genotypes are observed uniquely (Hillert et al., 2020). As genetic discovery and rich population datasets continue to expand, this challenge should decrease somewhat over time (C. F. Wright et al., 2019).

5.6.2. Utility over time and actionability

As genetic screening is possible from birth, while non-genetic methods may not be, age of viability for available screening methods should be evaluated. For late-onset diseases, early genetic screening may enable provision of preventative treatments to at-risk individuals or close monitoring for prodromal disease markers. For instance, rapid eye movement sleep behaviour disorder is a prodromal feature for Parkinson's disease (Ferini-Strambi, Marelli, Galbiati, Rinaldi, & Giora, 2014) and monitoring of at risk individuals identified within genetic screening may enable early intervention. The influence of time (e.g., relative to stage of disease) upon treatment viability and effectiveness must also be considered. For example, genetic therapy has potential utility for preventing or delaying onset of degenerative disorders (Amado & Davidson, 2021); treatment benefit may be greatest when this is given early.

The ultimate benefit of early identification of disease risk through genetic screening is contingent upon the actionability of the result. A framework of actionability (Hunter et al., 2016; Rehm et al., 2015), shown to align with laypersons' views on treatment acceptability (Paquin et al., 2019), states that actionability is determined by: disease likelihood and severity, intervention effectiveness in disease minimisation or prevention, and the consequence of the intervention to a person and risk if not performed. Each of these elements are critical considerations when selecting traits and markers for inclusion within

genetic screening protocols and for informing the clinical interpretation of test results. Efforts to compile relevant information about treatable genetic disorders are ongoing; the value of these repositories will expand as understanding of treatable diseases improves (Bick et al., 2021; Rehm et al., 2015).

5.7. Limitations

The Bayesian logic in the case studies simplifies genotype-phenotype relationships and cannot address all considerations relevant to clinical genetic testing. Variable phenotype expressivity is not considered despite being common. Other factors include: polygenicity and oligogenicity, pleiotropy, the role of genetic and environmental modifying factors, and that of additive genetic effects in recessive conditions and heterozygous carriers of pathogenic variants. Such influences can fundamentally impact both the probability that a disease will manifest and its severity after onset. For instance, although spinal muscular atrophy is caused by partial or complete biallelic deletion of the *SMN1* gene, having additional copies of *SMN2*, a *SMN1* homologue, reduces disease expressivity by mitigating loss of *SMN1* protein function (Wadman et al., 2020). Any results obtained within a genetic screening protocol must be interpreted within the wider context of that disease and its modifiers (Escott-Price & Schmidt, 2021). This can be problematic in neurodegenerative diseases, such as ALS, when their architecture is not yet fully understood.

5.8. Conclusion

We have shown that risk following a positive screening test result can be strikingly low for rare neurological diseases. Accordingly, to maximise the utility of screening, we suggest prioritising protocols of very high sensitivity and specificity, careful selection of markers for screening, giving regard to clinical interpretability, and secondary testing to confirm positive findings.

A key advantage of a genetic screening approach for late-onset diseases is that these markers can be examined across the lifespan. Hence, positive test results could be useful for targeting people for prevention, and for monitoring of prodromal features.

Although not new to consider disease risks within a Bayesian context, it is important to stress the considerations raised here at a time when governments evaluate implementation of genomic sequencing for population screening and as access to genetic testing outside healthcare settings increases. While genetic screening has many potential benefits, the limitations of such an approach should be properly understood. Policy makers must consider the impact of a positive test result on large numbers of people that will never develop a given disease, a particularly salient issue for late-onset diseases. Although not the present focus, the substantial ethical and social considerations raised in conjunction to screening must also be regarded (Dickinson et al., 2018; Howard et al., 2015; McCusker & Loy, 2017; Ross & Clayton, 2019).

Chapter 6. Identifying biological subtypes of ALS with latent class clustering analysis

6.1. Abstract

Background

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease characterised by a highly variable clinical presentation and multifaceted genetic and biological bases that translate into great patient heterogeneity. The identification of homogeneous subgroups of patients could favour the development of effective treatments, healthcare, and clinical trials. We aimed to identify and characterise homogenous clinical subgroups of ALS, examining whether they represent underlying biological trends.

Methods

Latent class clustering analysis, an unsupervised machine-learning method, was used to identify homogenous subpopulations in 6,523 people with ALS from Project MinE, using widely collected ALS-related clinical variables. The clusters were validated using 7,829 independent people from STRENGTH. We tested whether the identified subgroups were associated with biological trends in genetic variation across genes previously linked to ALS, polygenic risk scores for risk of ALS and related neuropsychiatric traits, and in expression data from post-mortem motor cortex samples.

Results

We identified five ALS subgroups based on patterns in clinical data which were general across international datasets. Distinct genetic trends were observed for variants in the *SOD1* and *C9orf72* genes, and across genes implicated in biological processes relevant for ALS. Polygenic risk scores for risk of ALS, schizophrenia and Parkinson's disease were also associated with distinct clusters. Gene expression analysis identified different altered biological processes across clusters.

Conclusion

ALS subgroups characterised by highly distinct clinical presentations were discovered and validated in two large independent international datasets. Such groups were also characterised by different underlying genetic architectures and biology. Our results showed

that data-driven patient stratification into more clinically and biologically homogeneous subtypes of ALS is possible and could help developing more effective and targeted approaches to the medical and biological study of ALS.

6.2. Background

Amyotrophic lateral sclerosis (ALS) is a devastating neurodegenerative disease characterised by progressive neuromuscular degeneration leading to death, typically from respiratory failure within three years of onset (R. H. Brown & Al-Chalabi, 2017). The disease affects upper and lower motor neurons in the brain and spinal cord and has an estimated lifetime risk of between 1 in 300-400 people.

ALS is clinically defined, yet its clinical presentation varies greatly. The mean age of onset is around 60 years although people may develop ALS at almost any adult age (Ryan et al., 2019). ALS normally leads to death within 3-5 years of onset, however, in some patients it can occur within a year of onset, and 5-10% of people live for over 10 years (Al-Chalabi & Hardiman, 2013; Chiò et al., 2009; Juneja, Pericak-Vance, Laing, Dave, & Siddique, 1997b). Between 60 and 70% of people first develop symptoms in the spinal innervated muscles, with others having bulbar, mixed, or (in about 3%) respiratory onset (Masrori & Van Damme, 2020; Ryan et al., 2019; Shellikeri et al., 2017; Shoesmith et al., 2007). The extent of involvement of upper and lower motor neurons varies, ranging between a pure upper motor neuron (primary lateral sclerosis; PLS) and pure lower motor neuron phenotype (progressive muscular atrophy; PMA), with most presentations being a mixture of the two (Al-Chalabi et al., 2016; Finegan, Chipika, Shing, Hardiman, & Bede, 2019). An overlap with frontotemporal dementia (FTD) is also recognised, with a joint diagnosis for up to 15% of people in some studies and cognitive dysfunction in around 50%, according to disease stage (Crockford et al., 2018; Zucchi et al., 2019).

The biological landscape of ALS is similarly heterogenous, with recognised monogenic, oligogenic, and polygenic contributions to disease (R. H. Brown & Al-Chalabi, 2017; Emily P McCann et al., 2021; McLaughlin et al., 2017; Restuadi et al., 2022; Shatunov & Al-Chalabi, 2021; van Rheenen et al., 2021). Known genetic variation explains disease for around 15-20% of people, implicating variants in over 40 genes as causal for or modifiers of ALS (Chia

et al., 2018; P. R. Mehta et al., 2022; Shatunov & Al-Chalabi, 2021). Likewise, ALS is associated with disruption to various biological processes, including cytoskeletal transport, RNA function, autophagy, and proteostasis (R. H. Brown & Al-Chalabi, 2017; Masrori & Van Damme, 2020; Weishaupt et al., 2016).

Unfortunately, the efficacy of existing treatments for ALS is limited; the most effective drug therapies extend life expectancy from onset by no more than a few months (Lacomblez et al., 1996; van Eijk et al., 2020). This poor efficacy may reflect the heterogeneity of the disease, and therefore more effective treatments may be discovered within a precision medicine framework. Existing research supports this hypothesis. For instance, treatment with lithium appears to extend survival trajectories specifically among people with an *UNC13A* variant (van Eijk et al., 2017). Further research examining the utility of gene therapies that aim to offset aberrant gene function associated with specific genetic variation is ongoing (Amado & Davidson, 2021). This is valuable for *SOD1*-ALS which appears biologically separate from non-*SOD1* ALS (Mackenzie et al., 2007) but may be suboptimal when the biological disease signature is indistinguishable between people with and without a given variant, as for *C9orf72* and non-*C9orf72* associated ALS (Humphrey et al., 2023), given the breadth of processes implicated in the disease.

Patient stratification is an effective avenue for discovery of biological mechanisms relevant to particular subgroups (Jones et al., 2015). Such stratification has been attempted previously. For instance, clusters of ALS have been identified based on biological trends in transcriptomic and neuroanatomical data, with evidence suggesting that these groups may be reflected in the phenotype (Bede, Murad, Lope, Hardiman, et al., 2022; Bede, Murad, Lope, Li Hi Shing, et al., 2022; Dukic et al., 2021; Eshima et al., 2023; Tam et al., 2019).

Clustering has also been applied based on clinical data. Five distinct clinical clusters, predictive of disease duration until death or censoring, were identified in a British cohort using latent class cluster analysis (LCA) (Ganesalingam et al., 2009). Independently, semi-supervised machine-learning applied to high-dimensional clinical data from two independent Italian cohorts identified clusters conforming to the following clinical

subgroups: bulbar, respiratory, flail arm, classical, pyramidal, and flail leg (Faghri et al., 2022).

No study to date has validated subgroups they identified using independent samples from different populations which greatly limits their applicability. Moreover, despite ALS being clinically defined, no study has examined whether data-driven clinical subgroups differ biologically beyond subgroups defined by individual gene variants. The identification of clinically and biologically homogenous subgroups that are robust and consistent across populations of patients will likely benefit development of precision medicine approaches for ALS extending beyond those targeting specific genetic vulnerabilities.

We accordingly built upon existing studies of clinically-based clustering in ALS by examining: (1) whether data-driven clusters defined by widely collected clinical measures could be identified and validated across international ALS cohorts; (2) clinical characteristics defining these clusters; (3) whether clusters differ biologically in terms of (i) the frequency of rare variants in genes previously associated with ALS, (ii) common genetic variation captured within polygenic risk scores (PRS) for risk of ALS and related neuropsychiatric disorders, and (iii) molecular signatures via gene expression levels; (4) the extent to which clinically driven clusters could be identified using data attainable around the time of diagnosis. Across these investigations we used data from the Project MinE (Project MinE ALS Sequencing Consortium, 2018), the Survival, Trigger and Risk, Epigenetic, eNvironmental and Genetic Targets for motor neuron Health (STRENGTH) consortia, and the King's College London (KCL) Brain Bank (Iacoangeli et al., 2021; Jones et al., 2021).

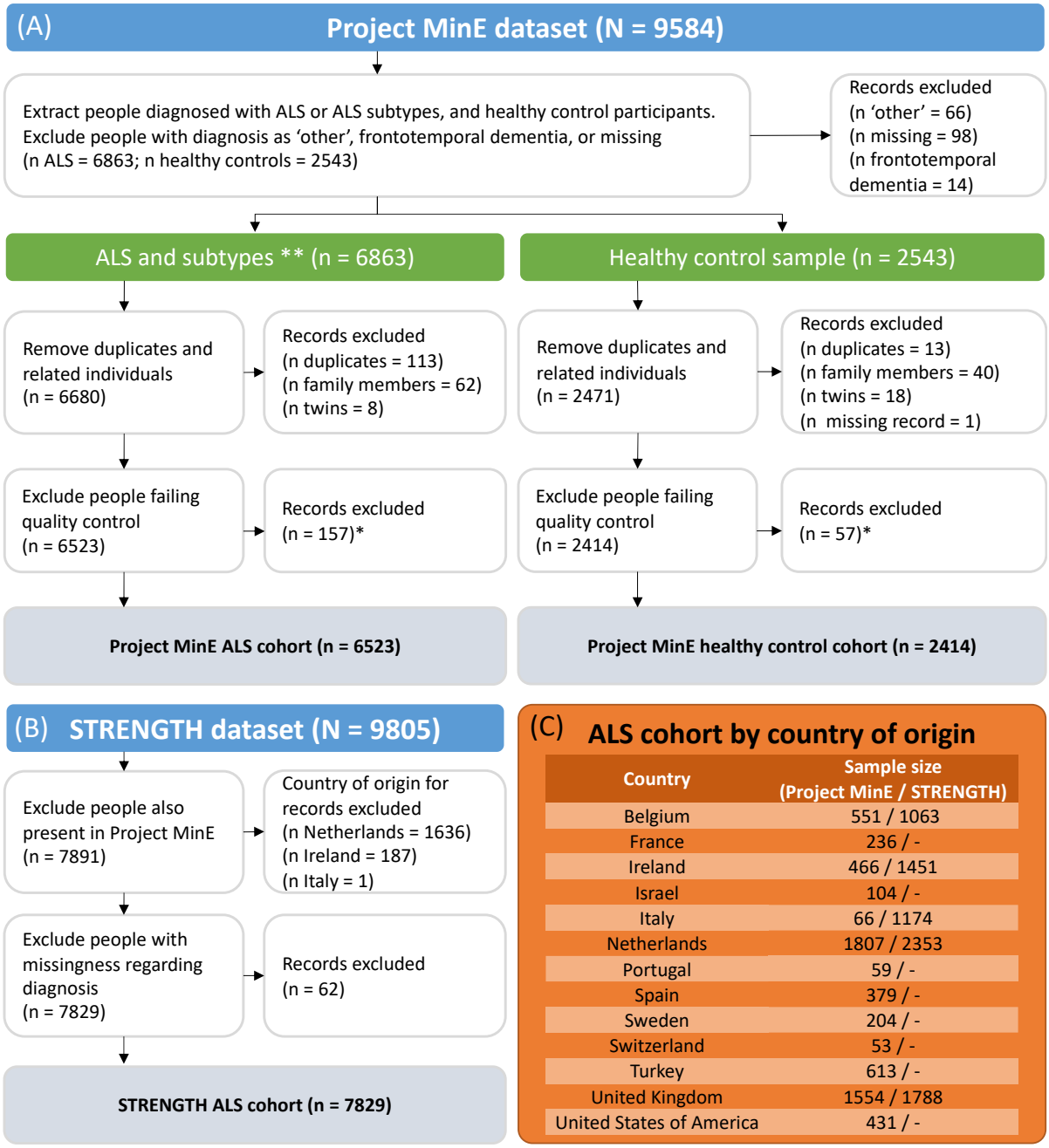
6.3. Methods

6.3.1. Sample

People with a diagnosis of ALS, PMA, or PLS were sampled from two international consortia: Project MinE (Project MinE ALS Sequencing Consortium, 2018) and STRENGTH. All samples underwent pre-processing (see Figure 6-1) to determine the sample entered into LCA and subsequent analyses (N: Project MinE = 6,523, STRENGTH = 7,829). Missingness was present in both cohorts (see Figure C-1). This is handled in LCA using a full information maximum

likelihood approach, which enables model fitting for records with incomplete data (T. Lee & Shi, 2021). Subsequent analyses employed a complete case analysis approach, retaining only people with information recorded for all variables used in that analysis.

Age- and sex-matched control participants from Project MinE (N = 2,414 after pre-processing; see Figure 6-1) were also sampled as a comparison group for analysis of trends in common genetic variation.



Country	Sample size (Project MinE / STRENGTH)
Belgium	551 / 1063
France	236 / -
Ireland	466 / 1451
Israel	104 / -
Italy	66 / 1174
Netherlands	1807 / 2353
Portugal	59 / -
Spain	379 / -
Sweden	204 / -
Switzerland	53 / -
Turkey	613 / -
United Kingdom	1554 / 1788
United States of America	431 / -

Figure 6-1. Summary of data processing and samples available by country for the Project MinE and STRENGH cohorts

*ALS = amyotrophic lateral sclerosis, PLS = primary lateral sclerosis, PMA = progressive muscular atrophy. *Quality control procedures for Project MinE have been described in previously (cf. (Project MinE ALS Sequencing Consortium, 2018; van der Spek et al., 2019; van Rheenen et al., 2021)); **ALS clinical diagnosis encompasses people recorded as: ALS, ALS/frontotemporal dementia, ALS/PLS, ALS/PMA, or progressive bulbar palsy (cf. (Al-Chalabi et al., 2016)); PLS and PMA are retained as distinct ALS subtypes in all analysis.* ***Panel A:*** *processing for people with ALS and healthy controls from the whole-genome sequencing cohort from Project MinE;* ***Panel B:*** *processing of the clinical ALS dataset from STRENGTH;* ***Panel C:*** *distribution of people in the final ALS cohort across countries.*

6.3.2. Study design

6.3.2.1. Clinical data

Phenotypic information from the Project MinE and STRENGTH datasets were used as features for clustering. Included variables are all frequently collected for people with ALS: sex at birth (male or female), site of onset (not-bulbar or bulbar), clinical diagnosis (ALS, PLS, or PMA), age of onset (years), disease duration (years) from onset until death or last status update with associated censoring status (alive or deceased), and delay from onset until diagnosis. Diagnostic delay was standardised by country to account for any inter-country differences (see Figure C-2). Standardisation was performed by centring each person's diagnostic delay by the per-country mean and scaling relative to the per-country standard deviation (see Table C-1).

6.3.2.2. Genetic data

Associations between clinically-defined clusters and biological trends were tested using data from Project MinE. Whole-genome sequence data were generated as previously described (Project MinE ALS Sequencing Consortium, 2018).

Information on rare genetic variation was extracted for a panel of 36 genes previously implicated in ALS: *ALS2, ANG, ANXA11, ATXN1, ATXN2, CFAP410* (formerly *C21orf2*), *CHCHD10, CHMP2B, C9orf72, DAO, DCTN1, ERBB4, FIG4, FUS, hnRNPA1, MATR3, MOBP, NEFH, NEK1, OPTN, PFN1, SCFD1, SETX, SIGMAR1, SOD1, SPG11, SQSTM1 (p62), TAF15, TARDBP, TBK1, TUBA4A, UBQLN2, UNC13A, VAPB, VCP, VEGFA*. Summaries of variants occurring across Project MinE samples are available in the databrowser (van der Spek et al., 2019).

Variation in *C9orf72* and *ATXN2* was reported regarding presence or absence of known pathogenic short tandem repeat expansions in each gene. Repeat expansion lengths were determined using Expansion Hunter (Dolzhenko et al., 2017). *C9orf72* expansions were confirmed using repeat-primed PCR, being classified as inconsistent if the dry and wet lab results did not match (Akimoto et al., 2014; Iacoangeli, Al Khleifat, Jones, et al., 2019). The minimum number of repeat units denoting presence of a repeat expansion was 30 for

C9orf72 and 28 for *ATXN2* (Sproviero et al., 2017). For *C9orf72,* inconsistent expansions were reported in 32 people and coded as missing.

Presence or absence of rare variants (MAF <0.01 in controls from both Project MinE and gnomAD v2.1.1 (Karczewski et al., 2020)) predicted by the Ensembl Variant Effect Predictor (McLaren et al., 2016) to have high or moderate impact upon gene function was recorded across the remaining 34 genes. Moderate impact variants included missense, in-frame insertions and deletions, and protein altering variants. High impact variants included stop lost and gained, start lost, transcript amplification, frameshift, transcript ablation and splice acceptor and donor variants.

Variants across all genes except *C9orf72* and *SOD1* were aggregated into burden groups (see Table C-2). The main burden groups described three functional pathways related to cellular processes disrupted in ALS: 'autophagy and proteostasis', 'RNA function', and 'cytoskeletal dynamics and axonal transport'. Genes were assigned to pathways according to the involvement of their protein products within them and to the processes which are disrupted by deleterious variation in each gene (see Table C-2); this was determined through literature review. A further 'any pathway' group, aggregating across the three functional pathways, was also defined. Each burden group was binary-coded according to presence or absence of variants in at least one gene assigned to the group.

The *SOD1* and *C9orf72* genes were analysed individually since their variants are the most frequent genetic causes of ALS, likely reflecting that they are implicated in various disease pathways (Balendra & Isaacs, 2018; Bunton-Stasyshyn, Saccon, Fratta, & Fisher, 2015), and each occurred with sufficient frequency to be tested individually.

Associations between clusters and common genetic variation was examined using PRS indicating risk for ALS and related neuropsychiatric diseases. PRS were derived from European ancestry genome-wide association study (GWAS) summary statistics of risk for ALS (van Rheenen et al., 2021), FTD (Ferrari et al., 2014), Alzheimer's disease (Kunkle et al., 2019), Parkinson's disease (Nalls et al., 2019), and schizophrenia (Trubetskoy et al., 2022).

Since samples from Project MinE are included within the ALS GWAS, PRS for this trait were generated based on GWAS summary statistics that exclude meta-analysis 'stratum 6', which includes most people from Project MinE. To ensure no sample overlap, analyses including the ALS PRS were performed using only those who were sampled within GWAS stratum 6. All available Project MinE samples were included in analyses based on PRS for other traits.

GWAS summary statistics were pre-processed as described in Chapter 3.2. PRS were calculated with SBayesR (Lloyd-Jones et al., 2019) under the reference-standardized approach of *GenoPredPipe,* whereby scores for the Project MinE target samples were standardised against the 1000 Genomes European ancestry population reference (Auton et al., 2015; Pain et al., 2021). SBayesR was applied using the default settings and the robust parameterisation option. Linkage disequilibrium was estimated using the pre-computed sparse matrices provided with the software, based on 50,000 individuals of European descent from the UK Biobank (Lloyd-Jones et al., 2019). We used the impute-n option of SBayesR to exclude variants with effective sample size >3 standard deviations from the median for GWAS summary statistics where per-SNP effective sample size was unavailable.

6.3.2.3. Gene expression and methylation data

The KCL BrainBank expression dataset consists of post-mortem bulk RNA-sequencing samples from the Medical Research Council (MRC) London Neurodegenerative Diseases Brain Bank at KCL. Frozen human post-mortem tissue was taken from the primary motor cortex of 112 patients whose genomes were sequenced as part of Project MinE.

Matching methylation data for BrainBank samples was also used here to perform phenotype analysis. DNA methylation was analysed using Illumina Infinium EPIC array following the standard Infinium HD array methylation protocol (Illumina).

The protocols for RNA-sequencing (Iacoangeli et al., 2021; Jones et al., 2021) and generating methylation data (Hop et al., 2022) have been described previously.

6.3.3. Procedure

6.3.3.1. Clustering of ALS clinical data

Latent Class Cluster Analysis (LCA) was applied with *Mplus* (v8.7) and the *MplusAutomation* R package (Hallquist & Wiley, 2018) (package v1.1.0; R v4.1.3) with the MLR estimator to identify data-driven clusters in clinical data. LCA is an unsupervised machine-learning approach with various benefits: information returned enables the immediate inspection of fit quality; data across categorical, ordinal, and continuous modalities can be combined, including time-to-event variables with censoring; an in-built full information maximum likelihood approach (T. Lee & Shi, 2021) enables class (cluster) assignment for people with incomplete records.

Project MinE was used as the discovery cohort, holding back independent samples from STRENGTH for model validation. After validation, the independent Project MinE and STRENGTH samples were pooled together and LCA was repeated within the joint dataset. We attempted to fit latent class models which considered between 1 and 9 latent classes for both discovery and joint samples.

Fitted models were compared first using Akaike (AIC) and Bayesian (BIC) information criterion, where smaller values indicate an improved model fit (Weller, Bowen, & Faubert, 2020). We additionally considered the theoretical interpretability and parsimony of the solution, which is typical for LCA to minimise identification of small and uninterpretable classes. The quality of the accepted fit was evaluated using entropy and minimum average probability of belonging to the assigned class. High entropy indicates that the identified classes describe the data well (Celeux & Soromenho, 1996); values of around 0.8 or higher indicate good model fit. Likewise, greater than 0.8 average probability belonging to the assigned class indicates that people conform well to class assignments.

To ensure that accepted models in each dataset were unbiased by use of full information maximum likelihood to approximate missing values in important model features, the final models were refitted using a subset of the samples, omitting people missing diagnostic delay and disease duration feature information.

To validate the best-fitting model identified using the discovery-sample, people from STRENGTH were assigned to classes by *Mplus* and the model fit statistics were re-evaluated.

As a further test of external validity for the accepted model from the discovery sample, we predicted class assignments in STRENGTH using a k-nearest neighbours algorithm. The ground truth for class membership was those assigned by *Mplus*. The k-nearest neighbours algorithm was trained upon the LCA model features using the Project MinE discovery sample (including people with censored disease duration). We allowed KNN to consider between 1 and 20 neighbours, running the algorithm for each value 20 times, and then performing a final prediction for the value with the highest mean accuracy for predicting class assignments in STRENGTH (see Figure C-3; Table C-5). Area under the receiver operating characteristic curve (AUC) was used to determine predictive performance for each class vs any other class. The k-nearest neighbours algorithm was implemented using the R *class* package (v7.3.20) (Venables & Ripley, 2002), and AUC was analysed using *pROC* (v1.18.0) (Robin et al., 2011).

Consistency between the discovery and joint dataset latent class models was established by inspecting cluster assignment overlap between them. The accepted joint-dataset model was used for subsequent analyses.

6.3.3.2. Clinical characterisation of clusters

Characteristics of the clusters identified via LCA were first examined using linear discriminant analysis as implemented within the *MASS* (v7.3.57) package (Venables & Ripley, 2002). This algorithm derives linear axes that maximise separation between the classes and associations between these axes and the predictor variables indicate which variables best distinguish the classes.

Accordingly, the analysis included all features used in LCA. Clinical diagnosis (ALS, PLS, or PMA) was dummy coded with ALS as the reference category. Age of onset and disease duration were standardised to have a mean of 0 and standard deviation of 1. Diagnostic delay was, as before, standardised by country of origin. Sex (male or female) and site of onset (bulbar or other) were not recoded. Relationships between linear discriminant axes

and clinical variables were examined using pooled within-group correlations, implemented within the *psych* package (v2.2.9) (Revelle, 2022) *statsBy* function.

Although linear discriminant analysis performs adequately with categorical data (Gilbert, 1968; Moore, 1973), it assumes that predictors are continuous and normally distributed. Multinomial logistic regression (Tabachnick & Fidell, 2019), which makes no such assumption, was therefore applied with stepwise feature selection to support the analysis. Multinomial logistic regression analysis was implemented via *nnet* (v7.3.17) (Venables & Ripley, 2002) with the same predictors and coding as for linear discriminant analysis. The *MASS* package *stepAIC* function was applied with using forward and backward feature selection to remove any unimportant features based on AIC.

Linear discriminant and multinomial logistic regression analysis methodologies do not account for censoring in data. To ensure that our application of these techniques was not biased by inclusion of people with censored disease duration, the analyses were performed both including and excluding these individuals.

Time-to-event/survival analysis was performed via *survival* (v3.3.1) (Therneau, 2022) and *survminer* (v0.4.9) (Kassambara, Kosinski, & Biecek, 2021) to examine the relationship between class and disease duration. Class was first used as a univariate predictor of disease duration and differences were compared using pairwise log-rank tests. A Cox proportional-hazards model was fitted using class and the other clinical variables to predict disease duration. Predictor variables were encoded as described above. Disease duration was not standardised.

6.3.3.3. Biological trends across clusters

Associations between the *k* clusters identified and rare variation in ALS-implicated genes were investigated using *kx2* Fisher's exact tests. Odds ratios for having a variant in a given cluster vs all other clusters were also determined.

Binary logistic regression models were used to compare PRS for each cluster against all other clusters and against healthy controls. Each analysis was performed with a given PRS as

a univariate predictor and after including the first five principal components of ancestry as covariates.

The R *stats* (v4.1.3) *(R Core Team, 2021)* package was used to perform Fisher's exact tests for the rare variant analysis and to fit binary logistic regression models for the PRS analyses. Odds ratios for class vs other comparisons in the rare variant analyses were derived using the *epitools* (v0.5.10.1) (Aragon, 2020) package *oddsratio.fisher* function.

To determine whether the clusters displayed alterations in biological processes, we performed differential expression, gene enrichment, and cell type analysis for the samples from Project MinE which had matching motor cortex expression data available (n = 88). Gene enrichment analysis was performed using the top 500 differentially expressed genes. Cell composition analysis (McKenzie et al., 2018) was performed for the following cell types: neurons, endothelial cells, astrocytes, microglia, oligodendrocytes, and oligodendrocyte progenitor cells.

We additionally compared omics-based relative to chronological age at the time of death across clusters, based on estimates of brain-specific transcriptional age and methylation-based biological age.

Further procedural details for analyses of gene expression and methylation data are provided in Appendix C.1.1.

6.3.3.4. Prediction of cluster membership using baseline data

Random forest and eXtreme Gradient Boosting classification algorithms were trained to examine whether cluster membership could be predicted using only information accessible around the time of first diagnosis. Six algorithms were trained across the two machine-learning methods with a multiclass classification objective, predicting assignments to the LCA-identified clusters, across three data configurations: (1) all clinical features used in LCA except disease duration (which cannot be assessed at this time); (2) clinical features from model 1 alongside rare and common genetic variables used to assess biological trends across classes; (3) clinical features from model 1, sample-matched with model 2.

SHapley Additive exPlanations (SHAP) (Covert & Lee, 2021; Lundberg & Lee, 2017) were used to examine which features were more influential upon machine-learning algorithm Class membership predictions. SHAP values were selected because of their model-agnostic approach to explaining predictions based upon the additive contribution of each feature across a dataset.

To further examine which features were important for prediction of exclusively Class 1 versus 2, the most frequent LCA clusters, 6 additional binary classification algorithms were trained across the two machine-learning methods and three data configurations after restricting only to people assigned to these classes.

Only people without missingness on included features were used in each multiclass (binary) analysis, therefore the total sample size was 12,508 (11,109) in the first data configuration and 3,226 (2,996) in the second and third.

Algorithms were trained with 10-fold cross-validation, repeated 10 times using pseudorandom seeding. Prediction performance was evaluated using the metrics of sensitivity, specificity, precision, and balanced accuracy.

Further procedural information is provided in Appendix C.1.2.

6.4. Results

6.4.1. Clustering of ALS clinical data

We identified that clinical subgroups of ALS were well described within a 5-class latent class model (see Table C-3; Table 6-1).

Table 6-1. Class membership characteristics for a 5-class model in the Project MinE, STRENGTH, and joint datasets

Numbers presented in bold refer to average posterior probability of belonging to the assigned class. The statistics in the discovery and validation datasets are for the 5-class model fitted to the discovery sample. The Joint dataset statistics are for the 5-class model fitted to the joint dataset which combines Project MinE and STRENGTH.

Dataset	Assigned class	N in class (% of dataset) based on		Average posterior probability of belonging to class				
		posterior probabilities	most likely class membership	1	2	3	4	5
Discovery (Project MinE)	1	3702.34 (0.568)	3952 (0.606)	**0.883**	0.106	0.01	0.001	0
	2	2110.49 (0.324)	2023 (0.310)	0.104	**0.818**	0.056	0.011	0.011
	3	527.84 (0.081)	409 (0.063)	0.01	0.075	**0.909**	0.006	0
	4	114.15 (0.018)	87 (0.013)	0	0.006	0.047	**0.944**	0.003
	5	68.18 (0.010)	52 (0.008)	0.002	0.118	0.016	0.036	**0.829**
Validation (STRENGTH)	1	3887.36 (0.497)	4126 (0.527)	**0.89**	0.101	0.008	0.001	0
	2	2606.34 (0.333)	2452 (0.313)	0.084	**0.866**	0.044	0.004	0.002
	3	1009.62 (0.130)	943 (0.120)	0.009	0.07	**0.912**	0.008	0
	4	263.43 (0.034)	252 (0.032)	0	0.001	0.041	**0.955**	0.003
	5	62.24 (0.008)	56 (0.007)	0.001	0.006	0	0.008	**0.984**
Joint	1	7027.66 (0.490)	7401 (0.516)	**0.9**	0.093	0.006	0.001	0
	2	5602.10 (0.390)	5470 (0.381)	0.067	**0.879**	0.044	0.006	0.004
	3	1319.10 (0.092)	1138 (0.079)	0.003	0.085	**0.898**	0.013	0.001
	4	302.60 (0.021)	259 (0.018)	0	0	0.04	**0.957**	0.003
	5	100.54 (0.007)	84 (0.006)	0	0.067	0.024	0.019	**0.889**

The 5-class model was first identified in the Project MinE discovery sample, with lower Akaike information criterion (AIC) and Bayesian Information Criterion (BIC) values than models with fewer classes (Table C-3). Latent class models did not converge when considering a higher number of classes. Acceptance of the 5-class model was supported by it having entropy (0.791) indicating that the model had reasonable certainty in classification, and by people having a high probability of belonging to their assigned classes (Table 6-1). An equivalent model, with high entropy (0.850), was identified when repeating LCA after restricting to only people without missingness in diagnostic delay and disease duration (see Table C-4).

The external validity of this solution was affirmed through application to independent samples from STRENGTH. High entropy (0.830) and high average class probability of belonging to assigned classes (Table 6-1) indicated that the solution fit these data well. Accurate prediction of class membership for people from STRENGTH within a k-nearest neighbours algorithm trained upon the clinical data from Project MinE samples further supported the validity of the subgroups identified (see Table C-5).

Since the initial 5-class model fitted both the discovery and validation datasets well, LCA was repeated using a joint dataset, pooling across independent samples from Project MinE and STRENGTH. Models of 1 to 9 classes were fitted and the 5-class model was accepted above those with additional classes since the external validity of a 5-class model had already been demonstrated and because improvements to AIC and BIC were not substantial with additional classes (Table C-3). Entropy in the 5-class solution remained high (0.836) and people had high probability of belonging to their assigned classes (Table 6-1). As before, a highly comparable model, with 0.881 entropy, was identified when repeating LCA after restricting to only people with diagnostic delay and disease duration reported (see Table C-4).

Equivalence between the 5-class models fitted to the Project MinE and joint datasets was determined by examining cluster similarity: ~91% of people from Project MinE and ~92% of STRENGTH were assigned to the equivalent cluster in the joint dataset model (see Figure C-4).

Accordingly, the joint dataset 5-class solution was accepted as the final model and used for subsequent analyses.

6.4.2. Clinical characterisation of clusters

We examined the clinical characteristics of the clinically-defined ALS subgroups identified with LCA. Table 6-2 and Figure 6-2 present descriptive statistics for clinical features across each class. Linear discriminant and multinomial logistic regression analyses highlight that diagnostic delay and disease duration were the main class delineators (see Figure 6-3; Table 6-3; Table C-7, Table C-8). All features were retained in the multinomial logistic regression

model analysing only people without censored disease duration; sex was dropped as a predictor when people with censored disease duration were included in the analysis (Table C-7, Table C-8).

Survival analysis further demonstrated the relationship between class and disease duration, indicating that the classes are each associated with distinct survival trajectories (Figure 6-2), and class remaining the most influential predictor of survival within a Cox proportional-hazards model after adjusting for other clinical features (Table C-9).

Table 6-2. Descriptive statistics for the clinical characteristics of people with ALS across the 5-class solution fitted to the joint dataset

[†]calculated within the survfit function of the R survival package (Therneau, 2022); see Figure 6-2 for Kaplan-Meier curves stratified by class. Patterns of missingness across each variable are shown in Figure C-1.

		Total	1	2	3	4	5
Number of people		14352	7401	5470	1138	259	84
N female (%)		5823 (0.41)	3250 (0.44)	1987 (0.36)	460 (0.4)	97 (0.37)	29 (0.35)
N Bulbar onset (%)		4063 (0.3)	2911 (0.42)	951 (0.18)	155 (0.14)	37 (0.15)	9 (0.12)
Mean age of onset in years (standard deviation)		60.96 (12.24)	64.74 (10.39)	56.85 (12.62)	59.26 (12.07)	56.11 (12.25)	39.55 (17.28)
Median diagnostic delay in years (inter-quartile range)		1 (0.58, 1.67)	0.72 (0.47, 1.03)	1.17 (0.75, 1.83)	3.48 (2.92, 4.34)	6.8 (5.06, 9.42)	16.92 (10.97, 21.51)
N with censored disease duration (%)		3534 (0.26)	428 (0.06)	2363 (0.45)	532 (0.47)	148 (0.57)	63 (0.75)
Median disease duration, years (inter-quartile range)		2.94 (1.86, 5.08)	1.97 (1.39, 2.53)	4.67 (3.7, 6.83)	7.02 (5.1, 9.81)	11.54 (8.84, 16.95)	21.75 (15.39, 29.58)
Median duration after accounting for censoring[†]		3.19	2.01	5.67	9.12	19.55	46.14
Clinical diagnosis	ALS	13115 (0.91)	7217 (0.98)	4795 (0.88)	894 (0.79)	155 (0.6)	54 (0.64)
	PLS	533 (0.04)	23 (0)	250 (0.05)	159 (0.14)	79 (0.31)	22 (0.26)
	PMA	704 (0.05)	161 (0.02)	425 (0.08)	85 (0.07)	25 (0.1)	8 (0.1)

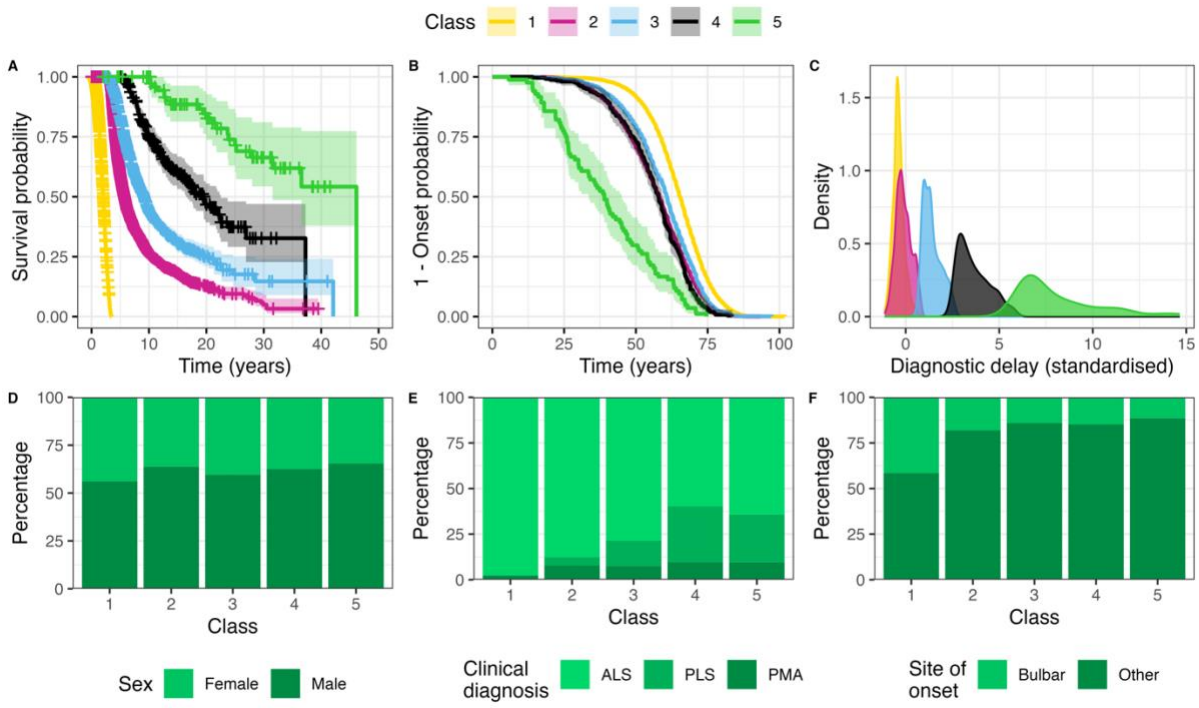


Figure 6-2. Trends in clinical features used in latent class analysis according to class

Panel A: *Kaplan-Meier curves for disease duration from onset until death or censoring; pairwise log-rank tests indicate that survival differs significantly between all classes ($p < 1x10^{-6}$ for all comparisons after false discovery rate adjustment for all pairwise tests performed). Orthogonal tick marks on the survival curves indicate censoring. Colouring around the curves indicates 95% confidence intervals.* ***Panel B:*** *Kaplan-Meier curves for age of onset.* ***Panel C:*** *density curves for diagnostic delay centred and scaled on the mean and standard deviation for diagnostic delay according to country of origin (see Figure C-2; Table C-1).* ***Panels D-F:*** *stacked bar-charts indicating distribution of the categorical variables sex (D), clinical diagnosis (E), and site of onset (F) across classes. ALS = amyotrophic lateral sclerosis; PLS = primary lateral sclerosis; PMA = progressive muscular atrophy.*

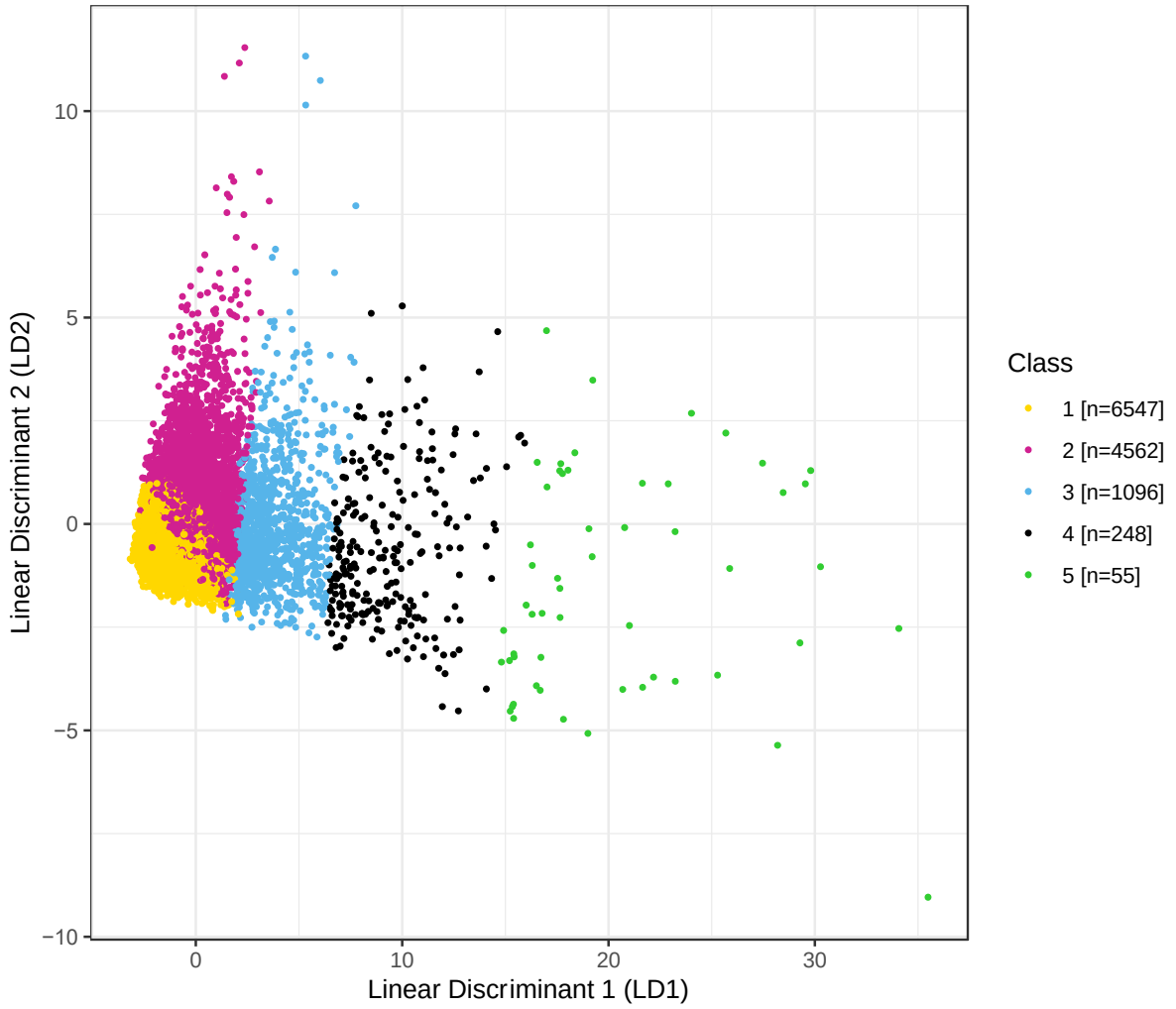


Figure 6-3. Distribution of people across the first two axes of linear discriminant analysis for all case-complete data

LD1 is highly correlated with diagnostic delay, while LD2 is associated primarily with disease duration (see Table 6-3). Figure C-5 presents a comparison figure for linear discriminant analysis when restricting to only people with non-censored disease duration. Figure C-6 shows this figure with people stratified by country of origin.

Table 6-3. Results of the linear discriminant analysis of all people with case-complete data

Proportion of trace describes the proportion of the separation between classes accounted for by each linear discriminant (LD) axis. Pooled within-group correlations greater than 0.5 are presented in bold and are considered variables associated with a given LD. Reference groups for categorical variables are: 'not-bulbar' for site of onset, 'male' for sex, 'amyotrophic lateral sclerosis' for clinical diagnosis. This analysis includes people with censored disease duration; Table C-6 presents comparable results for this analysis after restricting to only non-censored individuals. Figure 6-3 visualises the distribution of people and classes across the first two LD axes. PLS = primary lateral sclerosis; PMA = progressive muscular atrophy.

Statistic	**Variable**	**LD1**	**LD2**	**LD3**	**LD4**
Eigenvalue	-	130.01	31.57	4.80	2.28
Proportion of trace	-	0.943	0.056	1.28×10^{-3}	2.90×10^{-4}
Pooled within-group correlation	Diagnostic delay	**0.950**	-0.295	0.031	-0.084
	Age of onset	-0.081	-0.449	-0.388	-0.149
	Disease duration	0.406	**0.795**	0.024	0.198
	Site of onset (bulbar)	-0.065	-0.362	**0.520**	**0.634**
	Sex (female)	-0.009	-0.107	-0.052	-0.219
	Clinical diagnosis (PLS)	0.136	0.072	**-0.700**	**0.648**
	Clinical diagnosis (PMA)	0.034	0.181	0.008	-0.074

6.4.3. Biological trends across clusters

Fisher's exact tests were applied to examine associations between rare genetic variation and class for variants occurring within genes previously associated with ALS (see Table 6-4). *SOD1* variants and the *C9orf72* expansion differed in frequency across classes. For example, *C9orf72* expansions were overrepresented in Class 1 and underrepresented in Class 2; the opposite was observed for *SOD1* variants. The frequency of variants in genes linked to RNA processing and to cytoskeletal dynamics and axonal transport also differed across classes.

Table 6-4. Association between class and rare genetic variation in ALS-associated genes

Odds ratios are for having a variant in each class relative to all other classes; those with 95% confidence intervals (CI) which do not cross the null value of 1 are presented in bold. † false discovery rate (FDR) adjusted p-values <0.05 after adjusting for all Fisher's exact tests in the column are presented in bold. Δ Sample size differs for C9orf72 comparison, sample size in class: 1 = 3,374, 2 = 2,494, 3 = 322, 4 = 65, 5 = 36. §Any pathway refers to having a variant in any of the autophagy and proteostasis, RNA function, cytoskeletal dynamics and axonal transport pathways.

Genetic variation (P_{FDR}†)		**Class (sample size Δ)**				
		1 (3,401)	**2 (2,512)**	**3 (327)**	**4 (65)**	**5 (36)**
SOD1 (**2.04×10^{-4}**)	N with variant (freq)	13 (3.84×10^{-3})	31 (0.01)	6 (0.02)	2 (0.03)	1 (0.03)
	Odds ratio [95% CI]	**0.28 [0.14, 0.53]**	**2.16 [1.21, 3.93]**	2.37 [0.82, 5.62]	3.87 [0.45, 15.31]	3.43 [0.08, 21.27]
C9orf72 repeat expansion (**2.42×10^{-6}**)	N with variant (freq)	250 (0.07)	101 (0.04)	14 (0.04)	1 (0.02)	0 (0)
	Odds ratio [95% CI]	**1.93 [1.53, 2.44]**	**0.56 [0.44, 0.71]**	0.73 [0.39, 1.25]	0.25 [0.01, 1.46]	0 [0, 1.75]
Autophagy and proteostasis (0.546)	N with variant (freq)	309 (0.10)	241 (0.11)	24 (0.08)	6 (0.10)	5 (0.16)
	Odds ratio [95% CI]	0.96 [0.81, 1.15]	1.08 [0.9, 1.28]	0.77 [0.48, 1.18]	1 [0.35, 2.32]	1.59 [0.48, 4.15]
RNA function (**0.036**)	N with variant (freq)	319 (0.10)	244 (0.11)	19 (0.06)	8 (0.14)	0 (0)
	Odds ratio [95% CI]	1.02 [0.86, 1.21]	1.08 [0.91, 1.29]	**0.59 [0.35, 0.94]**	1.37 [0.56, 2.91]	0 [0, 1.05]
Cytoskeletal dynamics and axonal transport (**0.036**)	N with variant (freq)	461 (0.16)	416 (0.20)	45 (0.16)	9 (0.16)	4 (0.13)
	Odds ratio [95% CI]	**0.82 [0.71, 0.94]**	**1.27 [1.1, 1.46]**	0.92 [0.65, 1.27]	0.93 [0.4, 1.9]	0.72 [0.18, 2.04]
Any pathway § (**0.029**)	N with variant (freq)	979 (0.40)	799 (0.47)	81 (0.33)	23 (0.55)	8 (0.29)
	Odds ratio [95% CI]	0.9 [0.81, 1]	**1.17 [1.05, 1.31]**	**0.77 [0.58, 0.99]**	1.29 [0.74, 2.21]	0.67 [0.26, 1.52]

Binary logistic regression models were fitted to examine associations between class and PRS for ALS and related neuropsychiatric disorders (see Figure 6-4). Class 1 was significantly associated with higher PRS for risk of ALS, compared both to people in other classes and to healthy controls. Interestingly, PRSs for schizophrenia and Parkinson's disease were lower in Class 1 compared to other classes and were higher in Classes 2 and 5. PRS for schizophrenia were higher in most classes with respect to controls, and those for Parkinson's disease were higher for Classes 2 and 5.

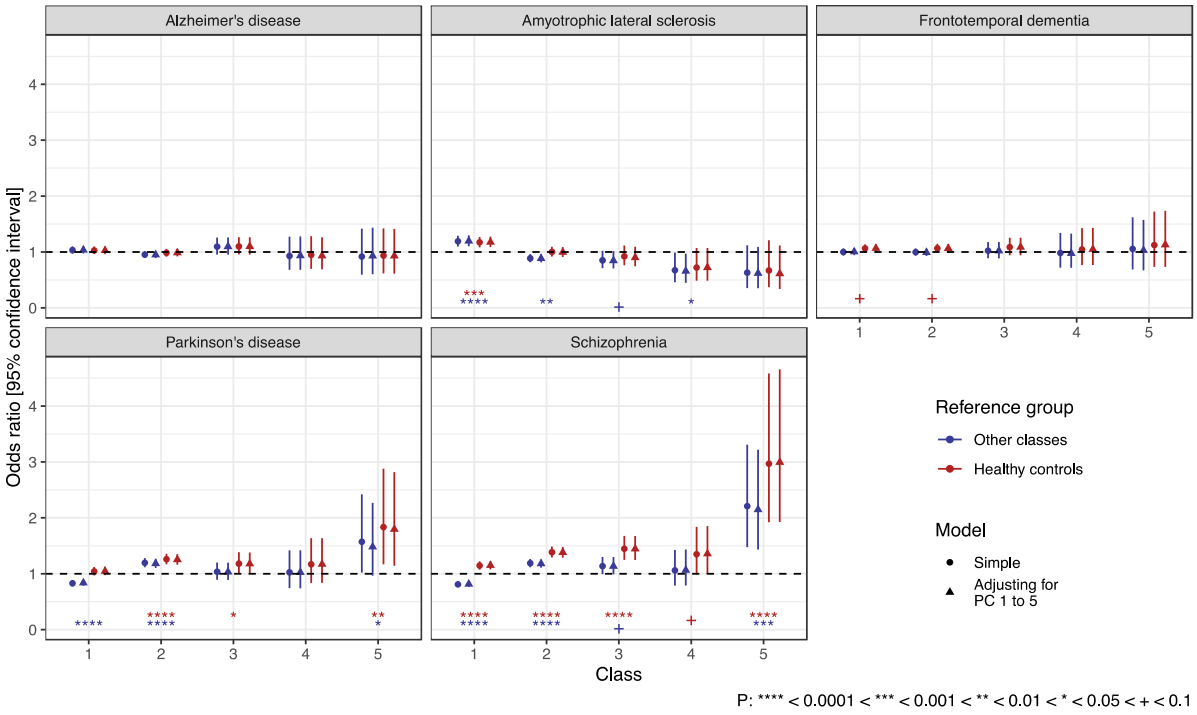


Figure 6-4. Odds ratios for association between polygenic risk scores (PRS) for neuropsychiatric diseases and class

Odds ratios are derived as exponentiated coefficients from binary logistic regression models, with the outcome variable of class vs all other classes in blue and class vs healthy controls in red. A circle denotes a 'simple' model including only the PRS indicated for that panel as a predictor, a triangle denotes a model which adjusts for the first five principal components (PC) of ancestry. Nominal statistical significance is indicated by asterisks for 'simple' models only, and colour coded according to the reference group. Sample sizes for each group in the PRS analyses [amyotrophic lateral sclerosis PRS sample size]: Class 1 = 3,464 [2,126], 2 = 2,421 [1,466], 3= 327 [203], 4 = 65 [45], 5 = 34 [20], controls = 2,371 [1,909].

RNA-sequencing data was available for 88 of 94 KCL BrainBank samples (a subset of Project MinE) assigned to classes via LCA (n: Class 1 = 70, Class 2 = 18). Differential expression analysis identified one gene, *MYCN*, significantly upregulated in Class 2 compared to Class 1 after independent hypothesis weighting correction ($\log_2$ Fold Change = 0.597, SE = 0.124, $p_{adjusted}$ = 0.033). The full differential expression results are available in Table C-10. Gene set enrichment analysis highlighted genes involved with cytoskeletal, extracellular matrix, muscle system, anatomical structure/cell development, and cell signalling related biological processes, as well as various microRNA targets, as over-represented among the 500 most significant differentially expressed genes. The results of this analysis are available in Table C-11.

Cell composition analysis performed upon the BrainBank Samples indicated that Class 2 had a significantly smaller proportion of endothelial cells (p = 0.00163) and microglia (p = 0.0442) than Class 1. Furthermore, although non-significant, transcriptional and biological age acceleration was, respectively, 3.1 and 1.4 years higher in class 2 (see Table 6-5).

Table 6-5. Results of cell composition and omics-based age analysis for BrainBank samples in Class 1 and Class 2 with matching motor cortex expression data

*Bold p-values denotes values <0.05. *Cell composition differences were assessed using the Wilcoxon rank-sum test (W statistic) and omics-based age was assessed using independent samples t-test (t statistic, degrees of freedom). SVD = singular value decomposition.*

		Mean ± SD		Test Statistic *	p-value
		Class 1 (N = 70)	Class 2 (N = 18)		
Cell Composition (SVD values)	Astrocytes	-0.00237 ± 0.0731	-0.0180 ± 0.0457	W = 679	0.616
	Endothelial Cells	-0.0131 ± 0.0391	-0.0385 ± 0.0208	W = 935	**0.00163**
	Neurons	0.0131 ± 0.0695	0.0422 ± 0.0542	W = 473	0.106
	Microglia	0.0107 ± 0.111	-0.0169 ± 0.0327	W = 825	**0.0442**
	Oligodendrocytes	-0.0193 ± 0.0407	0.00367 ± 0.115	W = 609	0.832
	Oligodendrocyte Progenitor Cells	0.0148 ± 0.0768	-0.00257 ± 0.0482	W = 661	0.752
Omics-based Age (years)	Transcriptional age acceleration	3.94 ± 10.2	6.99 ± 11.9	t = -1.085, df = 84	0.281
	Biological age acceleration	5.82 ± 3.74	7.25 ± 5.08	t = -1.305, df = 83	0.196

6.4.4. Prediction of cluster membership using baseline data

We examined the extent to which machine-learning algorithms could predict class membership using data available around the time of diagnosis. Random Forest and eXtreme Gradient Boosting algorithms were trained using clinical data only or a combination of clinical and genetic variables. Comparison of the area under the receiver operating characteristic curve for prediction of each class versus all other classes determined that the two approaches performed comparably (see Figure C-7; Figure C-8; Figure C-9). Table 6-6 presents performance metrics for the multiclass eXtreme Gradient Boosting algorithms trained upon each data configuration. Class membership was predicted with high accuracy for Classes 3-5 (see Table 6-6; Figure 6-5). Prediction of Classes 1 and 2 still performed reasonably, but with poorer specificity in Class 1 and sensitivity in Class 2; most

misclassification in these groups was for people in the opposing class (see Figure C-7; Figure C-8; Figure C-9). Algorithm performance was comparable across sample-matched datasets when using clinical features alone and using clinical and genetic measures.

Evaluation of feature importance using SHAP values identified diagnostic delay as the most influential feature upon predictions from either approach and that various other features contributed more greatly for prediction of Class 1 and 2 in particular (see Figure C-10).

Table 6-6. Performance of eXtreme Gradient Boosting classification algorithms for predicting class membership

*Class 5 was omitted from the algorithm when fewer than 20 people in the class remained in the dataset. Figure C-7, Figure C-8 and Figure C-9 present receiver operating-characteristic curves and corresponding area under the curve (AUC) for each class combination pairwise. †AUC values presented are for prediction of class vs all other classes (see Figure 6-5).*The 'first visit clinical data [matched]' rows describe an algorithm trained using features equivalent to the 'first-visit clinical data' model but after restricting the sample to match people included for the 'First-visit clinical data and genetic features' model.*

Model		Class				
		1	**2**	**3**	**4**	**5**
First-visit clinical data	Sample size	6547	4562	1096	248	55
	AUC†	0.847	0.827	0.999	0.999	1.000
	Sensitivity	0.77	0.68	0.97	0.99	1.00
	Specificity	0.76	0.81	0.99	1.00	1.00
	Precision	0.78	0.67	0.92	0.95	1.00
	Balanced Accuracy	0.77	0.74	0.98	0.99	1.00
First-visit clinical data and genetic features	Sample size	1813	1177	192	44	-
	AUC†	0.840	0.822	0.999	1.000	-
	Sensitivity	0.774	0.685	0.983	0.977	-
	Specificity	0.747	0.800	0.995	1.000	-
	Precision	0.797	0.663	0.925	0.977	-
	Balanced Accuracy	0.760	0.743	0.989	0.988	-
First-visit clinical data [matched]*	Sample size	1813	1177	192	44	-
	AUC†	0.839	0.820	0.999	1.000	-
	Sensitivity	0.776	0.681	0.984	0.977	-
	Specificity	0.743	0.802	0.995	1.000	-
	Precision	0.795	0.664	0.924	0.977	-
	Balanced Accuracy	0.760	0.742	0.989	0.988	-

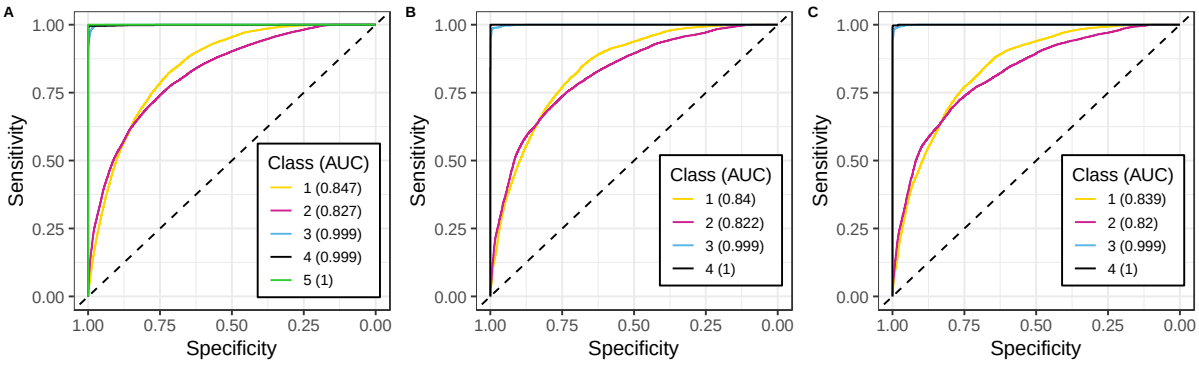


Figure 6-5. Receiver operator characteristic curves for performance of eXtreme Gradient Boosting algorithms in classifying each class versus all other classes

Curves shown indicate performance across all repeated cross-validation iterations. Each panel represents algorithms trained with different data configurations: A is upon clinical features available at diagnosis across all samples with complete clinical data (n = 12,508), B is these clinical features and measured genetic risks (n = 3,226), C is upon clinical features only, as in panel A, with a restricted sample to match the dataset of B (n = 3,226). Class 5 is excluded from classification in panels B and C owing to small sample size in these datasets. AUC = area under the curve.

6.5. Discussion

Latent class clustering analysis was applied to clinical data from large international ALS datasets to identify data-driven clinical subgroups and investigate whether biological differences exist between groups. Five distinct disease subgroups were identified, delineated primarily by diagnostic delay and disease duration, with importance also attributed to age and site of disease onset and clinical phenotype (ALS, PLS, or PMA). Notably, these clusters appear to generalise across countries (see Figure C-6).

Survival analysis indicated that the clusters were distinct regarding disease duration from onset until death; duration was greater for each respective class between 1 and 5. Diagnostic delay also generally increased with respect to class number, but, unlike disease duration, was similar in Class 1 and 2. Other features also aggregated unequally between clusters; bulbar onset was more frequent in Class 1; age of onset was later in Class 1 and earlier in Class 5, but comparable for Classes 2-4; the PLS clinical subtype was more frequent in Classes 4 and 5; the PMA subtype was more evenly distributed, but less frequent particularly in Class 1. Figure 6-2 visualises the clinical characteristics of each class.

The main clinical variables discriminating subgroups from this study are consistent with those distinguishing subgroups from a previous application of LCA in a UK ALS cohort (Ganesalingam et al., 2009). Both studies find diagnostic delay as an important discriminator of ALS subgroups and that subgroups have distinct survival trajectories. However, comparing class assignments of people from STRENGTH using the latent class model from each study demonstrates that the clusters are not the same (see Figure C-11). The most striking distinction is that most people assigned to Classes 1 and 2 using the present model are assigned to a single class by the model from the previous study. This difference can be explained by the inclusion of disease duration in present latent class model, which was withheld from clustering in the previous study and instead used as a tool for clinical validation.

Interestingly, neither the current or previous LCA-based clustering identified subgroups corresponding to clinically defined ALS subtypes (e.g., PMA or PLS), which supports the hypothesis that the validity of current clinically defined subtypes should be reconsidered. However, this may reflect a high rate of misdiagnosis or that the variables required to distinguish between clinical subtypes were not captured within these models. Use of high-dimensional clinical data identified ALS subgroups congruent with the classification system in a recent study (Chiò et al., 2011; Faghri et al., 2022). However, these subgroups would have fallen within the pure ALS diagnosis in our study and extreme subtypes such as PMA and PLS were not represented.

Analysis of rare genetic variants showed that the pathogenic *C9orf72* repeat expansion was more frequent in Class 1, consistent with evidence that the variant is associated with shorter disease duration (Byrne et al., 2012; N. A. Murphy et al., 2017; Umoh et al., 2016; van Rheenen et al., 2021). Putatively deleterious *SOD1* variants were overrepresented in Class 2, the characteristics of which was more representative of the disease phenotype associated with variants in this gene than for than Class 1 (Opie-Martin et al., 2022).

Variants in genes associated with cytoskeletal dynamics and axonal transport cell processes were more frequent in Class 2. Variants in genes associated with RNA function were also unequally distributed across groups, although no individual group appeared to drive the

association. These findings suggest that the ALS phenotype may present differently according to disruption of certain cellular processes. Recent studies support this possibility, reporting differences in disease progression and survival according to common variants associated with antioxidant and inflammatory disease pathways (Ravnik-Glavač et al., 2022) and according to expression-based clusters (Eshima et al., 2023). Associations between disruption to particular biological processes and the ALS phenotype warrant further investigation.

Trends in common genetic variation were examined using PRS for risk of ALS or related neuropsychiatric traits. Class 1 was associated with higher ALS PRS. Conversely, Classes 2 and 5 were associated with higher schizophrenia PRS, which was also generally higher across ALS subgroups compared to healthy controls. Classes 2 and 5 were also associated with higher, and Class 1 with lower, Parkinson's disease PRS. No associations emerged between class and PRS for FTD or Alzheimer's disease. The lack of association with FTD PRS may reflect the limited simple size of the only available GWAS for this trait (Ferrari et al., 2014), and thus an underpowered PRS.

The finding that ALS PRS were only higher than healthy controls in Class 1, the largest cluster, suggests that the GWAS primarily captures variants relevant for disease risk in this subgroup. Different variants may therefore be relevant for disease risk in the other subgroups. The possibility is supported by the association between Classes 2 and 5 and higher schizophrenia and Parkinson's disease PRS. These associations also suggest that genetic overlaps between ALS and other traits could be driven by certain disease subgroups (McLaughlin et al., 2017; van Rheenen et al., 2021).

Analysis of matching post-mortem motor cortex transcriptomic data available for a subset of Project MinE supported the differences in biological trends between Classes 1 and 2. Gene enrichment analysis identified many significant processes and pathways linked to cytoskeletal dynamics and axonal transport, congruent with our finding that variants in genes linked to this pathway were more frequent in Class 2. The variability in biological trends observed across these clinical subgroups supports the perspective that patient

stratification may be important for identifying biological disease mechanisms (Jones et al., 2015).

Lastly, we found that class membership could be predicted with only information attainable around the time of diagnosis using machine-learning classification algorithms (see Table 6-6; Figure 6-5). Classes 3 to 5 were clearly distinguished and poorer performance for Classes 1 and 2 is expected since their main delineator is disease duration which was excluded as a feature (see Table 6-2; Figure 6-2; Figure 6-3). Algorithms making predictions using both genetic and clinical features performed comparably to those trained using clinical features only. For rare variants in particular, this may reflect that presence of a given variant only informs predictions for a small proportion of the whole cohort.

A limitation of this study is that rare variant analysis was conducted under the assumption that identified variants are relevant for the risk or modification of the ALS phenotype. We included rare variants predicted to have a functional effect upon genes previously implicated in ALS. Such broad inclusion criteria will likely identify a range of variants with a spectrum of relevance to the disease. Without further supporting evidence, or larger sample sizes, it is difficult to ascertain the role of each individual variant (Richards et al., 2015).

A further limitation is that biological trends could only be tested to a limited degree. Analyses with transcriptomic and methylation data were substantially constrained by the availability of data for relatively few samples from Classes 1 and 2. The genetic analyses were similarly limited by the small sample size for certain clusters. However, our investigation drew upon one of the richest genomic resources for ALS currently available (Project MinE ALS Sequencing Consortium, 2018) and this constraint should lessen in future studies as resources continue to expand.

Future studies should refine the disease classifications we have developed. The identified subgroups were partly and differentially separated by diagnostic delay and disease duration. Progression has been identified as an indicator of disease duration in ALS (Labra, Menon, Byth, Morrison, & Vucic, 2016; Rutkove, 2015), and the present patterns suggest non-linearity of progression across the disease course, which has been recognised previously

(Ramamoorthy et al., 2021). Our measurement of diagnostic delay is likely an imperfect proxy for disease progression. Therefore, future studies should include measures of patterns in disease progression. Recognising overlaps between ALS and other conditions (van Rheenen et al., 2021; Zucchi et al., 2019), it is also pertinent to measure non-motor features of ALS, such as cognitive or behavioural change.

In conclusion, this chapter illustrated that data-driven approaches can identify distinct clinical subtypes of ALS and suggests that differences between these subgroups may reflect the role of distinct biological mechanisms, beyond individual gene variants, upon the phenotype. The study supports the perspective that data-driven patient stratification may aid identification of biological disease mechanisms and, therefore, that such approaches should be considered in the design of future ALS studies. Defining clinically and biologically meaningful subtypes of ALS has important implications for future research and clinical practice regarding: the inherent utility of an early and reliable prediction of disease prognosis for improving and personalising patient care; improved matching of people across clinical trial placebo and active arms to facilitate testing of treatment efficacy; precision medicine development owing to easier identification of disease processes relevant for particular subgroups.

Future research should aim to develop detailed understanding of ALS subtypes by employing multi-omic datasets to examine how they are reflected across the spectrum of genome to phenome.

Chapter 7. An online utility for comparative phenotype analysis in ALS

7.1. Abstract

Objective

Variants in the superoxide dismutase (*SOD1*) gene are among the most common genetic causes of amyotrophic lateral sclerosis. Reflecting the wide spectrum of putatively deleterious variants that have been reported to date, it has become clear that *SOD1*-linked ALS presents a highly variable age at symptom onset and disease duration.

Methods

Here we describe an open access web-tool for comparative phenotype analysis in ALS: https://sod1-als-browser.rosalind.kcl.ac.uk/. The tool contains a built-in dataset of clinical information from 1,383 people with ALS harbouring a *SOD1* variant resulting in one of 162 unique amino acid sequence alterations, and from a non-*SOD1* comparator ALS cohort of 13,469 individuals. We present two examples of analyses possible with this tool, testing how the ALS phenotype relates to *SOD1* variants which alter amino acid residue hydrophobicity and to distinct variants at the 94th residue of SOD1, where six are sampled.

Results and conclusions

The tool provides immediate access to the datasets and enables bespoke analysis of phenotypic trends associated with different protein variants, including the option for users to upload their own datasets for integration with the server data. This utility can be used to study *SOD1*-ALS and provides an analytical framework to study the differences between other user-uploaded ALS groups and our large reference database of *SOD1* and non-*SOD1* ALS. It is designed to be useful for clinicians and researchers, including those without programming expertise, and is highly flexible in the analyses that can be conducted.

7.2. Background

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease characterised by dysfunction and death of motor neurons leading to progressive muscle weakness and paralysis (R. H. Brown & Al-Chalabi, 2017). Its clinical presentation can vary greatly. For example, although people most frequently develop the first symptoms between 55 and 65 years of age, the disease can onset across all stages of adulthood. Similarly, time from

symptom onset until death is a median of 3 years for ALS but some people die within a year of onset, and 5-10% of people survive for more than 10 years (Al-Chalabi & Hardiman, 2013; Chiò et al., 2009; Juneja et al., 1997b).

A plethora of genetic factors can affect the risk of ALS or its phenotype, and mutations in specific genes can lead to distinct clinical outcomes. For example, a hexanucleotide repeat expansion in the *C9orf72* gene is the most common known cause of ALS and carriers of this mutation typically develop ALS earlier and with faster progressing symptoms than sporadic ALS patients (Byrne et al., 2012; N. A. Murphy et al., 2017; Umoh et al., 2016; van Rheenen et al., 2021). Furthermore, different mutations within the same gene can also lead to distinct forms of the disease. For example, over 180 variants in the superoxide dismutase (*SOD1*) gene (Bunton-Stasyshyn et al., 2015; M. Kalia et al., 2022; Opie-Martin et al., 2022) have been found in ALS patients. As well as affecting ALS risk, some of these variants have distinct effects on clinical features such as age of onset of motor symptoms and disease duration. For example, p.A5V and p.H44R have a marked effect on disease duration while p.G38R is associated with an early onset (Bali et al., 2017; E. P. McCann et al., 2017; Opie-Martin et al., 2022; Parton et al., 2002). Being able to characterise how genetic variants affect the clinical phenotype is essential for optimal development and design of healthcare, treatments, and trial stratification. However, the multitude of genetic factors involved in ALS and their rarity are great challenges for their individual study. To address these limitations, focussing on *SOD1* given the recent gene therapy trials (Miller et al., 2022), we recently collated data from the literature and specialised ALS centres globally on approximately 15,000 people with ALS, over 1,000 of whom harboured a variant in the *SOD1* gene (Opie-Martin et al., 2022).

In this chapter, we describe a web tool (https://sod1-als-browser.rosalind.kcl.ac.uk/) with upload facilities to allow people to perform comparative and bespoke phenotype analyses using data from a database of almost 15,000 people with ALS without need for informatics proficiency. The tool currently allows users to define and select subgroups of patients with or without variants in *SOD1*, to stratify by individual or groups of SOD1 variants, and to upload data to combine with our database in the analysis. To show the potential of this tool and how to use it, we present two example case studies which leverage the data from our

recent publication which is accessible to all users. The first example builds upon research suggesting that variants affecting protein hydrophobicity promote aggregation of mutant SOD1 (Tompa & Kadhirvel, 2020) and tests how alterations in amino acid hydrophobicity affect the ALS phenotype. The second example focuses specifically on variation at the 94th amino acid residue of SOD1, which is a site containing multiple reported variants, testing how the phenotype differs for each variant sampled.

7.3. Materials and methods

7.3.1. Dataset

The tool enables access to a dataset of 14,852 people with ALS, 1,383 of whom harbour a potentially deleterious non-synonymous *SOD1* gene variant (N without *SOD1* variant = 13,469). A total of 162 unique amino acid variants (canonical SOD1 sequence IDs: ENSEMBL = ENST00000270142.11, UniProt = P00441) are represented within these data (see Figure 7-1; Table D-1). The dataset is further described within our previous publication (Opie-Martin et al., 2022) and a summary of the disease characteristics associated with the 49 variants harboured by at least 5 people is provided on the site.

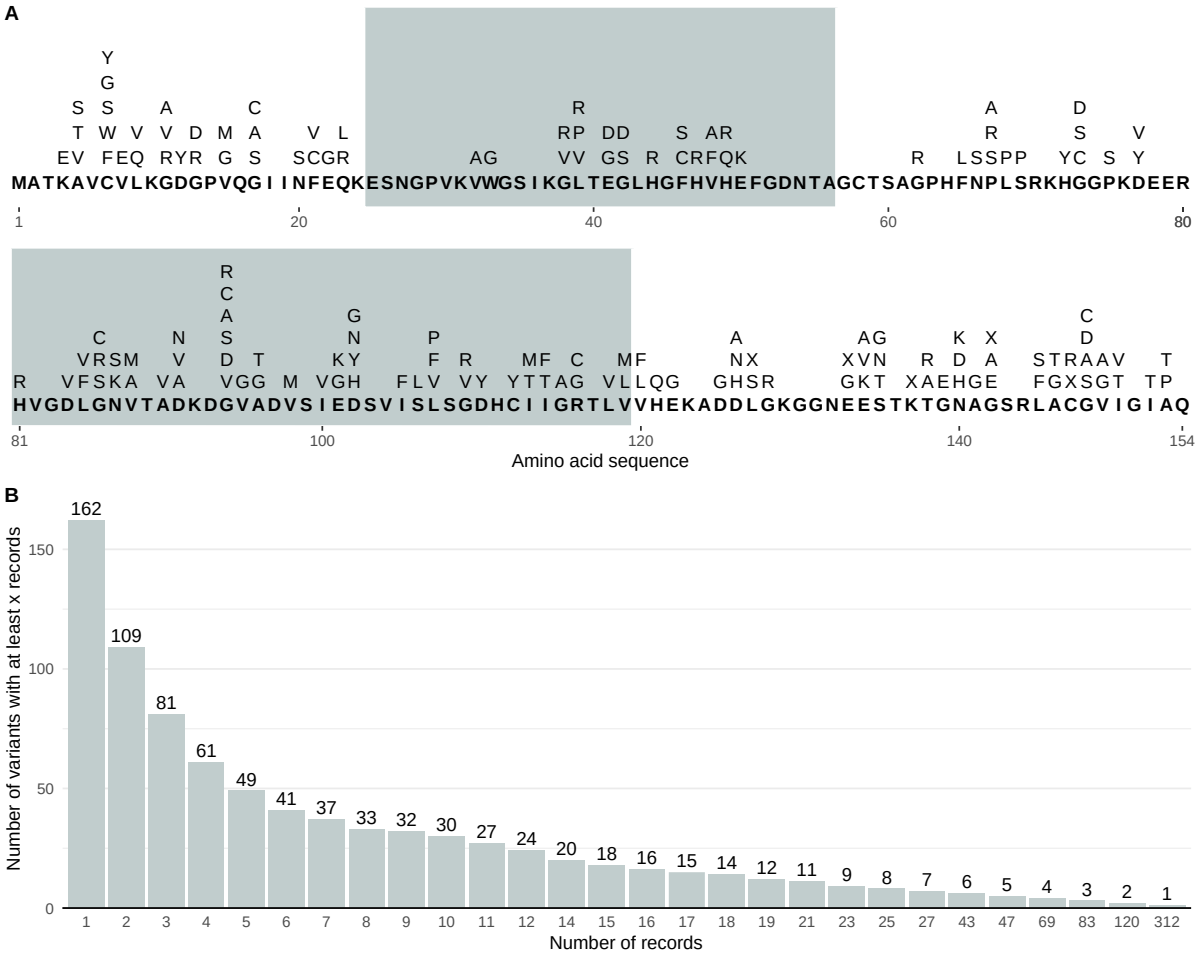


Figure 7-1. Variant characteristics for the native dataset

Panel A: *The canonical SOD1 amino acid sequence (bold) and variants recorded at each residue, denoted using IUPAC amino acid nomenclature (IUPAC-IUB Joint Commission on Biochemical Nomenclature, 1984), where 'X' indicates protein truncating variants. Alternating background shading indicates residues encoded from different exons of the SOD1 gene.*

Panel B: *The number of variants with at least a certain number of records available across different thresholds.*

7.3.2. Functionality

Survival analysis methods can be performed for (1) age at symptom onset, and (2) disease duration from symptom onset (with a corresponding censor variable indicating survival status). Kaplan-Meier and Cox proportional-hazards (CPH) approaches are both implemented and relevant descriptive statistics for the analysed sample are given by strata. Differences between strata in univariate analyses are examined using the log-rank test; global and pairwise log-rank tests are performed when more than two strata are defined. Analyses using CPH models are performed whenever two or more strata are defined or when a single stratum is specified and the user selects at least one of several covariates

which can be included in the regression model. Available covariates are clinical diagnosis, family disease history, sex, age of onset, site of onset, and sample source (continent of origin). Users can pick which covariates are included in the analysis depending on requirements and associations between selected survival analysis strata and available covariates can be tested.

Various analysis options are provided. The user can model survival for any number of individual SOD1 variants (including a 'no variant' option) and variants can be collapsed into groups of interest (including an 'any other SOD1 variant' option). We include three pre-defined options for grouping variants: by functional location (Opie-Martin et al., 2022) in the protein (across the dimer interface, electrostatic loop, zinc loop, and other) or according to the gene exon from which variants are transcribed. The final pre-defined analysis compares people with any SOD1 variant versus the 'no variant' group.

Users can further customise the analysis. They can filter by continent of origin and opt to stratify the analysis by sex, family history, site of onset, clinical diagnosis, and country or continent of origin. Time-dependent CPH analyses are also possible, allowing users to define timepoints at which the data are split. This functionality allows time-dependent coefficients to be modelled and enables analysis constrained to a certain timeframe (e.g., only of the first 12 months from symptom onset).

We allow users to upload supplemental data that is appended to the native sample, enriching the analysis possible within the tool. There are no restrictions regarding records that can be uploaded as supplemental data; users may provide data associated with *SOD1* variants both present in and absent from the native data or provide data from other groups of patients (e.g., for variants from other genes). Formatting instructions for supplemental data are provided on the site.

The results of the user's analyses are presented on the website, and we provide options to (1) download these within an HTML report and (2) download publication-ready versions of the figures produced, with customisable formatting.

7.3.3. Tool design

The tool is written in the R programming language (R v4.2.3) and developed using the R packages (versions) *shiny* (1.7.4), *shinyjs* (2.1.0), *shinycssloaders* (1.0.0), *tidyverse* (2.0.0), *rmarkdown* (2.21), *countrycode* (1.4.0), *kableExtra* (1.3.4), and *plotly* (4.10.1) (Allaire et al., 2022; Arel-Bundock, Enevoldsen, & Yetman, 2018; Attali, 2020; Chang et al., 2022; R Core Team, 2021; Sali & Attali, 2020; Sievert, 2020; Wickham et al., 2019; Zhu, 2021). Survival analyses are performed and visualised using *survival* (v3.5-5), and *survminer* (v0.4.9) (Kassambara et al., 2021; Therneau, 2023).

7.3.4. Examples of use

Here we present two examples of analyses possible within this tool. We examined differences in age of onset and disease duration between the strata of each example using Kaplan-Meier analyses and the log-rank test, and CPH models with robust variance estimation as implemented by *coxph* were applied to examine differences between strata before and after controlling for possible covariates. In the CPH models, we controlled for sex and age of onset when analysing disease duration, and sex only when analysing age of onset.

Case study 1 examined whether changes to amino acid hydrophobicity influenced age of ALS onset or disease duration from onset until death. Amino acids were grouped into three hydrophobicity categories (Sharer, 2014): hydrophobic (Amino acid IUPAC code (IUPAC-IUB Joint Commission on Biochemical Nomenclature, 1984): F, M, I, L, V), hydrophilic (D, E, H, K, R, N, Q), and intermediate (Y, W, P, G, A, S, T, C). Variants resulting in an amino acid substitution were then categorised based on the hydrophobicity group of the wild type and mutant amino acid; Table D-1 presents the assignment of groups and data availability across variants. To specifically examine the consequence of changes in hydrophobicity, three sets of analyses were conducted, each respective to variants occurring in residues that are hydrophilic, intermediate, or hydrophobic in the wild type protein. In each analysis, variants resulting in altered hydrophobicity were compared relative to variants where the mutant and wild type amino acids remained in the same hydrophobicity group. The p.A5V variant was excluded from these analyses since it is characterised by a particularly aggressive phenotype and accounted for the majority of records (n = 312) in the ‘intermediate to

hydrophobic' category. A broader hydrophobicity analysis across all groups was also conducted.

Case study 2 examined trends associated with variation at the 94th amino acid residue of SOD1, coding for a glycine in the wild type protein. Six variants were present at this locus. We first analysed differences in age of onset and disease duration associated with having any p.G94 variant vs any other SOD1 variant. Second, we compared p.G94 variants individually to non-p.G94 variants, aggregating across p.G94R, p.G94S, and p.G94V since they each contained fewer than 5 records.

Table 7-1 summarises characteristics of the data from both case studies.

*Table 7-1. Data summary for case studies. For disease duration analysis, restricted mean and median estimates are obtained from the survival curve and the standard error (SE) and 95% confidence interval (CI) indicates certainty in this estimate; the distribution of disease duration in non-censored individuals is reported directly. In age of onset analysis no person is censored, therefore estimates correspond with the raw descriptive statistics. *Minimum and maximum shown because disease duration was only available for two people. SD = standard deviation.*

Case study	Analysis stratum	Total sample size, N	N with age of onset	N with disease duration (N censored)	Age of onset in years		Disease duration in months		
					Mean [SD / SE]	Median [95% CI]	Median across non-censored people [25-75%]	Restricted mean estimate (months) [SE]	Median estimate [95% CI]
1: Amino acid hydrophobicity	Hydrophilic	118	115	106 (36)	47.96 [11.65 / 1.08]	47 [44, 51]	45.5 [18.25, 96]	157.93 [22.5]	85 [66, 125]
	Hydrophilic to intermediate	227	215	170 (65)	49.58 [13.61 / 0.93]	49 [47, 50.47]	60 [27, 108}	147.09 [14.28]	96 [84, 123]
	Hydrophilic to hydrophobic	17	17	15 (4)	49.77 [8.59 / 2.02]	48 [46.85, 56]	60 [32.5, 70.5]	96.59 [29.31]	65 [39, -]
	Hydrophobic	161	152	127 (51)	49.46 [11.39 / 0.92]	50 [48, 53]	42.21 [20.75, 84]	109.18 [17.09]	74 [50, 109]
	Hydrophobic to intermediate	221	202	154 (34)	50.82 [12.43 / 0.87]	49.5 [48, 51.19]	28.98 [17.01, 88.5]	85.22 [10.29]	45 [30, 72]
	Hydrophobic to hydrophilic	9	9	8 (0)	50.78 [8.69 / 2.73]	54 [49, -]	18 [9.5, 25.5]	26.37 [9.69]	18 [10, -]
	Intermediate	177	173	135 (36)	47.04 [13.58 / 1.03]	46 [44, 49]	24 [14.81, 53]	80.11 [14.09]	38.4 [24, 53]
	Intermediate to hydrophilic	96	90	76 (15)	49.06 [14.25 / 1.49]	48.5 [45, 51]	36 [21, 94]	101.59 [15.36]	45 [32, 84]
	Intermediate to hydrophobic	38	37	29 (9)	46.3 [13.93 / 2.26]	47 [43, 53]	71 [20, 114]	120.85 [30.43]	84 [56, 168]
2: p.G94 amino acid residue analysis	Non-p.G94 SOD1 variants	1320	1252	1034 (248)	49.07 [12.65 / 0.36]	49 [48, 50]	22.62 [12, 66]	88.43 [4.83]	37.59 [30, 44]
	Any p.G94 variant	63	63	52 (10)	46.06 [14.51 / 1.81]	45 [39, 49]	27.02 [19.91, 52.05]	75.21 [18.89]	32 [26, 53]
	p.G94A	27	27	26 (1)	48.7 [16.77 / 3.17]	48 [43, 61]	22 [16, 32]	33.44 [6.55]	22 [19, 32]
	p.G94C	14	14	9 (4)	40.46 [9.4 / 2.42]	37.5 [35, 51]	42.84 [37.49, 49.18]	221.68 [86.97]	235.4 [42.84, -]
	p.G94D	15	15	14 (5)	49.73 [15.17 / 3.78]	51 [45, 63]	46 [27.04, 74.88]	55.56 [8.64]	53 [31, -]
	p.G94R	1	1	1 (0)	34 [- / -]	34 [-, -]	55 [-, -]	55 [-]	55 [-, -]
	p.G94S	2	2	0	37.5 [10.61 / 5.3]	37.5 [30,-]	-	-	-
	p.G94V	4	4	2 (0)	41.25 [4.03 / 1.75]	41 [37, -]	114.5 [22, 207]*	114.5 [65.41]	114.5 [22, -]

7.4. Results

7.4.1. Amino acid hydrophobicity analysis

In case study 1, we examined how the ALS phenotype varied by changes in amino acid hydrophobicity. Across all amino acid substitutions sampled: 42.86% were variants which remained in the same hydrophobicity category as wild type SOD1; 42.11% were variants with a hydrophilic or hydrophobic amino acid in the wild type and an intermediate amino acid in the mutant protein; 12.59% were variants with an intermediate amino acid becoming hydrophilic or hydrophobic; and 2.44% were variants with substitutions from hydrophilic to hydrophobic amino acids or vice versa (see Table 7-1).

Age at symptom onset appeared roughly comparable across variants in all categories of the hydrophobicity analyses (see Table 7-2; Table D-2; Figure 7-2), with all groups having a mean age of onset between 46 and 51 years (Table 7-1).

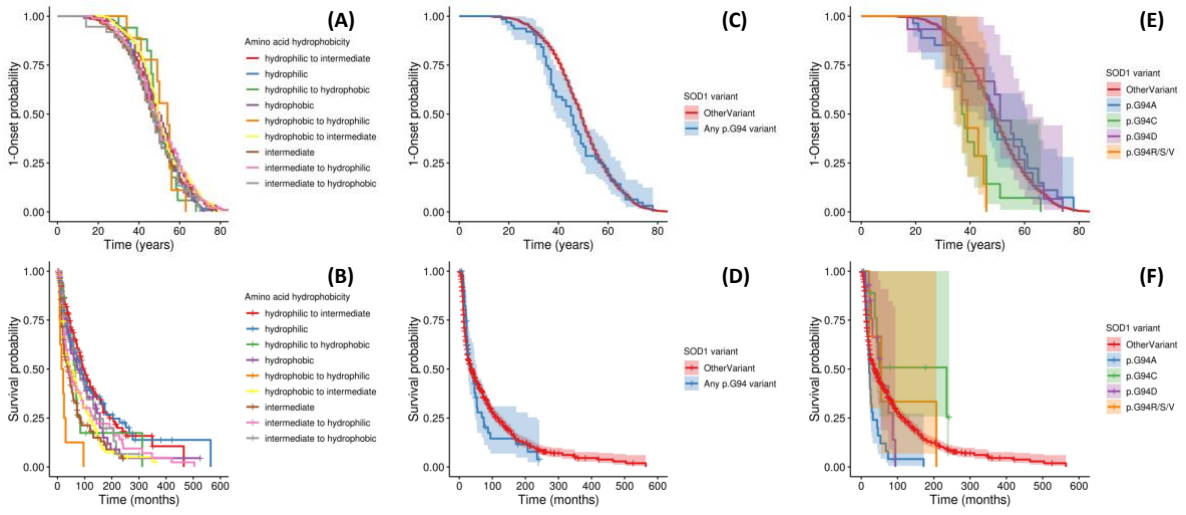


Figure 7-2. Kaplan-Meier survival curves for age of onset and disease duration analyses

Analysis shown: ***Panels A-B:*** *trends associated with wild type and variant amino acid hydrophobicity;* ***Panels C-D:*** *Any SOD1 p.G94 variant versus non-p.G94 SOD1 variants (OtherVariant).* ***Panels E-F:*** *individual p.G94 variants versus non-p.G94 SOD1 variants. Panels A, C, and E are for age of onset analysis, and B, D, and F describe disease duration. Panels A and B display all hydrophobicity groups in a single figure for each analysis and confidence intervals are not displayed to maximise visual clarity; Figure D-1 visualises trends in age of onset and disease duration for these groups after stratifying across panels according to the hydrophobicity group of the wild type residue.*

Table 7-2. Inferential statistics for survival analyses across case studies

*Bold values denote nominal p-values <0.05. *controlling for sex in the age of onset analysis and for sex and age of onset in the disease duration analysis. #Hazard ratios greater than 1 indicate earlier age of onset/shorter disease duration in the non-reference group. †No p.G94S variants were available for the disease duration analysis. CPH = Cox proportional-hazards*

Analysis	Case study (reference group)	Analysis stratum	Hazard ratio [95% confidence interval] #		P-value of difference between stratum and reference group	
			Unadjusted	Adjusting for covariates*	Log-rank test	CPH model
Age of onset	1 (Hydrophobic)	Hydrophobic to intermediate	0.825 [0.675, 1.01]	0.822 [0.673, 1]	0.097	0.0546
		Hydrophobic to hydrophilic	1.01 [0.677, 1.52]	1.01 [0.682, 1.49]	0.784	0.97
	1 (Intermediate)	Intermediate to hydrophobic	1.09 [0.792, 1.49]	1.1 [0.798, 1.51]	0.670	0.57
		Intermediate to hydrophilic	0.851 [0.657, 1.1]	0.851 [0.658, 1.1]	0.219	0.22
	1 (Hydrophilic)	Hydrophilic to intermediate	0.819 [0.659, 1.02]	0.819 [0.657, 1.02]	0.097	0.0734
		Hydrophilic to hydrophobic	0.995 [0.702, 1.41]	0.994 [0.702, 1.41]	0.979	0.974
	2 (Non-p.G94 SOD1 variants)	Any p.G94 variant	1.12 [0.831, 1.52]	1.12 [0.827, 1.52]	0.344	0.466
		p.G94A	0.857 [0.561, 1.31]	0.836 [0.544, 1.28]	0.436	0.415
		p.G94C	2.4 [1.21, 4.76]	2.48 [1.28, 4.83]	**6.73×10^{-4}**	**7.39×10^{-3}**
		p.G94D	0.891 [0.57, 1.39]	0.92 [0.593, 1.43]	0.659	0.708
		p.G94R/S/V	3.49 [2.27, 5.36]	3.32 [2.14, 5.15]	**5.66×10^{-4}**	**9.34×10^{-8}**
Disease duration	1 (Hydrophobic)	Hydrophobic to intermediate	1.4 [1.06, 1.85]	1.32 [1, 1.75]	**0.0202**	0.0503
		Hydrophobic to hydrophilic	4.36 [2.01, 9.48]	5.28 [2.87, 9.72]	**1.25×10^{-5}**	**9.19×10^{-8}**
	1 (Intermediate)	Intermediate to hydrophobic	0.63 [0.41, 0.968]	0.6 [0.373, 0.966]	0.0576	**0.0355**
		Intermediate to hydrophilic	0.719 [0.514, 1.01]	0.686 [0.497, 0.949]	0.0627	**0.0228**
	1 (Hydrophilic)	Hydrophilic to intermediate	0.924 [0.676, 1.26]	0.814 [0.591, 1.12]	0.619	0.209
		Hydrophilic to hydrophobic	1.53 [0.83, 2.8]	1.4 [0.753, 2.58]	0.296	0.289
	2 (Non-p.G94 SOD1 variants)	Any p.G94 variant	1.07 [0.823, 1.4]	1.14 [0.887, 1.46]	0.642	0.308
		p.G94A	1.75 [1.35, 2.28]	1.73 [1.34, 2.25]	**5.95×10^{-3}**	**3.00×10^{-5}**
		p.G94C	0.446 [0.197, 1.01]	0.518 [0.236, 1.13]	0.0672	0.100
		p.G94D	0.89 [0.589, 1.35]	0.946 [0.65, 1.38]	0.748	0.771
		p.G94R/S/V†	0.85 [0.415, 1.74]	0.907 [0.375, 2.19]	0.788	0.828

Disease duration analysis (see Table 7-2; Figure 7-2) however, suggested that alterations in amino acid hydrophobicity may affect disease prognosis following onset. Analysis of variants at residues which are hydrophobic in wild type SOD1 indicated that disease duration was shorter in substitutions to hydrophilic amino acids (p-value: log-rank test = 1.25×10^{-5}; CPH model = 9.19×10^{-8}) and that substitution into intermediate amino acids also tended towards shorter disease duration (p-value: log-rank test = 0.0202; CPH model = 0.0503). The estimated median [95% confidence interval] disease duration was 74 [50, 109] months for hydrophobic to hydrophobic substitutions, 45 [30, 72] for hydrophobic to intermediate, and 18 [10, NA] for hydrophobic to hydrophilic.

Among variants occurring in intermediate residues of wild type SOD1, becoming either hydrophilic or hydrophobic was associated with longer disease duration. The estimated median disease duration [95% confidence interval] for intermediate to intermediate amino acid substitutions was 38.4 [24, 53] months, under half that of intermediate to hydrophobic (84 [56, 168]) and shorter than intermediate to hydrophilic (45 [32, 84]) mutations.

Analysis of variants in hydrophilic residues of wild type SOD1 did not identify clear differences in disease duration between substitutions which remained hydrophilic and those which became intermediate (p-value: log-rank test = 0.619; CPH model = 0.209) or hydrophobic (p-value: log-rank test = 0.296; CPH model = 0.289). The estimated median disease duration [95% confidence interval] for hydrophilic to hydrophilic substitutions was 85 [66, 125] months, trending towards being shorter than in hydrophobic to intermediate (96 [84, 123]) and longer than in hydrophilic to hydrophobic (65 [39, NA]) substitutions.

Table D-3 presents an additional CPH model comparing all hydrophobicity groups relative to substitutions in residues with intermediate to intermediate amino acid substitutions. The analysis indicated that disease duration was shortest in this and the hydrophobic to hydrophilic substitution groups.

7.4.2. p.G94 amino acid residue analysis

In case study 2, we examined trends associated with variation in the 94th SOD1 residue. p.G94A was the most frequent variant at this locus and 5 other variants occurred in the dataset (see Table 7-1). This case study showed variant-specific trends in age of onset and disease duration, which were not discernible when aggregating across p.G94 variants, when compared with non-p.G94 SOD1 variants (see Table 7-2; Figure 7-2).

Age of onset was earlier than in the non-p.G94 SOD1 variant reference category only in the p.G94C (p-value: log-rank test = $6.73x10^{-4}$; CPH model = $7.39x10^{-4}$) and p.G94R/S/V (p-value: log-rank test = $5.66x10^{-4}$; CPH model = $9.34x10^{-8}$) groups; this difference appears considerable since the median age of onset for non-p.G94 SOD1 was over 10 years later than median onset in these two groups (see Table 7-1; Figure 7-2(E)).

The disease duration analysis indicated that only the p.G94A variant was associated with shorter time to death (p-value: log-rank test = $5.95x10^{-3}$; CPH model = $3.00x10^{-5}$). Inspection of hazard ratios suggests that p.G94C trended towards longer disease duration compared to non-p.G94 variants even after controlling for age of onset and sex (p-value: log-rank test = 0.0672; CPH model = 0.100). Although the median disease duration was longer for variants in the p.G94D and p.G94R/S/V variant groups, data were insufficient to test the association.

7.5. Discussion

We have developed a web-tool to facilitate bespoke investigations of the impact of *SOD1* variants upon the ALS phenotype, using survival analysis approaches. We have provided two examples of this tool's utility, examining differences in ALS age at symptom onset and disease duration according to (1) variants of varying impact upon residue hydrophobicity across SOD1 and (2) distinct variants at the 94th SOD1 residue.

This online facility has key benefits for research on the heterogenous ALS phenotype. First, it permits a user-friendly interface for performing survival analysis, with various options for customisation in accordance with the user's needs. Second, it provides access to a large in-built *SOD1*-ALS cohort and non-*SOD1* comparator population, which can be further enriched if users provide their own supplementary data.

The hydrophobicity analysis suggested that substitution variants altering residue hydrophobicity from hydrophobic to intermediate or hydrophilic are associated with a shorter disease prognosis compared to variants in these residues which remained hydrophobic across wild type and mutant SOD1. This aligns well with evidence that altered hydrophobicity promotes aggregation of the SOD1 protein (Tompa & Kadhirvel, 2020), and may reflect greater destabilisation and misfolding of SOD1 when variants cause more extreme alterations in hydrophobicity (Cordes & Sauer, 1999; Dyson, Wright, & Scheraga, 2006; Gidalevitz, Krupinski, Garcia, & Morimoto, 2009). Interestingly, variants of intermediate to intermediate amino acid substitutions were characterised by particularly short disease duration.

Hydrophobic to hydrophilic amino acid substitutions and vice versa were, notably, infrequent relative to other substitutions. Given that these would represent the most extreme hydrophobicity alterations, this could indicate a potential survivorship bias and that these substitutions may be sufficiently deleterious to be evolutionarily suppressed. This appears reasonable since SOD1 is highly conserved, with deficiency being linked to severe and early onset phenotypes (Ezer et al., 2021; Farrimond & Talbot, 2022; Park et al., 2019), and on the basis of variants in these hydrophobicity groups being entirely absent from the gnomAD v2.1.1 population database (Karczewski et al., 2020) (see Table D-4).

Analysis of the variants at p.G94 emphasised the extent to which individual SOD1 variants differentially influence the phenotype. Grouping together all p.G94 variants suggested that age of ALS onset and disease duration are comparable for people with variants at this residue and those with non-p.G94 SOD1 variants. Only by examining variants individually did we observe that p.G94A was associated with shorter, and p.G94C trended towards longer, disease duration than non-p.G94 *SOD1*-ALS. Likewise, p.G94C and the aggregation of p.G94R, p.G94S, and p.G94V were indicative of substantially earlier age of onset. These findings are consistent with the results of our previous analysis of *SOD1*-ALS, emphasising distinction between trends in age of onset and disease duration across individual variants (Opie-Martin et al., 2022). They highlight particularly the importance of making available

resources to allow variant-level analyses of the ALS phenotype associated with variation in *SOD1*.

The tool is not without limitation. Most notable is that a number of the 162 SOD1 variants sampled are harboured by very few individuals and thus are not sufficient for individual variant analysis with the native dataset alone. However, this issue can be somewhat circumvented by aggregating rarer variants into a single analysis stratum, and by the possibility of increasing the dataset with user-supplied data.

Certain considerations apply when providing supplementary data to the tool. Firstly, CPH models may only include covariates that are available in the native dataset. Second, records from supplementary data may overlap with native dataset. To reduce this possibility, the tool will automatically flag any people among the supplementary dataset who may be a duplicate of a person in the built-in dataset, checking for matches by country of origin, SOD1 amino acid change (if the user indicates that one is present), age of onset, site of onset, sex, and disease duration (if not censored). Users can also consult the cohort description provided and contact ALSoD (https://alsod.ac.uk/; Abel et al., 2012) with any concerns.

Overall, the open-access web-utility we provide (https://sod1-als-browser.rosalind.kcl.ac.uk/) has a potentially substantial benefit for ALS disease research and direct translational use for the design of patient stratification approaches, as well as being useful for mutation adjudication committees needing to make decisions on likely disease course with limited data. It permits an array of analysis options which can be readily implemented by users without any programming knowledge, and can be enriched by the provision of a supplementary dataset. Accordingly, this tool allows clinicians and researchers to circumvent many possible barriers they may otherwise face, for instance, regarding insufficient data availability or in preparing these data for analysis. The potential translational benefit of this tool is substantial, facilitating growth in understanding of the ALS phenotype which may aid the design and implementation of effective healthcare, treatments, and clinical trials.

Chapter 8. Examining genetic overlaps between neuropsychiatric diseases

8.1. Abstract

Continued methodological advances have enabled numerous statistical approaches for the analysis of summary statistics from genome-wide association studies. Genetic correlation analysis within specific regions enables a new strategy for identifying pleiotropy. Genomic regions with significant 'local' genetic correlations can be investigated further using state-of-the-art methodologies for statistical fine-mapping and variant colocalisation. We explored the utility of a genome-wide local genetic correlation analysis approach for identifying genetic overlaps between the candidate neuropsychiatric disorders, Alzheimer's disease, amyotrophic lateral sclerosis, frontotemporal dementia, Parkinson's disease, and schizophrenia. The correlation analysis identified several associations between traits, the majority of which were loci in the human leukocyte antigen (HLA) region. Colocalisation analysis suggested presence of a shared causal variant between amyotrophic lateral sclerosis and Alzheimer's disease in this region. Our study identified candidate loci that might play a role in multiple neuropsychiatric diseases and suggested that disease-implicated variants in these loci often differ between traits. Accordingly, this suggests the role of distinct mechanisms across diseases despite shared loci. The fine-mapping and colocalisation analysis protocol designed for this study has been implemented in a flexible analysis pipeline that produces HTML reports and is available at: https://github.com/ThomasPSpargo/COLOC-reporter.

8.2. Background

The genetic spectrum of neuropsychiatric disease is diverse and various overlaps exist between traits. For instance, genetic pleiotropy between amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) is increasingly recognised, and ALS is genetically correlated with Alzheimer's disease (AD), Parkinson's disease (PD), and schizophrenia (C. Li et al., 2021; Ranganathan et al., 2020; van Rheenen et al., 2021). Improving understanding of the genetic architecture underlying these complex diseases could facilitate future treatment discovery.

Advances in genomic research techniques have accelerated discovery of genetic variation associated with complex traits. Genome-wide association studies (GWAS), in particular, have enabled population-scale investigations of the genetic basis of human diseases and anthropometric measures (Abdellaoui, Yengo, Verweij, & Visscher, 2023). Summary-level results from GWAS are being shared alongside publications with increasing frequency over time (Reales & Wallace, 2023), and a breadth of approaches now exist for downstream analysis based on summary statistics which can facilitate their interpretation and provide further biological insight.

Genetic correlation analysis allows estimation of genetic overlap between traits (Bulik-Sullivan, Finucane, et al., 2015; Bulik-Sullivan, Loh, et al., 2015; Werme, van der Sluis, Posthuma, & de Leeuw, 2022; Y. Zhang et al., 2021). A 'global' genetic correlation approach gives a genome-wide average estimate of genetic relatedness. However, these relationships can be obscured when correlations in opposing directions cancel out genome-wide (Werme et al., 2022). Recent methods allow for a more nuanced analysis, of 'local' genetic correlations partitioned across the genome (Werme et al., 2022; Y. Zhang et al., 2021). This stratified approach to genome-wide analysis could prove effective for identifying pleiotropic regions and designing of subsequent analysis aiming to identify genetic variation shared between traits.

A number of techniques aim to disentangle causality within associated regions. This is important because the focus on single nucleotide polymorphisms (SNPs), which are markers of genetic variation, in GWAS produces results that can be difficult to interpret, and causal variants are typically unclear. More so, because of linkage disequilibrium (LD), GWAS associations often comprise large sets of highly correlated SNPs spanning large genomic regions. Statistical fine-mapping is a common approach for dissecting complex LD structures and finding variants with implications for a given trait among the tens or hundreds that might be associated in the region (Y. Zou, Carbonetto, Wang, & Stephens, 2022).

Interpretation of regions associated with multiple traits can also be challenging, since it is often unclear whether these overlaps are driven by the same causal variant. Statistical colocalisation analysis can disentangle association signals across traits to suggest whether

the overlaps result from shared or distinct causal genetic factors (Foley et al., 2021; Giambartolomei et al., 2018; Wallace, 2021). Traditionally this analysis was restricted by the assumption of at most one causal variant for each trait in the region. However, recent extensions to the method now permit analysis based on univariate fine-mapping results for the traits compared and, therefore, analysis of regions with multiple causal variants.

Accordingly, we conducted genome-wide local genetic correlation analysis across 5 neuropsychiatric traits with recognised phenotypic and genetic overlap (Beck et al., 2013; Ferrari et al., 2017; C. Li et al., 2021; Ranganathan et al., 2020; Weintraub & Mamikonyan, 2019): AD, ALS, FTD, PD, and schizophrenia. Loci highly correlated between trait pairs were further investigated with univariate fine-mapping and bivariate colocalisation techniques to examine variants driving these associations.

8.3. Methods

8.3.1. Sampled GWAS summary statistics

We leveraged publicly-accessible summary statistics from European ancestry GWAS meta-analyses of risk for AD (Kunkle et al., 2019), ALS (van Rheenen et al., 2021), FTD (Ferrari et al., 2014), PD (Nalls et al., 2019), and schizophrenia (Trubetskoy et al., 2022). European ancestry data were selected to avoid LD mismatch between the GWAS sample and reference data from an external European population.

8.3.2. Procedure

Figure 8-1 summarises the analysis protocol for this study; further details are provided below.

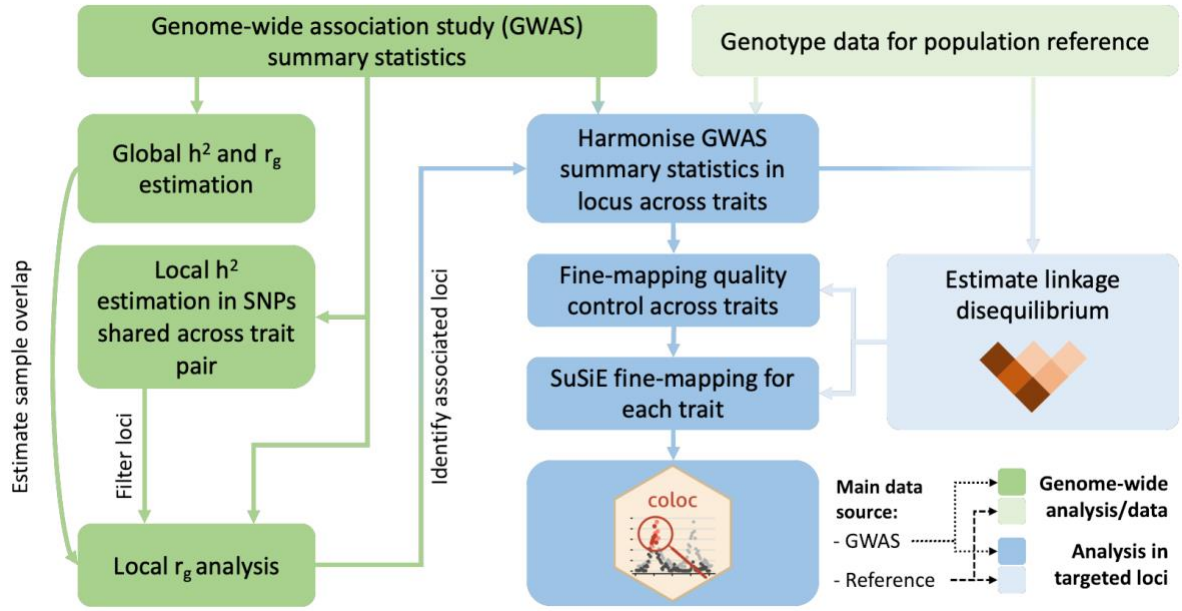


Figure 8-1. Overview of the analysis procedure for this study

SuSiE (sum of single effects) is a univariate fine-mapping approach implemented within the R package susieR. 'coloc' is an R package for bivariate colocalisation analysis between pairs of traits. h^2 = Heritability, r_g= bivariate genetic correlation. The analysis steps shaded in blue have been implemented within a readily applied analysis pipeline available on GitHub: https://github.com/ThomasPSpargo/COLOC-reporter.

8.3.2.1. Processing of GWAS summary statistics

A standard data cleaning protocol was applied to each set of summary statistics, as described in Chapter 3.2. As part of this protocol, the summary statistics were restricted to and harmonised with variants present within the 1000 Genomes phase 3 (1KG) European ancestry population (n = 503) reference dataset (Auton et al., 2015).

8.3.2.2. Genome-wide analyses

Global heritability and genetic correlations

LDSC (v1.0.1) (Bulik-Sullivan, Finucane, et al., 2015; Bulik-Sullivan, Loh, et al., 2015) was applied to estimate genome-wide univariate heritability (h^2) for each trait on the liability scale. The software was also applied to derive 'global' (i.e., genome-wide) genetic correlation estimates between trait pairs and estimate sample overlap from the bivariate intercept.

These analyses were performed with the default settings and using the HapMap3 (Altshuler et al., 2010) SNPs and the LD score files provided with the software, calculated in the 1KG European population.

Local genetic correlation analysis

LAVA (v0.1.0) (Werme et al., 2022) was applied to obtain local genetic correlation estimates across 2495 approximately independent blocks, delineating the genome based on patterns in LD. We used the blocks provided alongside the LAVA software which were derived from the 1KG European cohort. Bivariate intercepts from LDSC were provided to LAVA to estimate sample overlap between trait pairs.

In accordance with prior studies, genetic correlation analysis was performed following an initial filtering step. Univariate heritability was estimated for each genomic block across SNPs in-common between a pair of traits, and only loci with local h^2 p-values below a threshold of 2.004×10^{-5} (0.05/2495) in both traits continued to the bivariate analysis. This step ensures that univariate heritability is sufficient in both traits for a robust correlation estimate.

8.3.2.3. Targeted genetic analyses

Fine-mapping and colocalisation analysis

Statistical fine-mapping and colocalisation techniques were applied to further analyse associations between trait pairs in regions where the false discovery rate (FDR) adjusted p-value of local genetic correlation analysis was below 0.05 (after adjusting for all bivariate comparisons performed). Additional analysis was conducted at loci where significant correlations occurred between two trait pairs but not between the final pairwise comparison across the three implicated traits.

Fine-mapping was performed with *susieR* (v0.12.27) (G. Wang, Sarkar, Carbonetto, & Stephens, 2020; Y. Zou et al., 2022), which implements the 'sum of single effects' (SuSiE) model to represent statistical evidence of causal genetic variation within 'credible sets' and per-SNP posterior inclusion probabilities (PIPs). A 95% credible set indicates 95% certainty that at least one SNP included within the set has a causal association with the phenotype

and higher PIPs indicate greater posterior probability of being a causal variant within a credible set. Multiple credible sets are identified when the data suggest more than one independent causal signal.

Colocalisation analysis was implemented with *coloc* (v5.1.0.1) (Giambartolomei et al., 2014; Giambartolomei et al., 2018; Wallace, 2021), which calculates posterior probabilities that a causal variant exists for neither, one, or both of two compared traits, testing also whether evidence for a causal variant in both traits suggests a shared variant (i.e., hypothesis 4 (H4); colocalisation) or independent signals (Hypothesis 3 (H3)). Colocalisation analyses can be performed across all variants sampled in a region, under an assumption of at most one variant implicated per trait. It can also be performed using variants attributed to pairs of credible sets from SuSiE, relaxing the single variant assumption (Wallace, 2021). When evidence of a shared variant is found, the individual SNPs with the highest posterior probability for being that variant can be assessed. With a 95% confidence threshold, these are termed 95% credible SNPs.

Analysis pipeline

We conducted colocalisation and fine-mapping analysis within an open-access pipeline developed for this study using R (v4.2.2) (R Core Team, 2021): https://github.com/ThomasPSpargo/COLOC-reporter.

Briefly, in this workflow (see Figure 8-1), GWAS summary statistics are harmonised across analysed traits for a specified genomic region, including only variants in common between them and available within a reference population. An LD correlation matrix across sampled variants is derived from a reference population using PLINK (v1.90) (Purcell; Purcell et al., 2007).

Quality control is performed per-dataset prior to univariate fine-mapping analysis. Diagnostic tools provided with *susieR* are applied to test for consistency between the LD matrix and Z-scores from the GWAS and identify variants with a potential 'allele flip' (reversed effect estimate encoding) that can impact fine-mapping.

Fine-mapping is performed for each dataset with the *coloc* package *runsusie* function, which wraps around *susie_rss* from *susieR* and is configured to facilitate subsequent colocalisation analysis. Sample size (Effective sample size for binary traits) is specified as the median for SNPs analysed. Colocalisation analysis can be performed with the *coloc* functions *coloc.abf* and *coloc.susie* when fine-mapping yields at least one credible set for both traits and otherwise using *coloc.abf* only. Genes located near to credible sets from fine-mapping and credible SNPs from colocalisation analyses are identified via Ensembl and *biomaRt* (v2.54.0) (Cunningham et al., 2022; Durinck et al., 2005; Durinck, Spellman, Birney, & Huber, 2009).

Analysis parameters can be adjusted by the user in accordance with their needs. Various utilities are included to help interpretation of fine-mapping and colocalisation results, including identification of genes nearby to putatively causal signals, HTML reports to summarise completed analyses, and figures to visualise the results and compare the examined traits.

Current implementation

In this study, LD correlation matrices were derived from the 1KG European cohort. SNPs flagged for potential allele flip issues in either of the compared traits were removed from the analysis. Fine-mapping was performed with the *susie_rss* refine=TRUE option to avoid local maxima during convergence of the algorithm, leaving the other settings to the *runsusie* defaults. Colocalisation analysis was performed using the default priors for *coloc.susie* ($P_1=1x10^{-4}$, $P_2=1x10^{-4}$, $P_{12}=5x10^{-6}$).

Colocalisation and fine-mapping analyses were performed initially using the genomic blocks defined by LAVA, since these aim to define relatively independent LD partitions across the genome (Werme et al., 2022). If a 95% credible set could not be identified in one or both traits, we inspected local Manhattan plots for the region to determine whether potentially relevant signals occurred around the region boundaries. The analysis was repeated with a ±10Kb window around the LAVA-defined genomic region if p-values for SNPs at the edge of the block were $p<1x10^{-4}$ for both traits and the Manhattan plots were suggestive of a 'peak' not represented within the original boundaries.

8.4. Results

8.4.1. Genome-wide analyses

Descriptive information and heritability estimates for the sampled traits and GWAS are presented in Table 8-1. ALS had nominally significant global genetic correlations with schizophrenia (p=0.045), PD (p=0.013), and AD (p=0.006); no other bivariate genome-wide correlations were statistically significant (see Figure 8-2).

Table 8-1. Genome-wide association studies (GWAS) sampled

*Each GWAS is a GWAS meta-analysis of disease risk across people of European ancestry. *Proxy cases from the UK Biobank cohort. †Estimated from cumulative risk after age 45 after correcting for competing risk of mortality and assuming a lifespan of ~85 years. h^2 = heritability*

Trait	Estimated lifetime risk in population	GWAS			Liability scale h^2 (standard error)
		Reference	N Cases	N Controls	
Alzheimer's disease	1/10 (Chêne et al., 2015) †	Kunkle et al. (2019)	21,982	41,944	0.093 (0.0155)
Amyotrophic lateral sclerosis	1/350 (Alonso et al., 2009; Johnston et al., 2006)	van Rheenen et al. (2021)	27,205	110,881	0.0277 (0.003)
Frontotemporal dementia	1/742 (Coyle-Gilchrist et al., 2016)	Ferrari et al. (2014)	2,154	4,308	0.0329 (0.0283)
Parkinson's disease	1/37 (Parkinson's UK, 2017)	Nalls et al. (2019)	15,056 (+ 18,618 proxies*)	449,056	0.0506 (0.0046)
Schizophrenia	1/250 (Saha, Chant, Welham, & McGrath, 2005)	Trubetskoy et al. (2022)	53,386	77,258	0.1761 (0.0061)

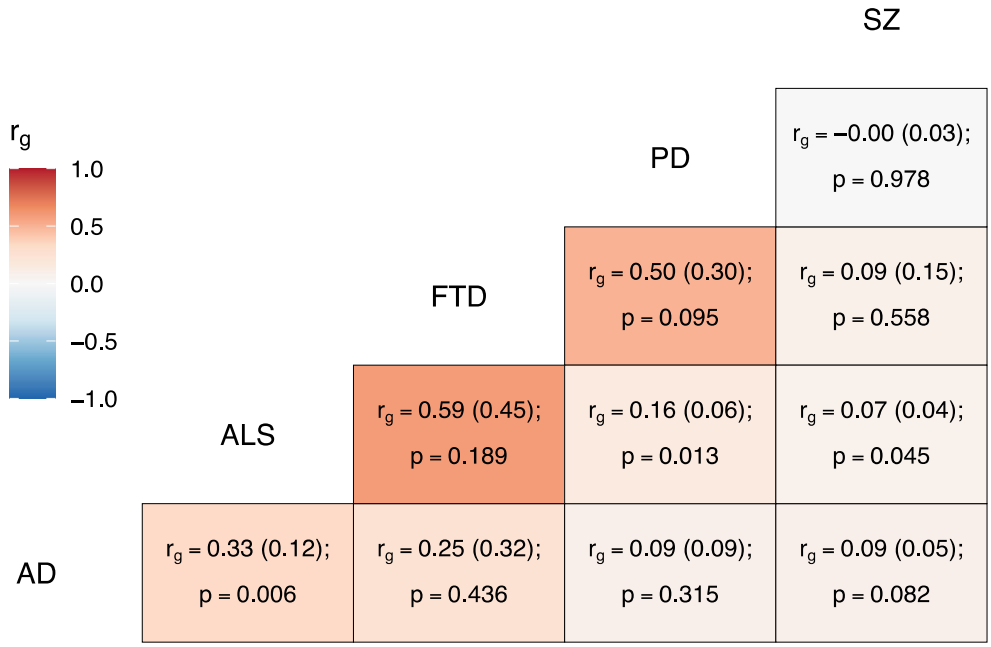


Figure 8-2. Genome-wide genetic correlation estimates between all trait pairs

The heatmap displays genetic correlations (r_g) each tile is labelled with the r_g estimate (standard error) and the p-value. AD = Alzheimer's disease, ALS = amyotrophic lateral sclerosis, FTD = frontotemporal dementia, PD = Parkinson's disease, SZ = schizophrenia.

A total of 605 local genetic correlation analyses were performed across all trait pairs in genomic regions where both traits passed the univariate heritability filtering step after restricting to SNPs sampled in both GWAS (see Table 8-2; Figure 8-3; Table E-1). The number of loci passing to bivariate analysis varied greatly across trait pairs and was congruent with the genome-wide heritability estimates (and their uncertainty) for each trait, reflecting differences in phenotypic variance explained by measured genetic variants and statistical power for each GWAS (see Table 8-1).

Table 8-2. Comparison of genome-wide SNP significance against local genetic correlation significance thresholds in all trait pairs and loci analysed

All loci analysed showed sufficient local univariate heritability across compared traits to allow bivariate correlation analysis. Subsequent fine-mapping and colocalisation analyses were performed in this study for regions with at least false discovery rate (FDR) adjusted significance for the local genetic correlation. SNP = single nucleotide polymorphism.

Number of traits in pair with genome-wide significant ($p < 5 \times 10^{-8}$) SNP in locus	Smallest significance threshold for local genetic correlation			
	Bonferroni ($p < 8.26 \times 10^{-5}$; 0.05/605)	FDR ($p_{fdr} < 0.05$)	Nominal ($p < 0.05$)	Non-significant ($p \geq 0.05$)
0	1	17	77	394
1	1	4	18	80
2	0	3	2	8

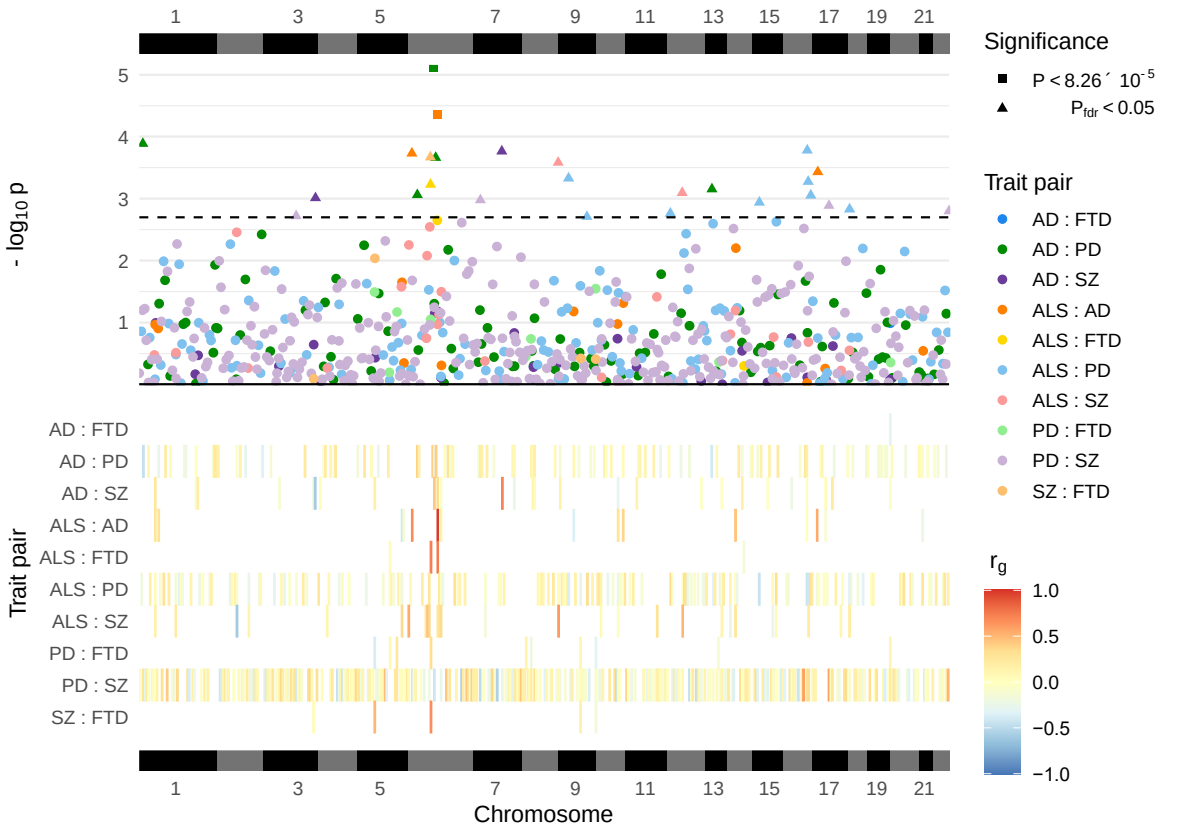


Figure 8-3. Local genetic correlation analyses between trait pairs

The lower panel displays a heatmap of genetic correlations (r_g) across genomic regions where any bivariate analyses were performed; white colouring indicates that the region was not analysed for a given trait pair owing to insufficient univariate heritability in one or both traits. The upper panel shows a Manhattan plot of p-values from each correlation analysis, denoting trait pairs by colour and comparisons passing defined significance thresholds by shape (square for a strict Bonferroni threshold and triangle for a false discovery rate (fdr) adjusted threshold); the hatched line indicates the threshold p-value above which P_{fdr}<0.05. The panels are both ordered by relative genomic position, with bars above and below indicating each chromosome. AD = Alzheimer's disease, ALS = amyotrophic lateral sclerosis, FTD = frontotemporal dementia, PD = Parkinson's disease, SZ = schizophrenia. Table E-1 provides a complete summary of local genetic correlation analyses performed.

Twenty-six bivariate comparisons were significant following FDR adjustment (p_{fdr} <0.05), two of which also passed the stringent Bonferroni threshold (p <8.26x10^{-5}; 0.05/605). While some regions included genome-wide significant SNPs (p <5x10^{-8}) for one or both traits, others occurred in regions where GWAS associations were weaker (see Table 8-2). Five of these associations occurred at loci within the human leukocyte antigen (HLA) region (GRCh37: Chr6:28.48-33.45Mb; 6p22.1-21.3 (Genome Reference Consortium)), and all five traits were implicated in at least one of these.

8.4.2. Targeted genetic analyses

Univariate fine-mapping and bivariate colocalisation analyses were subsequently performed to test for variants jointly implicated between trait pairs in regions with local genetic correlation P_{fdr} <0.05. The ALS and schizophrenia trait pair was additionally examined at Chr6:32.22-32.45Mb because significant genetic correlations were found between ALS and FTD and between schizophrenia and FTD at this locus. The correlation between ALS and schizophrenia at this locus had not been analysed owing to insufficient univariate heritability for ALS after restricting to SNPs in common with the schizophrenia GWAS.

Fine-mapping identified at least one 95% credible set for each of the compared traits for 7 of the 27 comparisons performed (see Table 8-3), and for one trait only in a further 5 (see Table E-2; Table E-3). This analysis suggested two credible sets for schizophrenia in the Chr12:56.99-58.75Mb locus, for AD in Chr6:32.45-32.54Mb, and (only when harmonised to SNPs in common with the ALS GWAS) for FTD in Chr6:32.22-32.45Mb (see Table E-3).

Colocalisation analyses performed across fine-mapping credible sets and across all SNPs in a region generally gave support to the equivalent hypothesis (Table 8-3; Table E-2). Moreover, comparisons suggesting a signal was present in one trait only were largely concordant with the identification of fine-mapping credible sets in only that trait (Table E-2). Figure E-1 compares per-SNP p-values across trait pairs for comparisons with evidence of a relevant signal in both traits. Figure E-2 shows patterns of LD across SNPs assigned to credible sets for these analyses.

Strong evidence was found for a shared variant between ALS and AD within the *HLA* region (Posterior probability of shared variant = 0.9; see Figure 8-4). The 95% credible SNPs for this association were distributed around the *MTCO3P1* pseudogene and rs9275477, the lead genome-wide significant SNP from the ALS GWAS in this region, had the highest posterior probability of being implicated in both traits. Figure E-3 presents sensitivity analysis showing that the result is robust to a range of values for the shared variant hypothesis prior probability.

The other comparisons that found fine-mapping credible sets in both traits suggested that overlaps from the correlation analysis were driven by distinct causal variants (see Table 8-3; Table E-2).

Univariate fine-mapping of PD and schizophrenia at Chr17:43.46-44.87Mb found large credible sets spanning many genes, including *MAPT* (Allen et al., 2014; Nakayama et al., 2019; Origone et al., 2018; Snowden et al., 2015) and *CRHR1* (Bigdeli et al., 2020; Cheng, Zhu, & Zhang, 2020) which have been previously implicated in the traits we have analysed. These expansive credible sets reflect the strong LD in the region and indicate a signal that is difficult to localise (see Figure E-2(F); Table E-3). The colocalisation analysis suggested independent variants for each trait despite many SNPs overlapping across their respective credible sets (see Figure E-2). Sensitivity analysis showed robust support for the two independent variants hypothesis across shared-variant hypothesis priors (Figure E-3).

Table 8-3. Colocalisation analysis conducted across 95% credible sets identified during univariate fine-mapping of trait pairs

*N SNPs refers to the number of SNPs present for both traits and the 1000 genomes reference panel in the region within colocalisation and fine-mapping analysis. *Indicates comparisons with genetic correlation analysis p <8.26x10^{-5} (0.05/605). $^{\Delta}$ Denotes locus extended by ±10kb for fine-mapping and colocalisation analysis. †Variant identified in colocalisation as having the highest posterior probability of being shared variant assuming hypothesis 4 is true (see Figure 8-4). §Differences in fine-mapping solutions across trait pairs in the Chr6:32.21-32.45Mb locus reflect differences in the SNPs retained after restricting to those in common between the compared GWAS $^{\varnothing}$ H0 = no causal variant for either trait, H1 = variant causal for trait 1, H2 = variant causal for trait 2, H3 = distinct causal variants for each trait, H4 = a shared causal variant between traits. PIP = posterior inclusion probability. AD = Alzheimer's disease, ALS = amyotrophic lateral sclerosis, FTD = frontotemporal dementia, PD = Parkinson's disease, SZ = schizophrenia.*

Trait		Genomic position (GRCh37)	Local genetic correlation estimate (95% confidence Interval)	Fine-mapping Credible set for trait		N SNPs	SNP with highest PIP for fine-mapping credible set (nearest gene; sense-strand base pair distance)		Posterior probability for hypothesis$^{\varnothing}$				
1	2			1	2		Trait 1	Trait 2	H0	H1	H2	H3	H4
AD	PD	Chr6:32576785-32639239 $^{\Delta}$	0.406 (0.197, 0.648)	1	1	958	rs9271247 (*HLA-DQA1*; +15,844)	rs3129751 (*HLA-DQA1*; +13,767)	<0.01	<0.01	<0.01	**0.95**	0.05
ALS	AD	Chr6:32629240-32682213 *	0.974 (0.717, 1.000)	1	1	475	rs9275477† (*MTCO3P1;* +1,260)	rs9275207 (*MTCO3P1*; +16,191)	<0.01	<0.01	<0.01	0.10	**0.90**
ALS	FTD	Chr6:32208902-32454577 §	0.723 (0.370, 1.000)	1	1	1709	rs9268833 (*HLA-DRB9*; 0)	rs1980493 (*BTNL2*; 0)	<0.01	<0.01	0.01	**0.99**	<0.01
					2			rs9767620 (*HLA-DRB9*; +1,498)	<0.01	<0.01	0.01	**0.99**	<0.01
ALS	SZ	Chr6:32208902-32454577 §	-	1	1	1711	rs9268833 (*HLA-DRB9*; 0)	rs9268219 (*C6orf10*; 0)	<0.01	<0.01	<0.01	**0.98**	<0.01
		Chr12:56987106-58748139	0.506 (0.218, 0.807)	1	1	2260	rs113247976 (*KIF5A;* 0)	rs12814239 (*LRP1*; 0)	<0.01	<0.01	<0.01	**1.00**	<0.01
					2			rs324017 (*NAB2*; 0)	<0.01	<0.01	<0.01	**1.00**	<0.01
PD	SZ	Chr17:43460501-44865832	0.595 (0.266, 0.950)	1	1	2453	rs58879558 (*MAPT*; 0)	rs62062288 (*MAPT*; 0)	<0.01	<0.01	<0.01	**0.81**	0.19
SZ	FTD	Chr6:32208902-32454577 §	0.669 (0.379, 0.990)	1	1	1657	rs9268219 (*C6orf10*; 0)	rs9268877 (*HLA-DRB9*; 0)	<0.01	<0.01	<0.01	**1.00**	<0.01

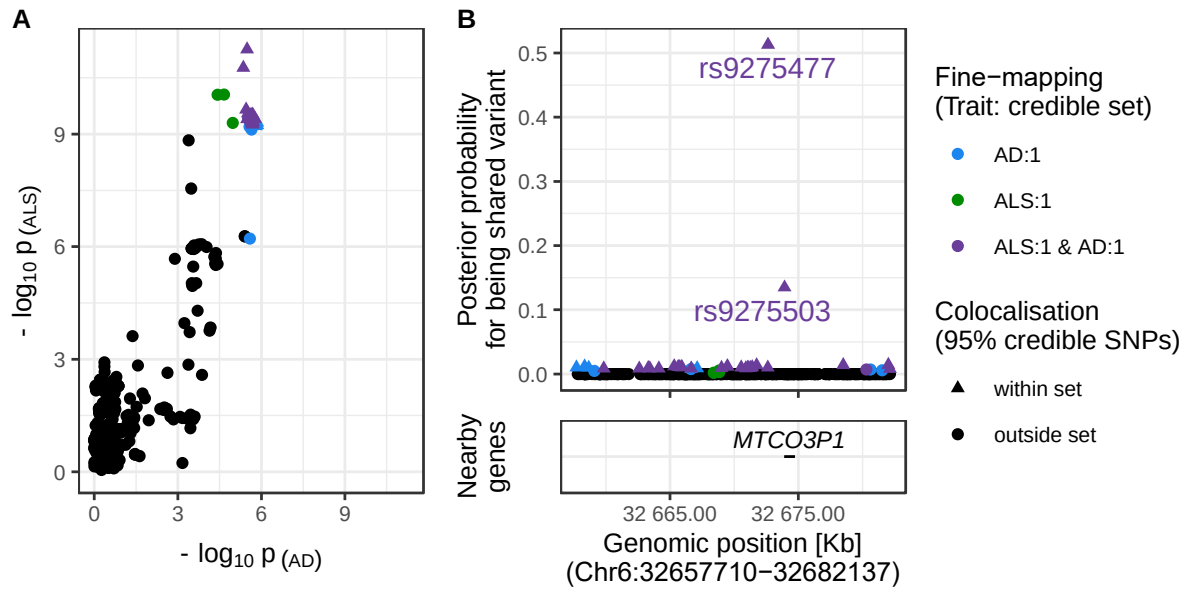


Figure 8-4. Evidence for colocalisation between amyotrophic lateral sclerosis (ALS) and Alzheimer's disease (AD) in the Chr6:32.63-32.68Mb region

***Panel A:** SNP-wise p-value distribution between ALS and AD across Chr6:32.63-32.68Mb, in which colocalisation analysis found 0.90 posterior probability of the shared variant hypothesis (see Table 8-3). **Panel B:** (upper) Per-SNP posterior probabilities for being a shared variant between ALS and AD, (lower) positions of HGNC gene symbols nearby to the 95% credible SNPs. Posterior probabilities for being a shared variant sum to 1 across all SNPs analysed and are predicated on the assumption that a shared variant exists; 95% credible SNPs are those spanned by the top 0.95 of posterior probabilities. The x-axis for Panel B is truncated by the base pair range of the credible SNPs and genomic positions are based on GRCh37.*

8.5. Discussion

We examined genetic overlaps between the neuropsychiatric conditions Alzheimer's disease, amyotrophic lateral sclerosis, frontotemporal dementia, Parkinson's disease, and schizophrenia. Associated genomic regions between pairs of traits were identified with local genetic correlation analysis and further analysed with statistical fine-mapping and colocalisation techniques.

Significant correlations were most frequent across genomic blocks within the *HLA* region, implicating each of the studied traits in at least one comparison. Several associated regions contained genes with known relevance for the traits studied, such as *KIF5A, MAPT, and CRHR1*. Colocalisation analysis found strong evidence for a shared genetic variant between ALS and AD in the Chr6:32.62-32.68Mb locus within *HLA*, while the other colocalisation analyses suggested causal signals distinct across traits, for one trait only, or for neither trait.

The tendency for association between traits around the *HLA* region is reasonable, since this is well-established hotspot for pleiotropy (Watanabe et al., 2019; Werme et al., 2022). HLA is particularly known for its role in immune response and it is implicated in various types of disease (Dendrou, Petersen, Rossjohn, & Fugger, 2018; Trowsdale & Knight, 2013). Mounting evidence has linked HLA and associated genetic variation to the traits we have analysed, and mechanisms underlying these associations are beginning to be understood (Al-Diwani, Pollak, Irani, & Lennox, 2017; Aliseychik, Andreeva, & Rogaev, 2018; Broce et al., 2018; Dendrou et al., 2018; Ferrari et al., 2016; Mokhtari & Lachman, 2016; Song et al., 2016; Trowsdale & Knight, 2013; Z.-X. Wang, Wan, & Xing, 2020; Yu et al., 2021). For instance, AD is associated with variants around the *HLA-DQA1* and *HLA-DRB1* genes and several SNPs in the non-coding region between them have been shown to modulate their expression (X. Zhang et al., 2022). Notably, one of the SNPs with a demonstrated regulatory role, rs9271247, had the highest probability of being causal for AD across the 95% credible set identified in the fine-mapping of the region.

Variants showing evidence for colocalisation between AD and ALS were distributed around the *MTCO3P1* pseudogene in the HLA class II non-coding region between *HLA-DQB1* and *HLA-DQB2. MTCO3P1* has been previously identified as one of the most pleiotropic genes in the GWAS catalog (Chesmore, Bartlett, & Williams, 2018; Sollis et al., 2022). Previous studies have suggested the relevance of this region in both traits. *HLA-DQB1* and *HLA-DQB2* are both upregulated in the spinal cord of people with ALS, alongside other genes implicated in various immunological processes for antigen processing and inflammatory response (Andrés-Benito, Moreno, Aso, Povedano, & Ferrer, 2017). HLA class II complexes, and their subcomponents, have been identified as upregulated in multiple brain regions of people with AD, using both gene and protein expression techniques (Aliseychik et al., 2018; Hopperton, Mohammad, Trépanier, Giuliano, & Bazinet, 2018).

Our analysis of this region gave stronger support for colocalisation between the ALS and AD GWAS than a previous study. The previous study defined a 100Kb window around the lead genome-wide significant SNP from the ALS GWAS, rs9275477, and found ~0.50 posterior probability for each of the shared and two independent variant(s) hypotheses (van Rheenen

et al., 2021). The difference between these studies reflects differences in processing of GWAS data; in this study all summary statistics underwent quality control to ensure only high-quality variants were retained.

More broadly, our analyses suggest that regions with strong genetic correlation between the five traits studied often result from adjacent but trait-specific signals, likely reflecting overlaps between LD blocks (Watanabe et al., 2019). Correlations also occurred in regions with weaker overall GWAS associations (see Table 8-2), where fine-mapping and colocalisation analyses did not suggest causal associations in one or either trait. Such patterns likely reflect a shared polygenic trend across the region, rather than associations attributable to discrete variants. Accordingly, other approaches may be better suited for identifying regions containing genetic variation jointly causal across diseases, including the traditional approach of testing regions around overlapping genome-wide significant variants.

This study has used gold-standard statistical tools to examine genetic relationships between traits. The local genetic correlation analysis approach enabled targeted investigation of genomic regions which appear to overlap between traits. The application of colocalisation analysis alongside a prior univariate fine-mapping step allowed for associations to be tested without conflating independent but nearby signals under the single-variant assumption of colocalisation analysis across all variants sampled in a region.

The study is not without limitation. We necessarily used the 1KG European reference population to estimate LD between SNPs. Fine-mapping is ideally performed with an LD matrix from the GWAS sample and is sensitive to misspecification when inconsistencies in LD occur between the reference and GWAS cohorts. Use of a reference population is not uncommon, however, and diagnostic tools available within the *susieR* package allow testing for inconsistencies between the reference and GWAS samples (Y. Zou et al., 2022). We accordingly implemented these tools centrally into our workflow and determined that the LD matrices from the 1KG reference were suitable for the data (estimates of Z-score and LD consistency are available in Table E-3). Nevertheless, repeating this study in under-represented populations would be an important future step to validate findings.

We employed statistical methods to identify and analyse genomic regions containing variants which might be jointly implicated across traits. These approaches provide useful associations between traits identified from large-scale genomic datasets. However, they alone are not sufficient for translation into clinical practice. Future studies should aim to extend any associations found by integrating functional and multi-omics datasets to gain mechanistic insights into observed trends and facilitate treatment discovery (Pain et al., 2023; X. Zhang et al., 2022).

The fine-mapping and colocalisation analysis pipeline we have used is available as an open-access resource on GitHub to facilitate application of these methods in future studies: https://github.com/ThomasPSpargo/COLOC-reporter. Specified genomic regions can be readily analysed by providing GWAS summary statistics for binary or quantitative traits of interest and a population-appropriate reference dataset for estimation of LD. Resources returned by the pipeline include detailed reports that overview the analyses performed.

Chapter 9. Summary and future directions

The work contributing towards this thesis consists of several investigations into the relationship between genotype and phenotype in amyotrophic lateral sclerosis (ALS). This chapter overviews the contribution of these studies towards understanding of the disease and future directions for ALS research.

9.1. Summary of findings

Chapter 4 and Chapter 5 focus on disease susceptibility. The former contributes a novel approach to calculating genetic penetrance, which is the probability of disease given a person harbours a certain variant, for autosomal dominant variants. The latter takes a broader view and uses a Bayesian framework to examine implications of genetic testing upon risk of a rare disease.

With the approach described in Chapter 4, we made novel estimates of penetrance for ALS associated with variants in the ALS-implicated genes, *SOD1*, *C9orf72*, and *FUS* (see Table 4-2*;* Table B-5). These estimates give an important insight into disease risk for people harbouring variants within these genes.

The analyses of Chapter 5 highlight the importance of considering inherent imperfections of testing for markers, genetic or otherwise, of disease risk. This is critical because understanding of increased liability of disease following a test that suggests presence of a disease-associated variant ties not only to penetrance but also the performance of the test and the other variant-disease characteristics. For instance, despite an estimated 0.701 penetrance, we calculated 0.109 probability of developing symptoms of ALS for a person found to harbour a deleterious *SOD1* variant in a test implemented within population-wide genetic screening (see Table 5-1). Likewise, for the *C9orf72* repeat expansion ($C9orf72^{RE}$), penetrance was estimated at 0.439 but ALS risk following a positive whole-genome sequencing screening result was 0.00519, increasing to 0.0198 following result confirmation from an appropriate laboratory test (Akimoto et al., 2014).

A key message from Chapter 5 is that the difference in disease risk following positive results from a diagnostic testing (e.g., testing performed based on some prior information about disease risk) and contextually-blind population screening scenario can be stark for rare outcomes such as ALS. Accordingly, it is important to reinforce positive findings with a secondary test and any results should be interpreted in the context of the prior belief about disease risk. Other broader considerations are addressed within the chapter.

Chapter 6 and Chapter 7 describe analyses of trends in ALS across different subgroups of people. The former focusses upon data-driven subgroups identified with a latent class clustering analysis approach. The latter focusses upon design and example application of a web-utility for analysing subgroups defined by *SOD1* gene variants resulting in a non-synonymous protein change.

Five distinct clusters of ALS were identified within clinical data during the investigation of data-driven subgroups from Chapter 6. Phenotypically, these 'classes' were primarily distinguished by their time from symptom onset to diagnosis (diagnostic delay) and by disease duration from symptom onset until death or censoring, although other variables also varied across them.

We found biological trends underlying these subgroups. Analysis of rare variants showed an uneven distribution of *SOD1* variants, $C9orf72^{RE}$, and of variants in genes associated with 'RNA function' and 'Cytoskeletal dynamics and axonal transport' processes across the classes. These were interpreted as indicating biological distinctions across the clinically-derived groups. The results of differential expression and functional enrichment analysis in a small subsample of the dataset supported this perspective.

Analysis of common genetic variation measured by polygenic risk scores (PRS) showed several interesting findings. PRS for risk of ALS were only higher than healthy control participants in Class 1, which contains about 50% of the dataset. This suggests that the variants captured by the ALS PRS may be biased towards single nucleotide polymorphisms associated with ALS susceptibility for those in Class 1, whilst other variants may be relevant to other subgroups. The result ties with the finding that other classes, particularly 2 and 5,

were associated with higher PRS for Parkinson's disease or schizophrenia susceptibility. This latter finding suggests that reported genetic overlaps between ALS and these neuropsychiatric traits could be driven by certain disease subgroups.

The *SOD1*-focussed study of Chapter 7 aimed to address the high degree of phenotypic variability associated with variants resulting in different SOD1 protein mutations. In one case study we exemplified the problem, demonstrating the variability in age of onset and disease duration associated with different variants at the p.G94 residue of SOD1. In a second case study, we tested the hypothesis that the ALS phenotype would differ according to the effect of different amino acid substitutions upon residue hydrophobicity. This analysis found that variants in residues classed as hydrophobic in wild type SOD1 which became more hydrophilic were associated with a shorter disease duration, suggestive of a more aggressive phenotype, than those remaining hydrophobic. Interestingly, we also observed shorter disease duration for residues classed as having intermediate hydrophobicity in wild type and mutant SOD1 compared to most other hydrophobicity groups.

Chapter 8 builds upon the evidence of genetic overlap between ALS and other traits from Chapter 6 and conducts a broader investigation of genetic variation shared between Alzheimer's disease, ALS, frontotemporal dementia, Parkinson's disease, and schizophrenia based on genome-wide association study summary statistics. The genome-wide local genetic correlation analysis suggested that overlaps between pairs of traits were most frequent within the chromosome 6 human leukocyte antigen (*HLA*) region. Further analysis of associated regions between trait pairs with statistical fine-mapping and colocalisation analysis techniques suggested a shared variant between Alzheimer's disease and ALS within HLA class II, located around the *MTCO3P1* pseudogene. These analyses suggested that the associations in other regions resulted from linkage between trait-specific variants or from a shared polygenic trend within a locus.

9.2. Future directions

Investigations performed across this thesis all contribute towards the common goal of improving understanding of the biological architecture underlying ALS and the heterogeneity of its clinical phenotype. Since these questions are broad, we have

contributed several open-access utilities to facilitate future investigations using these approaches.

In Chapter 4, we proposed a method for estimation of genetic penetrance which has relevance for many genetic variants implicated in autosomal dominant ALS, along with other autosomal dominant diseases. It enables a population-scale approach to estimate genetic penetrance that can be operated using data from only families affected by disease, avoiding biases typical for other population-scale approaches. The method can be applied using an R function available on GitHub (https://github.com/ThomasPSpargo/adpenetrance) or via a dedicated web-app (https://adpenetrance.rosalind.kcl.ac.uk/).

The 'SOD1-ALS-Browser' web-utility (https://sod1-als-browser.rosalind.kcl.ac.uk/) described within Chapter 7 was developed to provide immediate access to a large ALS dataset, containing people with and without SOD1 variants, that can be readily applied for bespoke analysis of the phenotype. The utility of this tool for elucidating phenotypic differences between groups of SOD1 variants has been shown in one recent project which defined two groups of variants according to their structural location and biophysical behaviour (M. Kalia et al., 2022).

The fine-mapping and colocalisation analysis workflow described in Chapter 8 provides a readily-implemented command-line tool for investigations of genetic overlap between traits based on genome-wide association study summary statistics. Our analysis focussed upon regions with suggested overlaps between neuropsychiatric diseases, but the pipeline is relevant to binary or quantitative traits and is well suited for analysis across multiple regions or traits, with flexible analysis settings according to the user's needs and with a detailed report overviewing each completed analysis. The pipeline can be accessed via GitHub: https://github.com/ThomasPSpargo/COLOC-reporter.

There are various important directions that future ALS research should pursue based on the investigations we present.

Further investigation of ALS risk associated with a spectrum of genetic variation is essential. Future studies should apply the method proposed in Chapter 4 to establish risks associated with additional variants linked to autosomal dominant ALS. Alternative approaches must also be implemented to assess other types of variation relevant for the disease. PRS may hold utility for this purpose. For instance, multi-trait PRS, combining SNP effect estimates for risk of ALS alongside those from other associated traits, have been shown to predict ALS case-control status and outperform PRS for ALS risk alone (Restuadi et al., 2022). These measures should however be implemented with caution as they (currently) account for a small proportion of variance in ALS risk and must be translated onto an absolute scale to be interpretable at an individual level (Pain et al., 2022; Restuadi et al., 2022). Given the complexity of ALS disease architecture, measures of individual disease risk would ideally integrate information from multiple risk factors, for instance, considering polygenic risk alongside other relevant monogenic or oligogenic variation.

Our identification and analysis of data-driven ALS subgroups opens several avenues for future research. Initial evidence of biological differences between these clusters should be followed up in future studies. It would be beneficial, for instance, to perform GWAS to identify loci that may differ between certain groups. This could be performed with the Project MinE dataset (Project MinE ALS Sequencing Consortium, 2018) but would only be reasonable for Classes 1 and 2 owing to the sample size requirements for GWAS and large disparity in the size of each subgroup. It would also be beneficial to extend investigations into other data modalities and examine the extent to which the subgroups and associations observed with pathways of ‘RNA function’ and ‘cytoskeletal dynamics and axonal transport’ translate into multi-omics datasets.

Future studies should also aim to refine the classification of disease subgroups we identified. Given our findings, it would be appropriate to integrate measures of disease progression and, given the clinical spectrum of ALS, measures of cognitive or behavioural change. A more comprehensive phenotypic dataset should improve resolution for establishing meaningful ALS subgroups and, in turn, facilitate investigations into the biological processes which distinguish them.

The biological overlap between ALS and other neuropsychiatric traits should also be further explored. Our findings align well with evidence implicating neuroinflammation within these traits (Al-Diwani et al., 2017; Chitnis & Weiner, 2017; McCauley & Baloh, 2019). A logical next step for follow-up investigations is to integrate functional datasets and experimental designs that directly test the relevance of processes associated with HLA in ALS. Further, given the association with Alzheimer's disease and evidence that anti-inflammatory medications can be protective against this disease, it is pertinent to investigate whether they also have a similar benefit for ALS (Kinney et al., 2018). Indeed, some existing work has suggested that anti-inflammatory medications have potential utility for treating the symptoms of ALS and therefore hold promise as future ALS therapies (Morello, Spampinato, Conforti, D'Agata, & Cavallaro, 2017).

References

Abdellaoui, A., Yengo, L., Verweij, K. J. H., & Visscher, P. M. (2023). 15 years of GWAS discovery: Realizing the promise. *American Journal of Human Genetics, 110*(2), 179-194. doi:https://doi.org/10.1016/j.ajhg.2022.12.011

Abel, O., Powell, J. F., Andersen, P. M., & Al-Chalabi, A. (2012). ALSoD: A user-friendly online bioinformatics tool for amyotrophic lateral sclerosis genetics. *Human Mutation, 33*(9), 1345-1351. doi:https://doi.org/10.1002/humu.22157

Adhikari, A. N., Gallagher, R. C., Wang, Y., Currier, R. J., Amatuni, G., Bassaganyas, L., . . . Brenner, S. E. (2020). The role of exome sequencing in newborn screening for inborn errors of metabolism. *Nature Medicine, 26*(9), 1392-1397. doi:https://doi.org/10.1038/s41591-020-0966-5

Akimoto, C., Volk, A. E., van Blitterswijk, M., Van den Broeck, M., Leblond, C. S., Lumbroso, S., . . . Kubisch, C. (2014). A blinded international study on the reliability of genetic testing for GGGGCC-repeat expansions in C9orf72 reveals marked differences in results among 14 laboratories. *Journal of Medical Genetics, 51*(6), 419. doi:https://doi.org/10.1136/jmedgenet-2014-102360

Al Sultan, A. A., Waller, R., Heath, P. R., & Kirby, J. (2016). The genetics of amyotrophic lateral sclerosis: current insights. *Degenerative Neurological and Neuromuscular Disease*, 49. doi:https://doi.org/10.2147/dnnd.s84956

Al-Chalabi, A. (2017). Perspective: Don't keep it in the family. *Nature, 550*(7676), S112-S112. doi:https://doi.org/10.1038/550S112a

Al-Chalabi, A., Calvo, A., Chio, A., Colville, S., Ellis, C. M., Hardiman, O., . . . Pearce, N. (2014). Analysis of amyotrophic lateral sclerosis as a multistep process: a population-based modelling study. *Lancet Neurology, 13*(11), 1108-1113. doi:https://doi.org/10.1016/s1474-4422(14)70219-4

Al-Chalabi, A., Fang, F., Hanby, M. F., Leigh, P. N., Shaw, C. E., Ye, W., & Rijsdijk, F. (2010). An estimate of amyotrophic lateral sclerosis heritability using twin data. *Journal of Neurology, Neurosurgery and Psychiatry, 81*(12), 1324. doi:https://doi.org/10.1136/jnnp.2010.207464

Al-Chalabi, A., & Hardiman, O. (2013). The epidemiology of ALS: a conspiracy of genes, environment and time. *Nature Reviews Neurology, 9*(11), 617-628. doi:https://doi.org/10.1038/nrneurol.2013.203

Al-Chalabi, A., Hardiman, O., Kiernan, M. C., Chiò, A., Rix-Brooks, B., & van den Berg, L. H. (2016). Amyotrophic lateral sclerosis: moving towards a new classification system. *Lancet Neurology, 15*(11), 1182-1194. doi:https://doi.org/10.1016/S1474-4422(16)30199-5

Al-Chalabi, A., & Lewis, C. M. (2011). Modelling the Effects of Penetrance and Family Size on Rates of Sporadic and Familial Disease. *Human Heredity, 71*(4), 281-288. doi:https://doi.org/10.1159/000330167

Al-Chalabi, A., van den Berg, L. H., & Veldink, J. (2017). Gene discovery in amyotrophic lateral sclerosis: implications for clinical management. *Nature Reviews Neurology, 13*(2), 96-104. doi:https://doi.org/10.1038/nrneurol.2016.182

Al-Chalabi, A., & Visscher, P. M. (2014). Motor neuron disease: Common genetic variants and the heritability of ALS. *Nature Reviews Neurology, 10*(10), 549-550. doi:https://doi.org/10.1038/nrneurol.2014.166

Al-Diwani, A. A. J., Pollak, T. A., Irani, S. R., & Lennox, B. R. (2017). Psychosis: an autoimmune disease? *Immunology, 152*(3), 388-401. doi:https://doi.org/10.1111/imm.12795

Aldred, M. A., Vijayakrishnan, J., James, V., Soubrier, F., Gomez-Sanchez, M. A., Martensson, G., . . . Trembath, R. C. (2006). *BMPR2* gene rearrangements account for a significant proportion of mutations in familial and idiopathic pulmonary arterial hypertension. *Human Mutation, 27*(2), 212-213. doi:10.1002/humu.9398

Aliseychik, M. P., Andreeva, T. V., & Rogaev, E. I. (2018). Immunogenetic Factors of Neurodegenerative Diseases: The Role of HLA Class II. *Biochemistry (Moscow), 83*(9), 1104-1116. doi:https://doi.org/10.1134/S0006297918090122

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., . . . Iannone, R. (2022). rmarkdown: Dynamic Documents for R (Version R package version 2.21.). Retrieved from https://rmarkdown.rstudio.com

Allen, M., Kachadoorian, M., Quicksall, Z., Zou, F., Chai, H. S., Younkin, C., . . . Alzheimer's Disease Genetics, C. (2014). Association of MAPT haplotypes with Alzheimer's disease risk and MAPT brain gene expression levels. *Alzheimer's Research & Therapy, 6*(4), 39. doi:https://doi.org/10.1186/alzrt268

Alonso, A., Logroscino, G., Jick, S. S., & Hernán, M. A. (2009). Incidence and lifetime risk of motor neuron disease in the United Kingdom: a population-based study. *European Journal of Neurology, 16*(6), 745-751. doi:https://doi.org/10.1111/j.1468-1331.2009.02586.x

ALS Variant Server. ALS Variant Server. Retrieved 02/2021 http://als.umassmed.edu/

Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., . . . Scientific, m. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature, 467*(7311), 52-58. doi:https://doi.org/10.1038/nature09298

Amado, D. A., & Davidson, B. L. (2021). Gene therapy for ALS: A review. *Molecular Therapy, 29*(12), 3345-3358. doi:https://doi.org/10.1016/j.ymthe.2021.04.008

Anders, S., Pyl, P. T., & Huber, W. (2014). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics, 31*(2), 166-169. doi:https://doi.org/10.1093/bioinformatics/btu638

Andersen, P. M. (2006). Amyotrophic lateral sclerosis associated with mutations in the CuZn superoxide dismutase gene. *Current Neurology and Neuroscience Reports, 6*(1), 37-46. doi:https://doi.org/10.1007/s11910-996-0008-9

Andrés-Benito, P., Moreno, J., Aso, E., Povedano, M., & Ferrer, I. (2017). Amyotrophic lateral sclerosis, gene deregulation in the anterior horn of the spinal cord and frontal cortex area 8: implications in frontotemporal lobar degeneration. *Aging, 9*(3), 823-851. doi:https://doi.org/10.18632/aging.101195

Aragon, T. J. (2020). epitools: Epidemiology Tools (Version 0.5.10.1).

Arel-Bundock, V., Enevoldsen, N., & Yetman, C. (2018). countrycode: An R package to convert country names and country codes. *Journal of Open Source Software, 3*(28), 848. doi:https://doi.org/10.21105/joss.00848

Attali, D. (2020). shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds (Version R package version 2.1.0). Retrieved from https://CRAN.R-project.org/package=shinyjs

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., . . . National Eye Institute, N. I. H. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:https://doi.org/10.1038/nature15393

Balendra, R., & Isaacs, A. M. (2018). C9orf72-mediated ALS and FTD: multiple pathways to disease. *Nature Reviews Neurology, 14*(9), 544-558. doi:https://doi.org/10.1038/s41582-018-0047-2

Bali, T., Self, W., Liu, J., Siddique, T., Wang, L. H., Bird, T. D., . . . Miller, T. M. (2017). Defining SOD1 ALS natural history to guide therapeutic clinical trial design. *Journal of Neurology, Neurosurgery and Psychiatry, 88*(2), 99. doi:https://doi.org/10.1136/jnnp-2016-313521

Bandres-Ciga, S., Noyce, A. J., Hemani, G., Nicolas, A., Calvo, A., Mora, G., . . . Traynor, B. J. (2019). Shared polygenic risk and causal inferences in amyotrophic lateral sclerosis. *Annals of Neurology, 85*(4), 470-481. doi:https://doi.org/10.1002/ana.25431

Beck, J., Poulter, M., Hensman, D., Rohrer, J. D., Mahoney, C. J., Adamson, G., . . . Mead, S. (2013). Large *C9orf72* hexanucleotide repeat expansions are seen in multiple neurodegenerative syndromes and are more frequent than expected in the UK population. *American Journal of Human Genetics, 92*(3), 345-353. doi:https://doi.org/10.1016/j.ajhg.2013.01.011

Bede, P., Murad, A., Lope, J., Hardiman, O., & Chang, K. M. (2022). Clusters of anatomical disease-burden patterns in ALS: a data-driven approach confirms radiological subtypes. *Journal of Neurology, 269*(8), 4404-4413. doi:https://doi.org/10.1007/s00415-022-11081-3

Bede, P., Murad, A., Lope, J., Li Hi Shing, S., Finegan, E., Chipika, R. H., . . . Chang, K. M. (2022). Phenotypic categorisation of individual subjects with motor neuron disease based on radiological disease burden patterns: A machine-learning approach. *Journal of the Neurological Sciences, 432*, 120079. doi:https://doi.org/10.1016/j.jns.2021.120079

Beeldman, E., Raaphorst, J., Klein Twennaar, M., de Visser, M., Schmand, B. A., & de Haan, R. J. (2016). The cognitive profile of ALS: a systematic review and meta-analysis update. *Journal of Neurology, Neurosurgery and Psychiatry, 87*(6), 611-619. doi:https://doi.org/10.1136/jnnp-2015-310734

Benn, D. E., Zhu, Y., Andrews, K. A., Wilding, M., Duncan, E. L., Dwight, T., . . . Clifton-Bligh, R. J. (2018). Bayesian approach to determining penetrance of pathogenic SDH variants. *Journal of Medical Genetics, 55*(11), 729-734. doi:https://doi.org/10.1136/jmedgenet-2018-105427

Bensimon, G., Lacomblez, L., & Meininger, V. (1994). A Controlled Trial of Riluzole in Amyotrophic Lateral Sclerosis. *New England Journal of Medicine, 330*(9), 585-591. doi:https://doi.org/10.1056/nejm199403033300901

Bertram, L., & Tanzi, R. E. (2005). The genetic epidemiology of neurodegenerative disease. *The Journal of Clinical Investigation, 115*(6), 1449-1457. doi:https://doi.org/10.1172/JCI24761

Bick, D., Bick, S. L., Dimmock, D. P., Fowler, T. A., Caulfield, M. J., & Scott, R. H. (2021). An online compendium of treatable genetic disorders. *American Journal of Medical Genetics Seminars in Medical Genetics, 187*(1), 48-54. doi:https://doi.org/10.1002/ajmg.c.31874

Biesecker, L. G. (2019). Genomic screening and genomic diagnostic testing-two very different kettles of fish. *Genome Medicine, 11*(1), 75. doi:https://doi.org/10.1186/s13073-019-0696-9

Bigdeli, T. B., Fanous, A. H., Li, Y., Rajeevan, N., Sayward, F., Genovese, G., . . . Harvey, P. D. (2020). Genome-Wide Association Studies of Schizophrenia and Bipolar Disorder in a

Diverse Cohort of US Veterans. *Schizophrenia Bulletin, 47*(2), 517-529. doi:https://doi.org/10.1093/schbul/sbaa133

Boeve, B. F., Boylan, K. B., Graff-Radford, N. R., DeJesus-Hernandez, M., Knopman, D. S., Pedraza, O., . . . Rademakers, R. (2012). Characterization of frontotemporal dementia and/or amyotrophic lateral sclerosis associated with the GGGGCC repeat expansion in C9ORF72. *Brain, 135*(Pt 3), 765-783. doi:https://doi.org/10.1093/brain/aws004

Bora, E. (2017). Meta-analysis of social cognition in amyotrophic lateral sclerosis. *Cortex, 88*, 1-7. doi:https://doi.org/10.1016/j.cortex.2016.11.012

Broce, I., Karch, C. M., Wen, N., Fan, C. C., Wang, Y., Hong Tan, C., . . . Sugrue, L. P. (2018). Immune-related genetic enrichment in frontotemporal dementia: An analysis of genome-wide association studies. *PLoS Medicine, 15*(1), e1002487. doi:https://doi.org/10.1371/journal.pmed.1002487

Brown, A.-L., Wilkins, O. G., Keuss, M. J., Hill, S. E., Zanovello, M., Lee, W. C., . . . Fratta, P. (2021). Common ALS/FTD risk variants in *UNC13A* exacerbate its cryptic splicing and loss upon TDP-43 mislocalization. *bioRxiv*, 2021.2004.2002.438170. doi:https://doi.org/10.1101/2021.04.02.438170

Brown, R. H., & Al-Chalabi, A. (2017). Amyotrophic Lateral Sclerosis. *New England Journal of Medicine, 377*(2), 162-172. doi:https://doi.org/10.1056/NEJMra1603471

Bulik-Sullivan, B. K., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., . . . Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control, C. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics, 47*(11), 1236-1241. doi:https://doi.org/10.1038/ng.3406

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., . . . Schizophrenia Working Group of the Psychiatric Genomics, C. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics, 47*(3), 291-295. doi:https://doi.org/10.1038/ng.3211

Bunton-Stasyshyn, R. K. A., Saccon, R. A., Fratta, P., & Fisher, E. M. C. (2015). SOD1 Function and Its Implications for Amyotrophic Lateral Sclerosis Pathology:New and Renascent Themes. *The Neuroscientist, 21*(5), 519-529. doi:https://doi.org/10.1177/1073858414561795

Burk, K., & Pasterkamp, R. J. (2019). Disrupted neuronal trafficking in amyotrophic lateral sclerosis. *Acta neuropathologica, 137*(6), 859-877. doi:https://doi.org/10.1007/s00401-019-01964-7

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., . . . Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature, 562*(7726), 203-209. doi:https://doi.org/10.1038/s41586-018-0579-z

Byrne, S., Elamin, M., Bede, P., Shatunov, A., Walsh, C., Corr, B., . . . Hardiman, O. (2012). Cognitive and clinical characteristics of patients with amyotrophic lateral sclerosis carrying a C9orf72 repeat expansion: a population-based cohort study. *Lancet Neurology, 11*(3), 232-240. doi:https://doi.org/10.1016/S1474-4422(12)70014-5

Byrne, S., Heverin, M., Elamin, M., Bede, P., Lynch, C., Kenna, K., . . . Hardiman, O. (2013). 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. *Annals of Neurology, 74*(5), 699-708. doi:https://doi.org/10.1002/ana.23969

Byrne, S., Jordan, I., Elamin, M., & Hardiman, O. (2013). Age at onset of amyotrophic lateral sclerosis is proportional to life expectancy. *Amyotrophic Lateral Sclerosis and*

Frontotemporal Degeneration, 14(7-8), 604-607. doi:https://doi.org/10.3109/21678421.2013.809122

Byrne, S., Walsh, C., Lynch, C., Bede, P., Elamin, M., Kenna, K., . . . Hardiman, O. (2011). Rate of familial amyotrophic lateral sclerosis: a systematic review and meta-analysis. *Journal of Neurology, Neurosurgery and Psychiatry, 82*(6), 623-627. doi:https://doi.org/10.1136/jnnp.2010.224501

Cady, J., Allred, P., Bali, T., Pestronk, A., Goate, A., Miller, T. M., . . . Baloh, R. H. (2015). Amyotrophic lateral sclerosis onset is influenced by the burden of rare variants in known amyotrophic lateral sclerosis genes. *Annals of Neurology, 77*(1), 100-113. doi:https://doi.org/10.1002/ana.24306

Cameron, D. L., Schröder, J., Penington, J. S., Do, H., Molania, R., Dobrovic, A., . . . Papenfuss, A. T. (2017). GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Research, 27*(12), 2050-2060. doi:https://doi.org/10.1101/gr.222109.117

Castellanos-Montiel, M. J., Chaineau, M., & Durcan, T. M. (2020). The Neglected Genes of ALS: Cytoskeletal Dynamics Impact Synaptic Degeneration in ALS. *Frontiers in Cellular Neuroscience, 14*. doi:https://doi.org/10.3389/fncel.2020.594975

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13*(2), 195-212. doi:https://doi.org/10.1007/BF01246098

Centers for Disease Control and Prevention. (2021). Use of Genomics in Newborn Screening Programs: The Promise and Challenges. Retrieved from https://www.cdc.gov/genomics/events/newborn_screening_2021.htm

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., . . . Borges, B. (2022). shiny: Web Application Framework for R (Version R package version 1.7.3.). Retrieved from https://CRAN.R-project.org/package=shiny

Chen, H., Kankel, M. W., Su, S. C., Han, S. W. S., & Ofengeim, D. (2018). Exploring the genetics and non-cell autonomous mechanisms underlying ALS/FTLD. *Cell Death & Differentiation, 25*(4), 648-662. doi:https://doi.org/10.1038/s41418-018-0060-4

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., . . . Yuan, J. (2022). xgboost: Extreme Gradient Boosting (Version R package version 1.7.1.1). Retrieved from https://github.com/dmlc/xgboost

Chêne, G., Beiser, A., Au, R., Preis, S. R., Wolf, P. A., Dufouil, C., & Seshadri, S. (2015). Gender and incidence of dementia in the Framingham Heart Study from mid-adult life. *Alzheimer's & Dementia, 11*(3), 310-320. doi:https://doi.org/10.1016/j.jalz.2013.10.005

Cheng, W. W., Zhu, Q., & Zhang, H. Y. (2020). Identifying Risk Genes and Interpreting Pathogenesis for Parkinson's Disease by a Multiomics Analysis. *Genes, 11*(9). doi:https://doi.org/10.3390/genes11091100

Chesmore, K., Bartlett, J., & Williams, S. M. (2018). The ubiquity of pleiotropy in human disease. *Human Genetics, 137*(1), 39-44. doi:https://doi.org/10.1007/s00439-017-1854-z

Chia, R., Chiò, A., & Traynor, B. J. (2018). Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications. *Lancet Neurology, 17*(1), 94-102. doi:https://doi.org/10.1016/S1474-4422(17)30401-5

Chiò, A., Battistini, S., Calvo, A., Caponnetto, C., Conforti, F. L., Corbo, M., . . . Surbone, A. (2014). Genetic counselling in ALS: facts, uncertainties and clinical suggestions.

Journal of Neurology, Neurosurgery and Psychiatry, 85(5), 478. doi:https://doi.org/10.1136/jnnp-2013-305546

Chiò, A., Calvo, A., Moglia, C., Mazzini, L., & Mora, G. (2011). Phenotypic heterogeneity of amyotrophic lateral sclerosis: a population based study. *Journal of Neurology, Neurosurgery and Psychiatry, 82*(7), 740-746. doi:10.1136/jnnp.2010.235952

Chiò, A., Logroscino, G., Hardiman, O., Swingler, R., Mitchell, D., Beghi, E., . . . On Behalf of the EURALS Consortium. (2009). Prognostic factors in ALS: A critical review. *Amyotrophic lateral sclerosis: official publication of the World Federation of Neurology Research Group on Motor Neuron Diseases, 10*(5-6), 310-323. doi:https://doi.org/10.3109/17482960802566824

Chiò, A., Logroscino, G., Traynor, B. J., Collins, J., Simeone, J. C., Goldstein, L. A., & White, L. A. (2013). Global epidemiology of amyotrophic lateral sclerosis: a systematic review of the published literature. *Neuroepidemiology, 41*(2), 118-130. doi:https://doi.org/10.1159/000351153

Chiò, A., Mazzini, L., Alfonso, S., Corrado, L., Canosa, A., Moglia, C., . . . Al-Chalabi, A. (2018). The multistep hypothesis of ALS revisited. *Neurology, 91*(7), e635. doi:https://doi.org/10.1212/WNL.0000000000005996

Chitnis, T., & Weiner, H. L. (2017). CNS inflammation and neurodegeneration. *The Journal of Clinical Investigation, 127*(10), 3577-3587. doi:https://doi.org/10.1172/JCI90609

Cirulli, E. T., Lasseigne, B. N., Petrovski, S., Sapp, P. C., Dion, P. A., Leblond, C. S., . . . Goldstein, D. B. (2015). Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science, 347*(6229), 1436-1441. doi:https://doi.org/10.1126/science.aaa3650

Cooper-Knock, J., Kirby, J., Highley, R., & Shaw, P. J. (2015). The Spectrum of C9orf72-mediated Neurodegeneration and Amyotrophic Lateral Sclerosis. *Neurotherapeutics, 12*(2), 326-339. doi:https://doi.org/10.1007/s13311-015-0342-1

Cordes, M. H. J., & Sauer, R. T. (1999). Tolerance of a protein to multiple polar-to-hydrophobic surface substitutions. *Protein Science, 8*(2), 318-325. doi:https://doi.org/10.1110/ps.8.2.318

Corris, P. A., & Seeger, W. (2020). 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. *American Journal of Physiology-Lung Cellular and Molecular Physiology, 318*(5), L992-L994. doi:https://doi.org/10.1152/ajplung.00098.2020

Covert, I., & Lee, S.-I. (2021). *Improving KernelSHAP: Practical Shapley Value Estimation Using Linear Regression*. Paper presented at the Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. https://proceedings.mlr.press/v130/covert21a.html

Coyle-Gilchrist, I. T. S., Dick, K. M., Patterson, K., Vázquez Rodríquez, P., Wehmann, E., Wilcox, A., . . . Rowe, J. B. (2016). Prevalence, characteristics, and survival of frontotemporal lobar degeneration syndromes. *Neurology, 86*(18), 1736. doi:https://doi.org/10.1212/WNL.0000000000002638

Crockford, C., Newton, J., Lonergan, K., Chiwera, T., Booth, T., Chandran, S., . . . Abrahams, S. (2018). ALS-specific cognitive and behavior changes associated with advancing disease stage in ALS. *Neurology, 91*(15), e1370. doi:https://doi.org/10.1212/WNL.0000000000006317

Cudkowicz, M. E., McKenna-Yasek, D., Sapp, P. E., Chin, W., Geller, B., Hayden, D. L., . . . Brown, R. H. (1997). Epidemiology of mutations in superoxide dismutase in amyotrophic lateral sclerosis. *Annals of Neurology, 41*(2), 210-221. doi:https://doi.org/10.1002/ana.410410212

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M R., Armean, Irina M., . . . Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research, 50*(D1), D988-D995. doi:https://doi.org/10.1093/nar/gkab1049

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics, 12*(7), 499-510. doi:https://doi.org/10.1038/nrg3012

DeJesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., . . . Rademakers, R. (2011). Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron, 72*(2), 245-256. doi:https://doi.org/10.1016/j.neuron.2011.09.011

Dekker, A. M., Diekstra, F. P., Pulit, S. L., Tazelaar, G. H. P., van der Spek, R. A., van Rheenen, W., . . . Veldink, J. H. (2019). Exome array analysis of rare and low frequency variants in amyotrophic lateral sclerosis. *Scientific Reports, 9*(1), 5931. doi:https://doi.org/10.1038/s41598-019-42091-3

Dendrou, C. A., Petersen, J., Rossjohn, J., & Fugger, L. (2018). HLA variation and disease. *Nature Reviews Immunology, 18*(5), 325-339. doi:https://doi.org/10.1038/nri.2017.143

Deng, H. X., Chen, W., Hong, S. T., Boycott, K. M., Gorrie, G. H., Siddique, N., . . . Siddique, T. (2011). Mutations in UBQLN2 cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia. *Nature, 477*(7363), 211-215. doi:https://doi.org/10.1038/nature10353

Dewey, F. E., Grove, M. E., Pan, C., Goldstein, B. A., Bernstein, J. A., Chaib, H., . . . Quertermous, T. (2014). Clinical Interpretation and Implications of Whole-Genome Sequencing. *JAMA, 311*(10), 1035-1045. doi:https://doi.org/10.1001/jama.2014.1717

Dickinson, J. A., Pimlott, N., Grad, R., Singh, H., Szafran, O., Wilson, B. J., . . . Bell, N. R. (2018). Screening: when things go wrong. *Canadian family physician Medecin de famille canadien, 64*(7), 502-508.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. doi:https://doi.org/10.1093/bioinformatics/bts635

Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., . . . Eberle, M. A. (2019). ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics, 35*(22), 4754-4756. doi:https://doi.org/10.1093/bioinformatics/btz431

Dolzhenko, E., van Vugt, J. J. F. A., Shaw, R. J., Bekritsky, M. A., van Blitterswijk, M., Narzisi, G., . . . Eberle, M. A. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Research, 27*(11), 1895-1903. doi:https://doi.org/10.1101/gr.225672.117

Dorsey, E. R., & Huntington Study Group, C. I. (2012). Characterization of a large group of individuals with huntington disease and their relatives enrolled in the COHORT study. *PLoS One, 7*(2), e29522-e29522. doi:https://doi.org/10.1371/journal.pone.0029522

Dukic, S., McMackin, R., Costello, E., Metzger, M., Buxo, T., Fasano, A., . . . Nasseroleslami, B. (2021). Resting-state EEG reveals four subphenotypes of amyotrophic lateral sclerosis. *Brain, 145*(2), 621-631. doi:https://doi.org/10.1093/brain/awab322

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics, 21*(16), 3439-3440. doi:https://doi.org/10.1093/bioinformatics/bti525

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols, 4*(8), 1184-1191. doi:https://doi.org/10.1038/nprot.2009.97

Dyson, H. J., Wright, P. E., & Scheraga, H. A. (2006). The role of hydrophobic interactions in initiation and propagation of protein folding. *Proceedings of the National Academy of Sciences of the United States of America, 103*(35), 13057-13061. doi:https://doi.org/10.1073/pnas.0605504103

Elbaz, A., Grigoletto, F., Baldereschi, M., Breteler, M. M., Manubens-Bertran, J. M., Lopez-Pousa, S., . . . Rocca, W. A. (1999). Familial aggregation of Parkinson's disease. *Neurology, 52*(9), 1876. doi:https://doi.org/10.1212/WNL.52.9.1876

Elden, A. C., Kim, H.-J., Hart, M. P., Chen-Plotkin, A. S., Johnson, B. S., Fang, X., . . . Gitler, A. D. (2010). Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature, 466*(7310), 1069-1075. doi:https://doi.org/10.1038/nature09320

Escott-Price, V., & Schmidt, K. M. (2021). Probability of Alzheimer's disease based on common and rare genetic variants. *Alzheimer's Research & Therapy, 13*(1), 140. doi:https://doi.org/10.1186/s13195-021-00884-7

Eshima, J., O'Connor, S. A., Marschall, E., Bowser, R., Plaisier, C. L., Smith, B. S., & Consortium, N. A. (2023). Molecular subtypes of ALS are associated with differences in patient prognosis. *Nature Communications, 14*(1), 95. doi:https://doi.org/10.1038/s41467-022-35494-w

Evans, J. D. W., Girerd, B., Montani, D., Wang, X.-J., Galiè, N., Austin, E. D., . . . Morrell, N. W. (2016). *BMPR2* mutations and survival in pulmonary arterial hypertension: an individual participant data meta-analysis. *The Lancet Respiratory Medicine, 4*(2), 129-137. doi:https://doi.org/10.1016/S2213-2600(15)00544-5

Ezer, S., Daana, M., Park, J. H., Yanovsky-Dagan, S., Nordström, U., Basal, A., . . . Harel, T. (2021). Infantile SOD1 deficiency syndrome caused by a homozygous *SOD1* variant with absence of enzyme activity. *Brain, 145*(3), 872-878. doi:https://doi.org/10.1093/brain/awab416

Faghri, F., Brunn, F., Dadu, A., Chiò, A., Calvo, A., Moglia, C., . . . Chiò, A. (2022). Identifying and predicting amyotrophic lateral sclerosis clinical subgroups: a population-based machine-learning study. *The Lancet Digital Health, 4*(5), e359-e369. doi:https://doi.org/10.1016/S2589-7500(21)00274-0

Fang, T., Jozsa, F., & Al-Chalabi, A. (2017). Nonmotor Symptoms in Amyotrophic Lateral Sclerosis: A Systematic Review. *International Review of Neurobiology, 134*, 1409-1441. doi:https://doi.org/10.1016/bs.irn.2017.04.009

Farg, M. A., Konopka, A., Soo, K. Y., Ito, D., & Atkin, J. D. (2017). The DNA damage response (DDR) is induced by the C9orf72 repeat expansion in amyotrophic lateral sclerosis. *Human Molecular Genetics, 26*(15), 2882-2896. doi:https://doi.org/10.1093/hmg/ddx170

Farhan, S. M. K., Howrigan, D. P., Abbott, L. E., Klim, J. R., Topp, S. D., Byrnes, A. E., . . . Consortium, C. R. (2019). Exome sequencing in amyotrophic lateral sclerosis implicates a novel gene, DNAJC7, encoding a heat-shock protein. *Nature Neuroscience, 22*(12), 1966-1974. doi:https://doi.org/10.1038/s41593-019-0530-0

Farrimond, L., & Talbot, K. (2022). A case of SOD1 deficiency: implications for clinical trials. *Brain, 145*(3), 805-806. doi:https://doi.org/10.1093/brain/awac063

Ferentinos, P., Paparrigopoulos, T., Rentzos, M., Zouvelou, V., Alexakis, T., & Evdokimidis, I. (2011). Prevalence of major depression in ALS: Comparison of a semi-structured interview and four self-report measures. *Amyotrophic Lateral Sclerosis, 12*(4), 297-302. doi:https://doi.org/10.3109/17482968.2011.556744

Ferini-Strambi, L., Marelli, S., Galbiati, A., Rinaldi, F., & Giora, E. (2014). REM Sleep Behavior Disorder (RBD) as a marker of neurodegenerative disorders. *Archives Italiennes de Biologie, 152*(2-3), 129-146. doi:https://doi.org/10.12871/000298292014238

Ferrari, R., Forabosco, P., Vandrovcova, J., Botía, J. A., Guelfi, S., Warren, J. D., . . . Consortium, U. K. B. E. (2016). Frontotemporal dementia: insights into the biological underpinnings of disease through gene co-expression network analysis. *Molecular Neurodegeneration, 11*(1), 21. doi:https://doi.org/10.1186/s13024-016-0085-4

Ferrari, R., Hernandez, D. G., Nalls, M. A., Rohrer, J. D., Ramasamy, A., Kwok, J. B. J., . . . Momeni, P. (2014). Frontotemporal dementia and its subtypes: a genome-wide association study. *Lancet Neurology, 13*(7), 686-699. doi:https://doi.org/10.1016/S1474-4422(14)70065-1

Ferrari, R., Wang, Y., Vandrovcova, J., Guelfi, S., Witeolar, A., Karch, C. M., . . . Desikan, R. S. (2017). Genetic architecture of sporadic frontotemporal dementia and overlap with Alzheimer's and Parkinson's diseases. *Journal of Neurology, Neurosurgery and Psychiatry, 88*(2), 152-164. doi:https://doi.org/10.1136/jnnp-2016-314411

Finegan, E., Chipika, R. H., Shing, S. L. H., Hardiman, O., & Bede, P. (2019). Primary lateral sclerosis: a distinct entity or part of the ALS spectrum? *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 20*(3-4), 133-145. doi:https://doi.org/10.1080/21678421.2018.1550518

Fogh, I., Ratti, A., Gellera, C., Lin, K., Tiloca, C., Moskvina, V., . . . Zheng, J. G. (2014). A genome-wide association meta-analysis identifies a novel locus at 17q11.2 associated with sporadic amyotrophic lateral sclerosis. *Human Molecular Genetics, 23*(8), 2220-2231. doi:https://doi.org/10.1093/hmg/ddt587

Foley, C. N., Staley, J. R., Breen, P. G., Sun, B. B., Kirk, P. D. W., Burgess, S., & Howson, J. M. M. (2021). A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature Communications, 12*(1), 764. doi:https://doi.org/10.1038/s41467-020-20885-8

Freischmidt, A., Wieland, T., Richter, B., Ruf, W., Schaeffer, V., Müller, K., . . . Weishaupt, J. H. (2015). Haploinsufficiency of TBK1 causes familial ALS and fronto-temporal dementia. *Nature Neuroscience, 18*(5), 631-636. doi:https://doi.org/10.1038/nn.4000

Fukunaga, K., Shinoda, Y., & Tagashira, H. (2015). The role of SIGMAR1 gene mutation and mitochondrial dysfunction in amyotrophic lateral sclerosis. *Journal of Pharmacological Sciences, 127*(1), 36-41. doi:https://doi.org/10.1016/j.jphs.2014.12.012

Ganesalingam, J., Stahl, D., Wijesekera, L., Galtrey, C., Shaw, C. E., Leigh, P. N., & Al-Chalabi, A. (2009). Latent Cluster Analysis of ALS Phenotypes Identifies Prognostically

Differing Groups. *PLoS One, 4*(9), e7107. doi:https://doi.org/10.1371/journal.pone.0007107

Gao, M., Zhu, L., Chang, J., Cao, T., Song, L., Wen, C., . . . Chen, F. (2023). Safety and Efficacy of Edaravone in Patients with Amyotrophic Lateral Sclerosis: A Systematic Review and Meta-analysis. *Clinical Drug Investigation, 43*(1), 1-11. doi:https://doi.org/10.1007/s40261-022-01229-4

Genome Reference Consortium. Human Genome Region MHC. Retrieved from https://www.ncbi.nlm.nih.gov/grc/human/regions/MHC?asm=GRCh37

Genome UK: The future of healthcare. (2020). https://www.gov.uk/government/publications/genome-uk-the-future-of-healthcare

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genetics, 10*(5), e1004383. doi:https://doi.org/10.1371/journal.pgen.1004383

Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., . . . Roussos, P. (2018). A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics, 34*(15), 2538-2545. doi:https://doi.org/10.1093/bioinformatics/bty147

Giannoccaro, M. P., Bartoletti-Stella, A., Piras, S., Pession, A., De Massis, P., Oppi, F., . . . Capellari, S. (2017). Multiple variants in families with amyotrophic lateral sclerosis and frontotemporal dementia related to C9orf72 repeat expansion: further observations on their oligogenic nature. *Journal of Neurology, 264*(7), 1426-1433. doi:https://doi.org/10.1007/s00415-017-8540-x

Gidalevitz, T., Krupinski, T., Garcia, S., & Morimoto, R. I. (2009). Destabilizing Protein Polymorphisms in the Genetic Background Direct Phenotypic Expression of Mutant SOD1 Toxicity. *PLoS Genetics, 5*(3), e1000399. doi:https://doi.org/10.1371/journal.pgen.1000399

Gilbert, E. S. (1968). On Discrimination Using Qualitative Variables. *Journal of the American Statistical Association, 63*(324), 1399-1412. doi:https://doi.org/10.1080/01621459.1968.10480936

Goldwurm, S., Tunesi, S., Tesei, S., Zini, M., Sironi, F., Primignani, P., . . . Pezzoli, G. (2011). Kin-cohort analysis of *LRRK2*-G2019S penetrance in Parkinson's disease. *Movement disorders : official journal of the Movement Disorder Society, 26*(11), 2144-2145. doi:10.1002/mds.23807

Greenway, M. J., Andersen, P. M., Russ, C., Ennis, S., Cashman, S., Donaghy, C., . . . Hardiman, O. (2006). *ANG* mutations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. *Nature Genetics, 38*(4), 411-413. doi:https://doi.org/10.1038/ng1742

Grotzinger, A. D., Fuente, J. d. l., Privé, F., Nivard, M. G., & Tucker-Drob, E. M. (2023). Pervasive Downward Bias in Estimates of Liability-Scale Heritability in Genome-wide Association Study Meta-analysis: A Simple Solution. *Biological Psychiatry*. doi:https://doi.org/10.1016/j.biopsych.2022.05.029

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 621-638. doi:https://doi.org/10.1080/10705511.2017.1402334

Hardiman, O., Al-Chalabi, A., Chio, A., Corr, E. M., Logroscino, G., Robberecht, W., . . . van den Berg, L. H. (2017). Amyotrophic lateral sclerosis. *Nature Reviews Disease Primers, 3*(1), 17071. doi:https://doi.org/10.1038/nrdp.2017.71

Healy, D. G., Falchi, M., O'Sullivan, S. S., Bonifati, V., Durr, A., Bressman, S., . . . Wood, N. W. (2008). Phenotype, genotype, and worldwide genetic penetrance of *LRRK2*-associated Parkinson's disease: a case-control study. *Lancet Neurology, 7*(7), 583-590. doi:https://doi.org/10.1016/S1474-4422(08)70117-0

Hillert, A., Anikster, Y., Belanger-Quintana, A., Burlina, A., Burton, B. K., Carducci, C., . . . Blau, N. (2020). The Genetic Landscape and Epidemiology of Phenylketonuria. *American Journal of Human Genetics, 107*(2), 234-250. doi:https://doi.org/10.1016/j.ajhg.2020.06.006

Hop, P. J., Zwamborn, R. A. J., Hannon, E., Shireby, G. L., Nabais, M. F., Walker, E. M., . . . Veldink, J. H. (2022). Genome-wide study of DNA methylation shows alterations in metabolic, inflammatory, and cholesterol pathways in ALS. *Science Translational Medicine, 14*(633), eabj0264. doi:https://doi.org/10.1126/scitranslmed.abj0264

Hopperton, K. E., Mohammad, D., Trépanier, M. O., Giuliano, V., & Bazinet, R. P. (2018). Markers of microglia in post-mortem brain samples from patients with Alzheimer's disease: a systematic review. *Molecular Psychiatry, 23*(2), 177-198. doi:https://doi.org/10.1038/mp.2017.246

Hörster, F., Kölker, S., Loeber, J. G., Cornel, M. C., Hoffmann, G. F., & Burgard, P. (2017). Newborn Screening Programmes in Europe, Arguments and Efforts Regarding Harmonisation: Focus on Organic Acidurias. *JIMD reports, 32*, 105-115. doi:https://doi.org/10.1007/8904_2016_537

Howard, H. C., Knoppers, B. M., Cornel, M. C., Wright Clayton, E., Sénécal, K., Borry, P., . . . the, P. H. G. F. (2015). Whole-genome sequencing in newborn screening? A statement on the continued importance of targeted approaches in newborn screening programmes. *European Journal of Human Genetics, 23*(12), 1593-1600. doi:https://doi.org/10.1038/ejhg.2014.289

Hughes, I., & Hase, T. (2010). *Measurements and their uncertainties: a practical guide to modern error analysis*. Oxford: Oxford University Press.

Humphrey, J., Venkatesh, S., Hasan, R., Herb, J. T., de Paiva Lopes, K., Küçükali, F., . . . Consortium, N. A. (2023). Integrative transcriptomic analysis of the amyotrophic lateral sclerosis spinal cord implicates glial activation and suggests new risk genes. *Nature Neuroscience, 26*(1), 150-162. doi:https://doi.org/10.1038/s41593-022-01205-3

Hunink, M. G. M., Weinstein, M. C., Wittenberg, E., Drummond, M. F., Pliskin, J. S., Wong, J. B., & Glasziou, P. P. (2014). *Decision Making in Health and Medicine: Integrating Evidence and Values* (2nd ed.). Cambridge: Cambridge University Press.

Hunter, J. E., Irving, S. A., Biesecker, L. G., Buchanan, A., Jensen, B., Lee, K., . . . on behalf of the ClinGen, R. (2016). A standardized, evidence-based protocol to assess clinical actionability of genetic disorders associated with genomic variation. *Genetics in Medicine, 18*(12), 1258-1268. doi:https://doi.org/10.1038/gim.2016.40

Iacoangeli, A., Al Khleifat, A., Jones, A. R., Sproviero, W., Shatunov, A., Opie-Martin, S., . . . Al-Chalabi, A. (2019). *C9orf72* intermediate expansions of 24–30 repeats are associated with ALS. *Acta Neuropathol Commun, 7*, 115. doi:https://doi.org/10.1186/s40478-019-0724-4

Iacoangeli, A., Al Khleifat, A., Sproviero, W., Shatunov, A., Jones, A. R., Morgan, S. L., . . . Al-Chalabi, A. (2019). DNAscan: personal computer compatible NGS analysis, annotation and visualisation. *BMC Bioinformatics, 20*(1), 213. doi:https://doi.org/10.1186/s12859-019-2791-8

Iacoangeli, A., Al Khleifat, A., Sproviero, W., Shatunov, A., Jones, A. R., Opie-Martin, S., . . . Al-Chalabi, A. (2019). ALSgeneScanner: a pipeline for the analysis and interpretation of DNA sequencing data of ALS patients. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 20*, 207-215. doi:https://doi.org/10.1080/21678421.2018.1562553

Iacoangeli, A., Fogh, I., Selvackadunco, S., Topp, S. D., Shatunov, A., van Rheenen, W., . . . Jones, A. R. (2021). SCFD1 expression Quantitative Trait Loci in Amyotrophic Lateral Sclerosis are differentially expressed. *Brain Communications*. doi:https://doi.org/10.1093/braincomms/fcab236

Illumina. (2019). Accuracy Improvements in Germline Small Variant Calling with the DRAGEN Platform.

IUPAC-IUB Joint Commission on Biochemical Nomenclature. (1984). Nomenclature and Symbolism for Amino Acids and Peptides: Recommendations 1983. *European Journal of Biochemistry, 138*(1), 9-37. doi:https://doi.org/10.1111/j.1432-1033.1984.tb07877.x

Jansen, M. E., Lister, K. J., van Kranen, H. J., & Cornel, M. C. (2017). Policy Making in Newborn Screening Needs a Structured and Transparent Approach. *Frontiers in Public Health, 5*(53). doi:https://doi.org/10.3389/fpubh.2017.00053

Johnson, J. O., Mandrioli, J., Benatar, M., Abramzon, Y., Van Deerlin, V. M., Trojanowski, J. Q., . . . Traynor, B. J. (2010). Exome Sequencing Reveals VCP Mutations as a Cause of Familial ALS. *Neuron, 68*(5), 857-864. doi:https://doi.org/10.1016/j.neuron.2010.11.036

Johnston, C. A., Stanton, B. R., Turner, M. R., Gray, R., Blunt, A. H.-M., Butt, D., . . . Al-Chalabi, A. (2006). Amyotrophic lateral sclerosis in an urban setting: A population based study of inner city London. *Journal of Neurology, 253*(12), 1642-1643. doi:https://doi.org/10.1007/s00415-006-0195-y

Jones, A. R., Iacoangeli, A., Adey, B. N., Bowles, H., Shatunov, A., Troakes, C., . . . Al-Chalabi, A. (2021). A HML6 endogenous retrovirus on chromosome 3 is upregulated in amyotrophic lateral sclerosis motor cortex. *Scientific Reports, 11*(1), 14283. doi:https://doi.org/10.1038/s41598-021-93742-3

Jones, A. R., Troakes, C., King, A., Sahni, V., De Jong, S., Bossers, K., . . . Al-Chalabi, A. (2015). Stratified gene expression analysis identifies major amyotrophic lateral sclerosis genes. *Neurobiology of Aging, 36*(5), 2006.e2001-2006.e2009. doi:https://doi.org/10.1016/j.neurobiolaging.2015.02.017

Juneja, T., Pericak-Vance, M. A., Laing, N. G., Dave, S., & Siddique, T. (1997a). Prognosis in Familial Amyotrophic Lateral Sclerosis. *Neurology, 48*(1), 55. doi:10.1212/WNL.48.1.55

Juneja, T., Pericak-Vance, M. A., Laing, N. G., Dave, S., & Siddique, T. (1997b). Prognosis in Familial Amyotrophic Lateral Sclerosis: Progression and survival in patients with glu100gly and ala4val mutations in Cu,Zn superoxide dismutase. *Neurology, 48*(1), 55. doi:https://doi.org/10.1212/WNL.48.1.55

Kalia, M., Miotto, M., Ness, D., Opie-Martin, S., Spargo, T. P., Di Rienzo, L., . . . Iacoangeli, A. (2022). Molecular dynamics analysis of Superoxide Dismutase 1 mutations suggests

decoupling between mechanisms underlying ALS onset and progression. *bioRxiv*, 2022.2012.2005.519128. doi:https://doi.org/10.1101/2022.12.05.519128

Kalia, S. S., Adelman, K., Bale, S. J., Chung, W. K., Eng, C., Evans, J. P., . . . on behalf of the, A. S. F. M. W. G. (2017). 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. *Genetics in Medicine, 19*(2), 249-255. doi:https://doi.org/10.1038/gim.2016.190

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., . . . Genome Aggregation Database, C. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature, 581*(7809), 434-443. doi:https://doi.org/10.1038/s41586-020-2308-7

Kassambara, A., Kosinski, M., & Biecek, P. (2021). survminer: Drawing Survival Curves using 'ggplot2' (Version 0.4.9). Retrieved from https://CRAN.R-project.org/package=survminer

Kay, C., Collins, J. A., Miedzybrodzka, Z., Madore, S. J., Gordon, E. S., Gerry, N., . . . Hayden, M. R. (2016). Huntington disease reduced penetrance alleles occur at high frequency in the general population. *Neurology, 87*(3), 282-288. doi:https://doi.org/10.1212/WNL.0000000000002858

Keller, M. F., Ferrucci, L., Singleton, A. B., Tienari, P. J., Laaksovirta, H., Restagno, G., . . . Nalls, M. A. (2014). Genome-Wide Analysis of the Heritability of Amyotrophic Lateral Sclerosis. *JAMA Neurology, 71*(9), 1123-1134. doi:https://doi.org/10.1001/jamaneurol.2014.1184

Kenna, K. P., McLaughlin, R. L., Byrne, S., Elamin, M., Heverin, M., Kenny, E. M., . . . Hardiman, O. (2013). Delineating the genetic heterogeneity of ALS using targeted high-throughput sequencing. *Journal of Medical Genetics, 50*(11), 776. doi:https://doi.org/10.1136/jmedgenet-2013-101795

Kenna, K. P., Van Doormaal, P. T. C., Dekker, A. M., Ticozzi, N., Kenna, B. J., Diekstra, F. P., . . . Landers, J. E. (2016). *NEK1* variants confer susceptibility to amyotrophic lateral sclerosis. *Nature Genetics, 48*(9), 1037-1042. doi:https://doi.org/10.1038/ng.3626

Kim, H. J., Kim, N. C., Wang, Y.-D., Scarborough, E. A., Moore, J., Diaz, Z., . . . Taylor, J. P. (2013). Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature, 495*(7442), 467-473. doi:https://doi.org/10.1038/nature11922

Kinney, J. W., Bemiller, S. M., Murtishaw, A. S., Leisgang, A. M., Salazar, A. M., & Lamb, B. T. (2018). Inflammation as a central mechanism in Alzheimer's disease. *Alzheimer's & Dementia: Translational Research & Clinical Interventions, 4*, 575-590. doi:https://doi.org/10.1016/j.trci.2018.06.014

Kirby, J., Goodall, E. F., Smith, W., Highley, J. R., Masanzu, R., Hartley, J. A., . . . Shaw, P. J. (2010). Broad clinical phenotypes associated with TAR-DNA binding protein (TARDBP) mutations in amyotrophic lateral sclerosis. *Neurogenetics, 11*(2), 217-225. doi:https://doi.org/10.1007/s10048-009-0218-9

Kirmeyer, S. E., & Hamilton, B. E. (2011). *Childbearing Differences Among Three Generations of U.S. Women*. Retrieved from https://www.cdc.gov/nchs/data/databriefs/db68.pdf

Kirov, G., Rees, E., Walters, J. T. R., Escott-Price, V., Georgieva, L., Richards, A. L., . . . Owen, M. J. (2014). The Penetrance of Copy Number Variations for Schizophrenia and Developmental Delay. *Biological Psychiatry, 75*(5), 378-385. doi:https://doi.org/10.1016/j.biopsych.2013.07.022

Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J., & Peterson, H. (2020). gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler [version 2; peer review: 2 approved]. *F1000Research, 9*(709). doi:https://doi.org/10.12688/f1000research.24956.2

Kondori, N. R., Paul, P., Robbins, J. P., Liu, K., Hildyard, J. C. W., Wells, D. J., & de Belleroche, J. S. (2018). Focus on the Role of D-serine and D-amino Acid Oxidase in Amyotrophic Lateral Sclerosis/Motor Neuron Disease (ALS). *Frontiers in Molecular Biosciences, 5*. doi:https://doi.org/10.3389/fmolb.2018.00008

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology, 20*(1), 117. doi:https://doi.org/10.1186/s13059-019-1720-5

Kuhn, M. (2022). caret: Classification and Regression Training (Version 6.0.93). Retrieved from https://CRAN.R-project.org/package=caret

Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., . . . Lieberman, A. P. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nature Genetics, 51*(3), 414-430. doi:https://doi.org/10.1038/s41588-019-0358-2

Labra, J., Menon, P., Byth, K., Morrison, S., & Vucic, S. (2016). Rate of disease progression: a prognostic biomarker in ALS. *Journal of Neurology, Neurosurgery & Psychiatry, 87*(6), 628. doi:https://doi.org/10.1136/jnnp-2015-310998

Lacomblez, L., Bensimon, G., Leigh, P. N., Guillet, P., & Meininger, V. (1996). Dose-ranging study of riluzole in amyotrophic lateral sclerosis. Amyotrophic Lateral Sclerosis/Riluzole Study Group II. *Lancet, 347*(9013), 1425-1431. doi:https://doi.org/10.1016/s0140-6736(96)91680-3

Lambrechts, D., Storkebaum, E., Morimoto, M., Del-Favero, J., Desmet, F., Marklund, S. L., . . . Carmeliet, P. (2003). VEGF is a modifier of amyotrophic lateral sclerosis in mice and humans and protects motoneurons against ischemic death. *Nature Genetics, 34*(4), 383-394. doi:https://doi.org/10.1038/ng1211

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., . . . Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research, 46*(D1), D1062-D1067. doi:https://doi.org/10.1093/nar/gkx1153

Langbehn, D. R., Brinkman, R. R., Falush, D., Paulsen, J. S., & Hayden, M. R. (2004). A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clinical Genetics, 65*(4), 267-277. doi:https://doi.org/10.1111/j.1399-0004.2004.00241.x

Larkin, E. K., Newman, J. H., Austin, E. D., Hemnes, A. R., Wheeler, L., Robbins, I. M., . . . Loyd, J. E. (2012). Longitudinal Analysis Casts Doubt on the Presence of Genetic Anticipation in Heritable Pulmonary Arterial Hypertension. *American Journal of Respiratory and Critical Care Medicine, 186*(9), 892-896. doi:https://doi.org/10.1164/rccm.201205-0886OC

Lattante, S., Ciura, S., Rouleau, G. A., & Kabashi, E. (2015). Defining the genetic connection linking amyotrophic lateral sclerosis (ALS) with frontotemporal dementia (FTD). *Trends in Genetics, 31*(5), 263-273. doi:https://doi.org/10.1016/j.tig.2015.03.005

Lattante, S., Pomponi, M. G., Conte, A., Marangi, G., Bisogni, G., Patanella, A. K., . . . Sabatelli, M. (2018). ATXN1 intermediate-length polyglutamine expansions are

associated with amyotrophic lateral sclerosis. *Neurobiology of Aging, 64*, 157.e151-157.e155. doi:https://doi.org/10.1016/j.neurobiolaging.2017.11.011

Lattante, S., Rouleau, G. A., & Kabashi, E. (2013). *TARDBP* and *FUS* Mutations Associated with Amyotrophic Lateral Sclerosis: Summary and Update. *Human Mutation, 34*(6), 812-826. doi:https://doi.org/10.1002/humu.22319

Lee, A. J., Wang, Y., Alcalay, R. N., Mejia-Santana, H., Saunders-Pullman, R., Bressman, S., . . . for the Michael, J. F. L. C. C. (2017). Penetrance estimate of LRRK2 p.G2019S mutation in individuals of non-Ashkenazi Jewish ancestry. *Movement Disorders, 32*(10), 1432-1438. doi:https://doi.org/10.1002/mds.27059

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., . . . Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics, 50*(8), 1112-1121. doi:https://doi.org/10.1038/s41588-018-0147-3

Lee, T., & Shi, D. (2021). A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. *Psychological Methods, 26*(4), 466-485. doi:https://doi.org/10.1037/met0000381

Li, C., Yang, T., Ou, R., & Shang, H. (2021). Overlapping Genetic Architecture Between Schizophrenia and Neurodegenerative Disorders. *Frontiers in Cell and Developmental Biology, 9*. doi:https://doi.org/10.3389/fcell.2021.797072

Li, Z., Wang, Y., & Wang, F. (2018). A study on fast calling variants from next-generation sequencing data using decision tree. *BMC Bioinformatics, 19*(1), 145. doi:https://doi.org/10.1186/s12859-018-2147-9

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18-22.

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., . . . Visscher, P. M. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nature Communications, 10*(1), 5086. doi:https://doi.org/10.1038/s41467-019-12653-0

Loeber, J. G., Burgard, P., Cornel, M. C., Rigter, T., Weinreich, S. S., Rupp, K., . . . Vittozzi, L. (2012). Newborn screening programmes in Europe; arguments and efforts regarding harmonization. Part 1 – From blood spot to screening result. *Journal of Inherited Metabolic Disease, 35*(4), 603-611. doi:https://doi.org/10.1007/s10545-012-9483-0

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550. doi:https://doi.org/10.1186/s13059-014-0550-8

Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions.* Paper presented at the Advances in Neural Information Processing Systems.

Mackenzie, I. R., Bigio, E. H., Ince, P. G., Geser, F., Neumann, M., Cairns, N. J., . . . Trojanowski, J. Q. (2007). Pathological TDP-43 distinguishes sporadic amyotrophic lateral sclerosis from amyotrophic lateral sclerosis with SOD1 mutations. *Annals of Neurology, 61*(5), 427-434. doi:https://doi.org/10.1002/ana.21147

Majounie, E., Renton, A. E., Mok, K., Dopper, E. G. P., Waite, A., Rollinson, S., . . . Traynor, B. J. (2012). Frequency of the *C9orf72* hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. *Lancet Neurology, 11*(4), 323-330. doi:https://doi.org/10.1016/s1474-4422(12)70043-1

Majumder, M. A., Guerrini, C. J., & McGuire, A. L. (2021). Direct-to-Consumer Genetic Testing: Value and Risk. *Annual Review of Medicine, 72*(1), 151-166. doi:https://doi.org/10.1146/annurev-med-070119-114727

Marogianni, C., Rikos, D., Provatas, A., Dadouli, K., Ntellas, P., Tsitsi, P., . . . Xiromerisiou, G. (2019). The role of C9orf72 in neurodegenerative disorders: a systematic review, an updated meta-analysis, and the creation of an online database. *Neurobiology of Aging*, 1.e1-1.e10. doi:https://doi.org/10.1016/j.neurobiolaging.2019.04.012

Marriott, H., Spargo, T. P., Khleifat, A. A., Fogh, I., Andersen, P. M., Başak, N. A., . . . Iacoangeli, A. (2022). Mutations in the tail domain of the neurofilament heavy chain gene increase the risk of amyotrophic lateral sclerosis. *medRxiv*, 2022.2011.2003.22281905. doi:https://doi.org/10.1101/2022.11.03.22281905

Masrori, P., & Van Damme, P. (2020). Amyotrophic lateral sclerosis: a clinical review. *European Journal of Neurology, 27*(10), 1918-1929. doi:https://doi.org/10.1111/ene.14393

Mayer, M. (2023). kernelshap: Kernel SHAP (Version R package version 0.3.3). Retrieved from https://github.com/mayer79/kernelshap

McCann, E. P., Henden, L., Fifita, J. A., Zhang, K. Y., Grima, N., Bauer, D. C., . . . Blair, I. P. (2021). Evidence for polygenic and oligogenic basis of Australian sporadic amyotrophic lateral sclerosis. *Journal of Medical Genetics, 58*(2), 87-95. doi:https://doi.org/10.1136/jmedgenet-2020-106866

McCann, E. P., Williams, K. L., Fifita, J. A., Tarr, I. S., O'Connor, J., Rowe, D. B., . . . Blair, I. P. (2017). The genotype–phenotype landscape of familial amyotrophic lateral sclerosis in Australia. *Clinical Genetics, 92*(3), 259-266. doi:https://doi.org/10.1111/cge.12973

McCauley, M. E., & Baloh, R. H. (2019). Inflammation in ALS/FTD pathogenesis. *Acta Neuropathologica, 137*(5), 715-730. doi:https://doi.org/10.1007/s00401-018-1933-9

McCusker, E. A., & Loy, C. T. (2017). Huntington Disease: The Complexities of Making and Disclosing a Clinical Diagnosis After Premanifest Genetic Testing. *Tremor and Other Hyperkinetic Movements, 7*, 467. doi:https://doi.org/10.7916/D8PK0TDD

McKenzie, A. T., Wang, M., Hauberg, M. E., Fullard, J. F., Kozlenkov, A., Keenan, A., . . . Zhang, B. (2018). Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Scientific Reports, 8*(1), 8868. doi:https://doi.org/10.1038/s41598-018-27293-5

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., . . . Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology, 17*(1), 122. doi:https://doi.org/10.1186/s13059-016-0974-4

McLaughlin, R. L., Schijven, D., van Rheenen, W., van Eijk, K. R., O'Brien, M., Kahn, R. S., . . . Schizophrenia Working Group of the Psychiatric Genomics, C. (2017). Genetic correlation between amyotrophic lateral sclerosis and schizophrenia. *Nature Communications, 8*(1), 14774. doi:https://doi.org/10.1038/ncomms14774

McLaughlin, R. L., Vajda, A., & Hardiman, O. (2015). Heritability of Amyotrophic Lateral Sclerosis: Insights From Disparate Numbers. *JAMA Neurology, 72*(8), 857-858. doi:https://doi.org/10.1001/jamaneurol.2014.4049

Mehta, P., Kaye, W., Raymond, J., Punjani, R., Larson, T., Cohen, J., . . . Horton, K. (2018). Prevalence of Amyotrophic Lateral Sclerosis — United States, 2015. *Morbidity and Mortality Weekly Report, 67*(46). doi:https://doi.org/10.15585/mmwr.mm6746a1

Mehta, P. R., Iacoangeli, A., Opie-Martin, S., van Vugt, J., Al Khleifat, A., Bredin, A., . . . Al-Chalabi, A. (2022). The impact of age on genetic testing decisions in amyotrophic

lateral sclerosis. *Brain, 145*(12), 4440-4447. doi:https://doi.org/10.1093/brain/awac279

Miller, T. M., Cudkowicz, M. E., Genge, A., Shaw, P. J., Sobue, G., Bucelli, R. C., . . . Fradette, S. (2022). Trial of Antisense Oligonucleotide Tofersen for *SOD1* ALS. *New England Journal of Medicine, 387*(12), 1099-1110. doi:https://doi.org/10.1056/NEJMoa2204705

Minikel, E. V., Vallabh, S. M., Lek, M., Estrada, K., Samocha, K. E., Sathirapongsasuti, J. F., . . . MacArthur, D. G. (2016). Quantifying prion disease penetrance using large population control cohorts. *Science Translational Medicine, 8*(322), 322ra329. doi:https://doi.org/10.1126/scitranslmed.aad5169

Mokhtari, R., & Lachman, H. M. (2016). The Major Histocompatibility Complex (MHC) in Schizophrenia: A Review. *Journal of Clinical & Cellular Immunology, 7*(6). doi:https://doi.org/10.4172/2155-9899.1000479

Montuschi, A., Iazzolino, B., Calvo, A., Moglia, C., Lopiano, L., Restagno, G., . . . Chiò, A. (2015). Cognitive correlates in amyotrophic lateral sclerosis: a population-based study in Italy. *Journal of Neurology, Neurosurgery and Psychiatry, 86*(2), 168. doi:https://doi.org/10.1136/jnnp-2013-307223

Moore, D. H. (1973). Evaluation of Five Discrimination Procedures for Binary Variables. *Journal of the American Statistical Association, 68*(342), 399-404. doi:https://doi.org/10.1080/01621459.1973.10482440

Moorthie, S., Hall, A., Janus, J., Brigden, T., Villiers, C. B. d., Blackburn, L., . . . Kroese, M. (2021). *Polygenic scores and clinical utility*. Retrieved from https://www.phgfoundation.org/report/polygenic-scores-and-clinical-utility

Morello, G., Spampinato, A. G., Conforti, F. L., D'Agata, V., & Cavallaro, S. (2017). Selection and Prioritization of Candidate Drug Targets for Amyotrophic Lateral Sclerosis Through a Meta-Analysis Approach. *Journal of Molecular Neuroscience, 61*(4), 563-580. doi:https://doi.org/10.1007/s12031-017-0898-9

Morgan, S., & Orrell, R. W. (2016). Pathogenesis of amyotrophic lateral sclerosis. *British Medical Bulletin, 119*(1), 87-98. doi:https://doi.org/10.1093/bmb/ldw026

Morgan, S., Shatunov, A., Sproviero, W., Jones, A. R., Shoai, M., Hughes, D., . . . Al-Chalabi, A. (2017). A comprehensive analysis of rare genetic variation in amyotrophic lateral sclerosis in the UK. *Brain, 140*(6), 1611-1618. doi:https://doi.org/10.1093/brain/awx082

Morita, M., Al-Chalabi, A., Andersen, P. M., Hosler, B., Sapp, P., Englund, E., . . . Brown, R. H. (2006). A locus on chromosome 9p confers susceptibility to ALS and frontotemporal dementia. *Neurology, 66*(6), 839. doi:https://doi.org/10.1212/01.wnl.0000200048.53766.b4

Müller, K., Brenner, D., Weydt, P., Meyer, T., Grehl, T., Petri, S., . . . Weishaupt, J. H. (2018). Comprehensive analysis of the mutation spectrum in 301 German ALS families. *Journal of Neurology, Neurosurgery & Psychiatry, 89*(8), 817-827. doi:https://doi.org/10.1136/jnnp-2017-317611

Murphy, J., Factor-Litvak, P., Goetz, R., Lomen-Hoerth, C., Nagy, P. L., Hupf, J., . . . Als, C. (2016). Cognitive-behavioral screening reveals prevalent impairment in a large multicenter ALS cohort. *Neurology, 86*(9), 813-820. doi:https://doi.org/10.1212/WNL.0000000000002305

Murphy, N. A., Arthur, K. C., Tienari, P. J., Houlden, H., Chio, A., & Traynor, B. J. (2017). Age-related penetrance of the *C9orf72* repeat expansion. *Scientific Reports, 7*(1), 2116. doi:https://doi.org/10.1038/s41598-017-02364-1

Murray, M. F., Evans, J. P., Angrist, M., Chan, K., Uhlmann, W., Doyle, D. L., . . . Imhof, S. (2018). A proposed approach for implementing genomics-based screening programs for healthy adults. *NAM Perspectives*. doi:https://doi.org/10.31478/201812a

Murray, M. F., Evans, J. P., & Khoury, M. J. (2019). DNA-Based Population Screening: Potential Suitability and Important Knowledge Gaps. *JAMA*. doi:https://doi.org/10.1001/jama.2019.18640

Nakayama, S., Shimonaka, S., Elahi, M., Nishioka, K., Oji, Y., Matsumoto, S. E., . . . Hattori, N. (2019). Tau aggregation and seeding analyses of two novel MAPT variants found in patients with motor neuron disease and progressive parkinsonism. *Neurobiology of Aging*. doi:https://doi.org/10.1016/j.neurobiolaging.2019.02.016

Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., . . . Zhang, F. (2019). Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurology, 18*(12), 1091-1102. doi:https://doi.org/10.1016/S1474-4422(19)30320-5

Nel, M., Mahungu, A. C., Monnakgotla, N., Botha, G. R., Mulder, N. J., Wu, G., . . . Heckmann, J. M. (2022). Revealing the Mutational Spectrum in Southern Africans With Amyotrophic Lateral Sclerosis. *Neurology Genetics, 8*(1), e654. doi:https://doi.org/10.1212/NXG.0000000000000654

Nguyen, H. P., Van Broeckhoven, C., & van der Zee, J. (2018). ALS Genes in the Genomic Era and their Implications for FTD. *Trends in Genetics, 34*(6), 404-423. doi:https://doi.org/10.1016/j.tig.2018.03.001

Nishimura, A. L., Mitne-Neto, M., Silva, H. C. A., Richieri-Costa, A., Middleton, S., Cascio, D., . . . Zatz, M. (2004). A Mutation in the Vesicle-Trafficking Protein VAPB Causes Late-Onset Spinal Muscular Atrophy and Amyotrophic Lateral Sclerosis. *American Journal of Human Genetics, 75*(5), 822-831. doi:https://doi.org/10.1086/425287

Office for National Statistics. (2020). *Childbearing for women born in different years*. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/conceptionandfertilityrates/datasets/childbearingforwomenbornindifferentyearsreferencetable

Opie-Martin, S., Iacoangeli, A., Topp, S. D., Abel, O., Mayl, K., Mehta, P. R., . . . Shaw, C. E. (2022). The *SOD1*-mediated ALS phenotype shows a decoupling between age of symptom onset and disease duration. *Nature Communications, 13*(1), 6901. doi:https://doi.org/10.1038/s41467-022-34620-y

Origone, P., Geroldi, A., Lamp, M., Sanguineri, F., Caponnetto, C., Cabona, C., . . . Mandich, P. (2018). Role of *MAPT* in Pure Motor Neuron Disease: Report of a Recurrent Mutation in Italian Patients. *Neurodegenerative Diseases, 18*(5-6), 310-314. doi:https://doi.org/10.1159/000497820

Orlacchio, A., Babalini, C., Borreca, A., Patrono, C., Massa, R., Basaran, S., . . . Kawarai, T. (2010). SPATACSIN mutations cause autosomal recessive juvenile amyotrophic lateral sclerosis. *Brain, 133*(2), 591-598. doi:https://doi.org/10.1093/brain/awp325

Otto, P. A., & Horimoto, A. R. V. R. (2012). Penetrance rate estimation in autosomal dominant conditions. *Genetics and Molecular Biology, 35*(3), 583-588. doi:https://doi.org/10.1590/S1415-47572012005000051

Pain, O., Gillett, A. C., Austin, J. C., Folkersen, L., & Lewis, C. M. (2022). A tool for translating polygenic scores onto the absolute scale using summary statistics. *European Journal of Human Genetics, 30*, 339-348. doi:https://doi.org/10.1038/s41431-021-01028-z

Pain, O., Glanville, K. P., Hagenaars, S. P., Selzam, S., Fürtjes, A. E., Gaspar, H. A., . . . Lewis, C. M. (2021). Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genetics, 17*(5), e1009021. doi:https://doi.org/10.1371/journal.pgen.1009021

Pain, O., Jones, A., Khleifat, A. A., Agarwal, D., Hramyka, D., Karoui, H., . . . Al-Chalabi, A. (2023). Harnessing Transcriptomic Signals for Amyotrophic Lateral Sclerosis to Identify Novel Drugs and Enhance Risk Prediction. *medRxiv*, 2023.2001.2018.23284589. doi:https://doi.org/10.1101/2023.01.18.23284589

Pang, S. Y.-Y., Hsu, J. S., Teo, K.-C., Li, Y., Kung, M. H. W., Cheah, K. S. E., . . . Ho, S.-L. (2017). Burden of rare variants in ALS genes influences survival in familial and sporadic ALS. *Neurobiology of Aging, 58*, 238.e239-238.e215. doi:https://doi.org/10.1016/j.neurobiolaging.2017.06.007

Paquin, R. S., Mittendorf, K. F., Lewis, M. A., Hunter, J. E., Lee, K., Berg, J. S., . . . Goddard, K. A. B. (2019). Expert and lay perspectives on burden, risk, tolerability, and acceptability of clinical interventions for genetic disorders. *Genetics in Medicine, 21*(11), 2561-2568. doi:https://doi.org/10.1038/s41436-019-0524-z

Park, J. H., Elpers, C., Reunert, J., McCormick, M. L., Mohr, J., Biskup, S., . . . Marquardt, T. (2019). SOD1 deficiency: a novel syndrome distinct from amyotrophic lateral sclerosis. *Brain, 142*(8), 2230-2237. doi:https://doi.org/10.1093/brain/awz182

Parkinson's UK. (2017). *The Incidence and Prevalence of Parkinson's in the UK: Results from the Clinical Practice Research Datalink Reference Report*. Retrieved from https://www.parkinsons.org.uk/professionals/resources/incidence-and-prevalence-parkinsons-uk-report

Parton, M. J., Broom, W., Andersen, P. M., Al-Chalabi, A., Nigel Leigh, P., Powell, J. F., . . . Consortium, D. A. S. A. (2002). D90A-*SOD1* mediated amyotrophic lateral sclerosis: A single founder for all cases with evidence for a Cis-acting disease modifier in the recessive haplotype. *Human Mutation, 20*(6), 473. doi:https://doi.org/10.1002/humu.9081

Paulson, H. (2018). Repeat expansion diseases. *Handbook of Clinical Neurology, 147*, 105-123. doi:https://doi.org/10.1016/B978-0-444-63233-3.00009-9

Pensato, V., Magri, S., Bella, E. D., Tannorella, P., Bersano, E., Sorarù, G., . . . Gellera, C. (2020). Sorting Rare ALS Genetic Variants by Targeted Re-Sequencing Panel in Italian Patients: *OPTN*, *VCP*, and *SQSTM1* Variants Account for 3% of Rare Genetic Forms. *Journal of Clinical Medicine, 9*(2). doi:https://doi.org/10.3390/jcm9020412

Perrone, F., Cacace, R., Van Mossevelde, S., Van den Bossche, T., De Deyn, P. P., Cras, P., . . . Van Broeckhoven, C. (2018). Genetic screening in early-onset dementia patients with unclear phenotype: relevance for clinical diagnosis. *Neurobiology of Aging, 69*, 292.e297-292.e214. doi:https://doi.org/10.1016/j.neurobiolaging.2018.04.015

Phukan, J., Elamin, M., Bede, P., Jordan, N., Gallagher, L., Byrne, S., . . . Hardiman, O. (2012). The syndrome of cognitive impairment in amyotrophic lateral sclerosis: a population-based study. *Journal of Neurology, Neurosurgery and Psychiatry, 83*(1), 102. doi:https://doi.org/10.1136/jnnp-2011-300188

Placek, K., Baer, G. M., Elman, L., McCluskey, L., Hennessy, L., Ferraro, P. M., . . . McMillan, C. T. (2019). *UNC13A* polymorphism contributes to frontotemporal disease in

sporadic amyotrophic lateral sclerosis. *Neurobiology of Aging, 73*, 190-199. doi:https://doi.org/10.1016/j.neurobiolaging.2018.09.031

Privé, F., Aschard, H., Ziyatdinov, A., & Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics, 34*(16), 2781-2787. doi:https://doi.org/10.1093/bioinformatics/bty185

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery, 9*(3), e1301. doi:https://doi.org/10.1002/widm.1301

Project MinE ALS Sequencing Consortium. (2018). Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *European Journal of Human Genetics, 26*(10), 1537-1546. doi:https://doi.org/10.1038/s41431-018-0177-4

Purcell, S. PLINK (Version 1.9.0). Retrieved from http://pngu.mgh.harvard.edu/purcell/plink/

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., . . . Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics, 81*(3), 559-575. doi:https://doi.org/10.1086/519795

R Core Team. (2021). R: A language and environment for statistical computing: R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Ramamoorthy, D., Severson, K., Ghosh, S., Sachs, K., ALS, A., Glass, J. D., . . . Fraenkel, E. (2021). Identifying Patterns of ALS Progression from Sparse Longitudinal Data. *medRxiv*, 2021.2005.2013.21254848. doi:https://doi.org/10.1101/2021.05.13.21254848

Ranganathan, R., Haque, S., Coley, K., Shepheard, S., Cooper-Knock, J., & Kirby, J. (2020). Multifaceted Genes in Amyotrophic Lateral Sclerosis-Frontotemporal Dementia. *Frontiers in Neuroscience, 14*. doi:https://doi.org/10.3389/fnins.2020.00684

Ratnavalli, E., Brayne, C., Dawson, K., & Hodges, J. R. (2002). The prevalence of frontotemporal dementia. *Neurology, 58*(11), 1615. doi:https://doi.org/10.1212/WNL.58.11.1615

Ravnik-Glavač, M., Goričar, K., Vogrinc, D., Koritnik, B., Lavrenčič, J. G., Glavač, D., & Dolžan, V. (2022). Genetic Variability of Inflammation and Oxidative Stress Genes Affects Onset, Progression of the Disease and Survival of Patients with Amyotrophic Lateral Sclerosis. *Genes, 13*(5). doi:https://doi.org/10.3390/genes13050757

Reales, G., & Wallace, C. (2023). Sharing GWAS summary statistics results in more citations. *Communications Biology, 6*(1), 116. doi:https://doi.org/10.1038/s42003-023-04497-8

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., . . . Watson, M. S. (2015). ClinGen — The Clinical Genome Resource. *New England Journal of Medicine, 372*(23), 2235-2242. doi:https://doi.org/10.1056/NEJMsr1406261

Ren, X., & Kuan, P. F. (2020). RNAAgeCalc: A multi-tissue transcriptional age calculator. *PLoS One, 15*(8), e0237006. doi:https://doi.org/10.1371/journal.pone.0237006

Renton, A. E., Majounie, E., Waite, A., Simon-Sanchez, J., Rollinson, S., Gibbs, J. R., . . . Traynor, B. J. (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron, 72*(2), 257-268. doi:https://doi.org/10.1016/j.neuron.2011.09.010

Restuadi, R., Garton, F. C., Benyamin, B., Lin, T., Williams, K. L., Vinkhuyzen, A., . . . McRae, A. F. (2022). Polygenic risk score analysis for amyotrophic lateral sclerosis leveraging cognitive performance, educational attainment and schizophrenia. *European Journal of Human Genetics, 30*(5), 532-539. doi:https://doi.org/10.1038/s41431-021-00885-y

Revelle, W. (2022). psych: Procedures for Personality and Psychological Research (Version R package version 2.2.9.). Northwestern University, Evanston, Illinois, USA. Retrieved from https://CRAN.R-project.org/package=psych

Rhoades, R., Jackson, F., & Teng, S. (2019). Discovery of rare variants implicated in schizophrenia using next-generation sequencing. *Journal of Translational Genetics and Genomics, 3*, 1. doi:https://doi.org/10.20517/jtgg.2018.26

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., . . . on behalf of the, A. L. Q. A. C. (2015). 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. *Genetics in Medicine, 17*(5), 405-423. doi:https://doi.org/10.1038/gim.2015.30

Rinaldo, P., Zafari, S., Tortorelli, S., & Matern, D. (2006). Making the case for objective performance metrics in newborn screening by tandem mass spectrometry. *Mental Retardation and Developmental Disabilities Research Reviews, 12*(4), 255-261. doi:https://doi.org/10.1002/mrdd.20130

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*(1), 77. doi:https://doi.org/10.1186/1471-2105-12-77

Rosen, D. R., Bowling, A. C., Patterson, D., Usdin, T. B., Sapp, P., Mezey, E., . . . Brown, R. H., Jr. (1994). A frequent ala 4 to val superoxide dismutase-1 mutation is associated with a rapidly progressive familial amyotrophic lateral sclerosis. *Human Molecular Genetics, 3*(6), 981-987. doi:https://doi.org/10.1093/hmg/3.6.981

Rosen, D. R., Siddique, T., Patterson, D., Figlewicz, D. A., Sapp, P., Hentati, A., . . . Brown, R. H. (1993). Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature, 362*(6415), 59-62. doi:https://doi.org/10.1038/362059a0

Ross, L. F., & Clayton, E. W. (2019). Ethical Issues in Newborn Sequencing Research: The Case Study of BabySeq. *Pediatrics, 144*(6), e20191031. doi:https://doi.org/10.1542/peds.2019-1031

Rowland, L. P. (2001). How Amyotrophic Lateral Sclerosis Got Its Name: The Clinical-Pathologic Genius of Jean-Martin Charcot. *Archives of Neurology, 58*(3), 512-515. doi:https://doi.org/10.1001/archneur.58.3.512

Rutkove, S. B. (2015). Clinical Measures of Disease Progression in Amyotrophic Lateral Sclerosis. *Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics, 12*(2), 384-393. doi:https://doi.org/10.1007/s13311-014-0331-9

Ryan, M., Heverin, M., Doherty, M. A., Davis, N., Corr, E. M., Vajda, A., . . . Hardiman, O. (2018). Determining the incidence of familiality in ALS. *Neurology Genetics, 4*(3), e239. doi:https://doi.org/10.1212/NXG.0000000000000239

Ryan, M., Heverin, M., McLaughlin, R. L., & Hardiman, O. (2019). Lifetime Risk and Heritability of Amyotrophic Lateral Sclerosis. *JAMA Neurology, 76*(11), 1367-1374. doi:https://doi.org/10.1001/jamaneurol.2019.2044

Saeed, M., Yang, Y., Deng, H. X., Hung, W. Y., Siddique, N., Dellefave, L., . . . Siddique, T. (2009). Age and founder effect of SOD1 A4V mutation causing ALS. *Neurology, 72*(19), 1634-1639. doi:https://doi.org/10.1212/01.wnl.0000343509.76828.2a

Saelaert, M., Mertes, H., Moerenhout, T., De Baere, E., & Devisch, I. (2019). Criteria for reporting incidental findings in clinical exome sequencing – a focus group study on professional practices and perspectives in Belgian genetic centres. *BMC Medical Genomics, 12*(1), 123. doi:https://doi.org/10.1186/s12920-019-0561-0

Saha, S., Chant, D., Welham, J., & McGrath, J. (2005). A Systematic Review of the Prevalence of Schizophrenia. *PLoS Medicine, 2*(5), e141. doi:https://doi.org/10.1371/journal.pmed.0020141

Sali, A., & Attali, D. (2020). shinycssloaders: Add Loading Animations to a 'shiny' Output While It's Recalculating (Version R package version 1.0.0.). Retrieved from https://CRAN.R-project.org/package=shinycssloaders

Schulze, A., Lindner, M., Kohlmüller, D., Olgemöller, K., Mayatepek, E., & Hoffmann, G. F. (2003). Expanded newborn screening for inborn errors of metabolism by electrospray ionization-tandem mass spectrometry: results, outcome, and implications. *Pediatrics, 111*(6), 1399-1406. doi:https://doi.org/10.1542/peds.111.6.1399

Senol-Cosar, O., Schmidt, R. J., Qian, E., Hoskinson, D., Mason-Suares, H., Funke, B., & Lebo, M. S. (2019). Considerations for clinical curation, classification, and reporting of low-penetrance and low effect size variants associated with disease risk. *Genetics in Medicine, 21*(12), 2765-2773. doi:https://doi.org/10.1038/s41436-019-0560-8

Sharer, J. D. (2014). Amino Acid Disorders. In M. J. Aminoff & R. B. Daroff (Eds.), *Encyclopedia of the Neurological Sciences (Second Edition)* (pp. 136-147). Oxford: Academic Press.

Shatunov, A., & Al-Chalabi, A. (2021). The genetic architecture of ALS. *Neurobiology of Disease, 147*, 105156. doi:https://doi.org/10.1016/j.nbd.2020.105156

Shatunov, A., Mok, K., Newhouse, S., Weale, M. E., Smith, B., Vance, C., . . . Al-Chalabi, A. (2010). Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: a genome-wide association study. *Lancet Neurology, 9*(10), 986-994. doi:https://doi.org/10.1016/S1474-4422(10)70197-6

Shefner, J. M., Al-Chalabi, A., Baker, M. R., Cui, L.-Y., de Carvalho, M., Eisen, A., . . . Kiernan, M. C. (2020). A proposal for new diagnostic criteria for ALS. *Clinical Neurophysiology*. doi:https://doi.org/10.1016/j.clinph.2020.04.005

Shellikeri, S., Karthikeyan, V., Martino, R., Black, S. E., Zinman, L., Keith, J., & Yunusova, Y. (2017). The neuropathological signature of bulbar-onset ALS: A systematic review. *Neuroscience and Biobehavioral Reviews, 75*, 378-392. doi:https://doi.org/10.1016/j.neubiorev.2017.01.045

Sheppard, P., & Monden, C. (2020). When does family size matter? Sibship size, socioeconomic status and education in England. *Evolutionary Human Sciences, 2, e51*, 1-21. doi:https://doi.org/10.1017/ehs.2020.54

Shino, M. Y., McGuire, V., Van Den Eeden, S. K., Tanner, C. M., Popat, R., Leimpeter, A., . . . Nelson, L. M. (2010). Familial aggregation of Parkinson's disease in a multiethnic community-based case-control study. *Movement Disorders, 25*(15), 2587-2594. doi:https://doi.org/10.1002/mds.23361

Shireby, G. L., Davies, J. P., Francis, P. T., Burrage, J., Walker, E. M., Neilson, G. W. A., . . . Mill, J. (2020). Recalibrating the epigenetic clock: implications for assessing biological

age in the human cortex. *Brain, 143*(12), 3763-3775. doi:https://doi.org/10.1093/brain/awaa334

Shoesmith, C. L., Findlater, K., Rowe, A., & Strong, M. J. (2007). Prognosis of amyotrophic lateral sclerosis with respiratory onset. *Journal of Neurology, Neurosurgery, & Psychiatry, 78*(6), 629-631. doi:https://doi.org/10.1136/jnnp.2006.103564

Sievert, C. (2020). *Interactive web-based data visualization with R, plotly, and shiny*. Florida: CRC Press.

Simpson, C. L., & Al-Chalabi, A. (2006). Amyotrophic lateral sclerosis as a complex genetic disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 1762*(11), 973-985. doi:https://doi.org/10.1016/j.bbadis.2006.08.001

Sjoberg, D. (2022). ggsankey: Sankey, Alluvial and Sankey Bump Plots (Version 0.0.99999).

Smith, B. N., Topp, S. D., Fallini, C., Shibata, H., Chen, H. J., Troakes, C., . . . Shaw, C. E. (2017). Mutations in the vesicular trafficking protein annexin A11 are associated with amyotrophic lateral sclerosis. *Science Translational Medicine, 9*(388). doi:https://doi.org/10.1126/scitranslmed.aad9157

Snowden, J. S., Adams, J., Harris, J., Thompson, J. C., Rollinson, S., Richardson, A., . . . Pickering-Brown, S. (2015). Distinct clinical and pathological phenotypes in frontotemporal dementia associated with MAPT, PGRN and C9orf72 mutations. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 16*(7-8), 497-505. doi:https://doi.org/10.3109/21678421.2015.1074700

Solberg, O. K., Filkuková, P., Frich, J. C., & Feragen, K. J. B. (2018). Age at Death and Causes of Death in Patients with Huntington Disease in Norway in 1986-2015. *Journal of Huntington's disease, 7*(1), 77-86. doi:https://doi.org/10.3233/JHD-170270

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., . . . Harris, Laura W. (2022). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research, 51*(D1), D977-D985. doi:https://doi.org/10.1093/nar/gkac1010

Song, S., Miranda, C. J., Braun, L., Meyer, K., Frakes, A. E., Ferraiuolo, L., . . . Kaspar, B. K. (2016). Major histocompatibility complex class I molecules protect motor neurons from astrocyte-induced toxicity in amyotrophic lateral sclerosis. *Nature Medicine, 22*(4), 397-403. doi:https://doi.org/10.1038/nm.4052

Southern, K. W., Munck, A., Pollitt, R., Travert, G., Zanolla, L., Dankert-Roelse, J., & Castellani, C. (2007). A survey of newborn screening for cystic fibrosis in Europe. *Journal of Cystic Fibrosis, 6*(1), 57-65. doi:https://doi.org/10.1016/j.jcf.2006.05.008

Spargo, T. P., Opie-Martin, S., Bowles, H., Lewis, C. M., Iacoangeli, A., & Al-Chalabi, A. (2022). Calculating variant penetrance from family history of disease and average family size in population-scale data. *Genome Medicine, 14*, 141. doi:https://doi.org/10.1186/s13073-022-01142-7

Sproviero, W., Shatunov, A., Stahl, D., Shoai, M., van Rheenen, W., Jones, A. R., . . . Al-Chalabi, A. (2017). ATXN2 trinucleotide repeat length correlates with risk of ALS. *Neurobiology of Aging, 51*, 178.e171-178.e179. doi:https://doi.org/10.1016/j.neurobiolaging.2016.11.010

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7 ed.). Boston, MA: Pearson.

Takahashi, Y., Fukuda, Y., Yoshimura, J., Toyoda, A., Kurppa, K., Moritoyo, H., . . . Tsuji, S. (2013). ERBB4 Mutations that Disrupt the Neuregulin-ErbB4 Pathway Cause Amyotrophic Lateral Sclerosis Type 19. *American Journal of Human Genetics, 93*(5), 900-905. doi:https://doi.org/10.1016/j.ajhg.2013.09.008

Tam, O. H., Rozhkov, N. V., Shaw, R., Kim, D., Hubbard, I., Fennessey, S., . . . Gale Hammell, M. (2019). Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. *Cell Reports, 29*(5), 1164-1177.e1165. doi:https://doi.org/10.1016/j.celrep.2019.09.066

Tarini, B. A., Christakis, D. A., & Welch, H. G. (2006). State Newborn Screening in the Tandem Mass Spectrometry Era: More Tests, More False-Positive Results. *Pediatrics, 118*(2), 448. doi:https://doi.org/10.1542/peds.2005-2026

Tazelaar, G. H. P., Boeynaems, S., De Decker, M., van Vugt, J. J. F. A., Kool, L., Goedee, H. S., . . . van Es, M. A. (2020). *ATXN1* repeat expansions confer risk for amyotrophic lateral sclerosis and contribute to TDP-43 mislocalization. *Brain Communications*. doi:https://doi.org/10.1093/braincomms/fcaa064

Thakur, K., Tiwari, A., Sharma, K., Modgil, S., Khosla, R., & Anand, A. (2020). Angiogenesis-Centered Molecular Cross-Talk in Amyotrophic Lateral Sclerosis Survival: Mechanistic Insights. *Critical Reviews in Eukaryotic Gene Expression, 30*(2), 137-151. doi:https://doi.org/10.1615/CritRevEukaryotGeneExpr.2020031020

The U.S.–Venezuela Collaborative Research Project, & Wexler, N. S. (2004). Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proceedings of the National Academy of Sciences of the United States of America, 101*(10), 3498. doi:https://doi.org/10.1073/pnas.0308679101

The UniProt, C. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research, 49*(D1), D480-D489. doi:https://doi.org/10.1093/nar/gkaa1100

Thenappan, T., Ryan, J. J., & Archer, S. L. (2012). Evolving epidemiology of pulmonary arterial hypertension. *American Journal of Respiratory and Critical Care Medicine, 186*(8), 707-709. doi:https://doi.org/10.1164/rccm.201207-1266ED

Therneau, T. (2022). A Package for Survival Analysis in R (Version 3.3.1). Retrieved from https://CRAN.R-project.org/package=survival

Therneau, T. (2023). A Package for Survival Analysis in R (Version 3.5-5). Retrieved from https://CRAN.R-project.org/package=survival

Tierney, N., Cook, D., McBain, M., & Fay, C. (2021). naniar: Data Structures, Summaries, and Visualisations for Missing Data (Version 0.6.1). Retrieved from https://CRAN.R-project.org/package=naniar

Tompa, D. R., & Kadhirvel, S. (2020). Changes in hydrophobicity mainly promotes the aggregation tendency of ALS associated SOD1 mutants. *International Journal of Biological Macromolecules, 145*, 904-913. doi:https://doi.org/10.1016/j.ijbiomac.2019.09.181

Trowsdale, J., & Knight, J. C. (2013). Major Histocompatibility Complex Genomics and Human Disease. *Annual Review of Genomics and Human Genetics, 14*(1), 301-323. doi:https://doi.org/10.1146/annurev-genom-091212-153455

Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., . . . Bertolino, A. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature, 604*(7906), 502-508. doi:https://doi.org/10.1038/s41586-022-04434-5

Turner, M. R., Al-Chalabi, A., Chio, A., Hardiman, O., Kiernan, M. C., Rohrer, J. D., . . . Talbot, K. (2017). Genetic screening in sporadic ALS and FTD. *Journal of Neurology, Neurosurgery and Psychiatry, 88*(12), 1042-1044. doi:https://doi.org/10.1136/jnnp-2017-315995

Turner, M. R., Barnwell, J., Al-Chalabi, A., & Eisen, A. (2012). Young-onset amyotrophic lateral sclerosis: historical and other observations. *Brain, 135*(9), 2883-2891. doi:https://doi.org/10.1093/brain/aws144

Umoh, M. E., Fournier, C., Li, Y., Polak, M., Shaw, L., Landers, J. E., . . . Glass, J. D. (2016). Comparative analysis of C9orf72 and sporadic disease in an ALS clinic population. *Neurology, 87*(10), 1024-1030. doi:https://doi.org/10.1212/wnl.0000000000003067

Vajda, A., McLaughlin, R. L., Heverin, M., Thorpe, O., Abrahams, S., Al-Chalabi, A., & Hardiman, O. (2017). Genetic testing in ALS: A survey of current practices. *Neurology, 88*(10), 991-999. doi:https://doi.org/10.1212/wnl.0000000000003686

van Blitterswijk, M., Baker, M. C., DeJesus-Hernandez, M., Ghidoni, R., Benussi, L., Finger, E., . . . Rademakers, R. (2013). C9ORF72 repeat expansions in cases with previously identified pathogenic mutations. *Neurology, 81*(15), 1332-1341. doi:https://doi.org/10.1212/WNL.0b013e3182a8250c

van Blitterswijk, M., Mullen, B., Nicholson, A. M., Bieniek, K. F., Heckman, M. G., Baker, M. C., . . . Rademakers, R. (2014). TMEM106B protects C9ORF72 expansion carriers against frontotemporal dementia. *Acta Neuropathologica, 127*(3), 397-406. doi:https://doi.org/10.1007/s00401-013-1240-4

van Blitterswijk, M., van Es, M. A., Hennekam, E. A. M., Dooijes, D., van Rheenen, W., Medic, J., . . . van den Berg, L. H. (2012). Evidence for an oligogenic basis of amyotrophic lateral sclerosis. *Human Molecular Genetics, 21*(17), 3776-3784. doi:https://doi.org/10.1093/hmg/dds199

van der Spek, R. A. A., van Rheenen, W., Pulit, S. L., Kenna, K. P., van den Berg, L. H., & Veldink, J. H. (2019). The project MinE databrowser: bringing large-scale whole-genome sequencing in ALS to researchers and the public. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 20*, 432-440. doi:https://doi.org/10.1080/21678421.2019.1606244

van Eijk, R. P. A., Jones, A. R., Sproviero, W., Shatunov, A., Shaw, P. J., Leigh, P. N., . . . Van Es, M. A. (2017). Meta-analysis of pharmacogenetic interactions in amyotrophic lateral sclerosis clinical trials. *Neurology, 89*(18), 1915-1922. doi:https://doi.org/10.1212/wnl.0000000000004606

van Eijk, R. P. A., Kliest, T., McDermott, C. J., Roes, K. C. B., Van Damme, P., Chiò, A., . . . van den Berg, L. H. (2020). TRICALS: creating a highway toward a cure. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 1-6. doi:https://doi.org/10.1080/21678421.2020.1788092

van Es, M. A., Hardiman, O., Chiò, A., Al-Chalabi, A., Pasterkamp, R. J., Veldink, J. H., & van den Berg, L. H. (2017). Amyotrophic lateral sclerosis. *The Lancet, 390*(10107), 2084-2098. doi:https://doi.org/10.1016/s0140-6736(17)31287-4

van Es, M. A., Veldink, J. H., Saris, C. G., Blauw, H. M., van Vught, P. W., Birve, A., . . . van den Berg, L. H. (2009). Genome-wide association study identifies 19p13.3 (*UNC13A*) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. *Nature Genetics, 41*(10), 1083-1087. doi:https://doi.org/10.1038/ng.442

van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pulit, S. L., . . . Veldink, J. H. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature Genetics, 48*(9), 1043-1048. doi:https://doi.org/10.1038/ng.3622

van Rheenen, W., van der Spek, R. A. A., Bakker, M. K., van Vugt, J. J. F. A., Hop, P. J., Zwamborn, R. A. J., . . . Consortium, S. (2021). Common and rare variant association

analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nature Genetics, 53*(12), 1636-1648. doi:https://doi.org/10.1038/s41588-021-00973-1

Vance, C., Al-Chalabi, A., Ruddy, D., Smith, B. N., Hu, X., Sreedharan, J., . . . Shaw, C. E. (2006). Familial amyotrophic lateral sclerosis with frontotemporal dementia is linked to a locus on chromosome 9p13.2-21.3. *Brain, 129*(Pt 4), 868-876. doi:https://doi.org/10.1093/brain/awl030

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer.

Vucic, S., Higashihara, M., Sobue, G., Atsuta, N., Doi, Y., Kuwabara, S., . . . Kiernan, M. C. (2020). ALS is a multistep process in South Korean, Japanese, and Australian patients. *Neurology, 94*(15), e1657. doi:https://doi.org/10.1212/WNL.0000000000009015

Wadman, R. I., Jansen, M. D., Stam, M., Wijngaarde, C. A., Curial, C. A. D., Medic, J., . . . van der Pol, W. L. (2020). Intragenic and structural variation of the SMN locus and clinical variability of spinal muscular atrophy. *Brain Communications, 2*(2), fcaa075. doi:https://doi.org/10.1093/braincomms/fcaa075

Wainschtein, P., Jain, D., Zheng, Z., Aslibekyan, S., Becker, D., Bi, W., . . . Consortium, N. T.-O. f. P. M. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics, 54*(3), 263-273. doi:https://doi.org/10.1038/s41588-021-00997-7

Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genetics, 17*(9), e1009440. doi:https://doi.org/10.1371/journal.pgen.1009440

Wang, G., Sarkar, A., Carbonetto, P., & Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 82*(5), 1273-1300. doi:https://doi.org/10.1111/rssb.12388

Wang, Z.-X., Wan, Q., & Xing, A. (2020). HLA in Alzheimer's Disease: Genetic Association and Possible Pathogenic Roles. *Neuromolecular Medicine, 22*(4), 464-473. doi:https://doi.org/10.1007/s12017-020-08612-4

Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T. J. C., . . . Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics, 51*(9), 1339-1348. doi:https://doi.org/10.1038/s41588-019-0481-0

Weintraub, D., & Mamikonyan, E. (2019). The Neuropsychiatry of Parkinson Disease: A Perfect Storm. *American Journal of Geriatric Psychiatry, 27*(9), 998-1018. doi:https://doi.org/10.1016/j.jagp.2019.03.002

Weishaupt, J. H., Hyman, T., & Dikic, I. (2016). Common Molecular Pathways in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. *Trends in Molecular Medicine, 22*(9), 769-783. doi:https://doi.org/10.1016/j.molmed.2016.07.005

Weller, B. E., Bowen, N. K., & Faubert, S. J. (2020). Latent Class Analysis: A Guide to Best Practice. *Journal of Black Psychology, 46*(4), 287-311. doi:https://doi.org/10.1177/0095798420930932

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., . . . Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing

improves variant detection and assembly of a human genome. *Nature Biotechnology, 37*(10), 1155-1162. doi:https://doi.org/10.1038/s41587-019-0217-9

Werme, J., van der Sluis, S., Posthuma, D., & de Leeuw, C. A. (2022). An integrated framework for local genetic correlation analysis. *Nature Genetics, 54*(3), 274-282. doi:https://doi.org/10.1038/s41588-022-01017-y

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software, 21*(12), 1-20. doi:https://doi.org/10.18637/jss.v021.i12

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software, 40*(1), 1-29. doi:https://doi.org/10.18637/jss.v040.i01

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*: Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., . . . Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. doi:https://doi.org/10.21105/joss.01686

Willemse, S. W., Harley, P., van Eijk, R. P. A., Demaegd, K. C., Zelina, P., Pasterkamp, R. J., . . . van Es, M. A. (2023). *UNC13A* in amyotrophic lateral sclerosis: from genetic association to therapeutic target. *Journal of Neurology, Neurosurgery & Psychiatry*, jnnp-2022-330504. doi:https://doi.org/10.1136/jnnp-2022-330504

World Bank, W. D. I. (2020). Fertility rate, total (births per woman). Retrieved 12/01/2021 https://databank.worldbank.org/reports.aspx?source=2&series=SP.DYN.TFRT.IN

Wright, A. L., Della Gatta, P. A., Le, S., Berning, B. A., Mehta, P., Jacobs, K. R., . . . Walker, A. K. (2021). Riluzole does not ameliorate disease caused by cytoplasmic TDP-43 in a mouse model of amyotrophic lateral sclerosis. *European Journal of Neuroscience, 54*(6), 6237-6255. doi:https://doi.org/10.1111/ejn.15422

Wright, C. F., West, B., Tuke, M., Jones, S. E., Patel, K., Laver, T. W., . . . Weedon, M. N. (2019). Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting. *American Journal of Human Genetics, 104*(2), 275-286. doi:https://doi.org/10.1016/j.ajhg.2018.12.015

Young, J. J., Lavakumar, M., Tampi, D., Balachandran, S., & Tampi, R. R. (2017). Frontotemporal dementia: latest evidence and clinical implications. *Therapeutic Advances in Psychopharmacology, 8*(1), 33-48. doi:https://doi.org/10.1177/2045125317739818

Yu, E., Ambati, A., Andersen, M. S., Krohn, L., Estiar, M. A., Saini, P., . . . Gan-Or, Z. (2021). Fine mapping of the HLA locus in Parkinson's disease in Europeans. *npj Parkinson's Disease, 7*(1), 84. doi:https://doi.org/10.1038/s41531-021-00231-5

Yüce, Ö., & West Stephen, C. (2013). Senataxin, Defective in the Neurodegenerative Disorder Ataxia with Oculomotor Apraxia 2, Lies at the Interface of Transcription and the DNA Damage Response. *Molecular and Cellular Biology, 33*(2), 406-417. doi:https://doi.org/10.1128/MCB.01195-12

Zarei, S., Carr, K., Reiley, L., Diaz, K., Guerra, O., Altamirano, P. F., . . . Chinea, A. (2015). A comprehensive review of amyotrophic lateral sclerosis. *Surgical neurology international, 6*, 171-171. doi:https://doi.org/10.4103/2152-7806.169561

Zhang, X., Zou, M., Wu, Y., Jiang, D., Wu, T., Zhao, Y., . . . Li, G. (2022). Regulation of the Late Onset Alzheimer's Disease Associated *HLA-DQA1/DRB1* Expression. *American Journal of Alzheimer's Disease and Other Dementias, 37*, 15333175221085066. doi:https://doi.org/10.1177/15333175221085066

Zhang, Y., Lu, Q., Ye, Y., Huang, K., Liu, W., Wu, Y., . . . Zhao, H. (2021). SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. *Genome Biology, 22*(1), 262. doi:https://doi.org/10.1186/s13059-021-02478-w

Zhu, H. (2021). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax (Version R package version 1.3.4). Retrieved from https://CRAN.R-project.org/package=kableExtra

Zou, Y., Carbonetto, P., Wang, G., & Stephens, M. (2022). Fine-mapping from summary data with the “Sum of Single Effects” model. *PLoS Genetics, 18*(7), e1010299. doi:https://doi.org/10.1371/journal.pgen.1010299

Zou, Z.-Y., Zhou, Z.-R., Che, C.-H., Liu, C.-Y., He, R.-L., & Huang, H.-P. (2017). Genetic epidemiology of amyotrophic lateral sclerosis: a systematic review and meta-analysis. *Journal of Neurology, Neurosurgery and Psychiatry, 88*, 540-549. doi:https://doi.org/10.1136/jnnp-2016-315018

Zucchi, E., Ticozzi, N., & Mandrioli, J. (2019). Psychiatric Symptoms in Amyotrophic Lateral Sclerosis: Beyond a Motor Neuron Disorder. *Frontiers in Neuroscience, 13*, 175-175. doi:https://doi.org/10.3389/fnins.2019.00175

Appendix A. Chapter 4 supplementary materials

Appendix A.1. Supplemental methods

Appendix A.1.1. Penetrance calculation procedure

Here follows an outline of the present approach to penetrance estimation. This method is available as an R function (R v4.1.2) accessible at https://github.com/ThomasPSpargo/adpenetrance/.

Step 1

To calculate penetrance using this method, we must identify the rate at which one of the defined disease states (familial, sporadic, unaffected, affected) occurs in families harbouring the variant sampled across a valid combination of two or three of these states (see Table 4-1). This rate is denoted as $R(X)$, and X can be any one of the four disease states for which variant information were provided.

Definitions:

- Familial = more than one family member affected
- Sporadic = only one family member affected
- Unaffected = no family member affected
- Affected = at least one family member affected – familial or sporadic not specified.

In Step 1, we determine $R(X)$ as it is observed within input data, $R(X)^{obs}$. If known, $R(X)^{obs}$ can be specified directly, alongside a corresponding indication of the states from which this estimate is derived. If the familial state is represented within input data, then state X is familial. If only the sporadic and unaffected states are represented, then state X is sporadic. If the affected and unaffected states are represented, then state X is affected.

$R(X)^{obs}$ can also be derived as a weighted proportion of heterozygous variant frequency estimates drawn from samples of unrelated people from two or three of the familial, sporadic, and unaffected disease states or the affected and unaffected states. When variant frequency estimates for the familial or sporadic states are included, the frequency of familial, $P(F|A)$, and sporadic, $P(S|A)$, disease among the affected population, A, must

feature in weightings; note that, as familial and sporadic states are binary outcomes within the affected population, $P(S|A) = 1 - P(F|A)$. Where the unaffected or affected groups are represented, baseline (e.g., lifetime) risk of a population member being affected, $P(A)^{pop}$, must be included within weightings.

In this weighted proportion calculation, we respectively denote variant frequencies for familial, sporadic, unaffected, and affected states as $M_{F,S,U,A}$, to be weighted by the factors $W_{F,S,U,A}$. Given that representation of any two or three of the familial, sporadic, and unaffected disease states or the affected and unaffected states can be used estimate $R(X)^{obs}$, we let the familial, sporadic, unaffected, and affected states be arbitrarily denoted as the states X, Y, and Z. Accordingly, letting $M_{F,S,U,A}$ and $W_{F,S,U,A}$ arbitrarily be $M_{X,Y,Z}$ and $W_{X,Y,Z}$ for the states X, Y and Z,

Equation A-1

$$R(X)^{obs} = \frac{M_X W_X}{M_X W_X + M_Y W_Y}$$

if data are given for a valid combination of two disease states, or

Equation A-2

$$R(X)^{obs} = \frac{M_X W_X}{M_X W_X + M_Y W_Y + M_Z W_z}$$

if data are given for the familial, sporadic, and unaffected disease states. Note that all 4 states cannot be specified together as the familial and sporadic states are subsumed within the affected state. For this reason, it is also unsuitable to represent the affected state alongside data for either or both of the familial or sporadic states. Table 4-1 presents all possible disease state combinations and outlines how the associated weighting factors should be defined to calculate $R(X)^{obs}$.

Step 2

A lookup table to which $R(X)^{obs}$ can be compared for penetrance estimation is generated here. This table stores a series of $R(X)$ values that would be expected at a given value of penetrance, f_i, in a population with average sibship size N, and (optionally) the residual disease risk g for people who do not harbour the variant, which can be calculated according to Equation 4-9 but is assumed to be 0 by default. We denote this series of $R(X)$ values as

$R(X)_i^{ex}$. The sibship size N must be defined alongside the data provided for Step 1 and should represent the average sibship size of the sample from which $R(X)^{obs}$ is determined.

$P(familial)$, $P(sporadic)$, and $P(unaffected)$ are first calculated, following Equation 4-5, Equation 4-6, and Equation 4-7

Equation 4-7, for a sequence of f values, $f_i = (0.0000, 0.0001, \ldots, 1.0000)$, at a specified N and g. This produces a series of values for each disease state: $P(familial)_i$, $P(sporadic)_i$, and $P(unaffected)_i$. If the affected state has been specified in Step 1, we calculate the probability of being affected, $P(affected)$, in accordance with equation 8 at each f_i: $P(affected)_i = P(familial)_i + P(sporadic)_i$. $R(X)_i^{ex}$ can then be derived as an unweighted proportion from $P(familial)_i$, $P(sporadic)_i$, $P(unaffected)_i$, and $P(affected)_i$, taking those disease states which were previously represented in Step 1 and ensuring X represents the same state as before. The lookup table is next constructed, storing an index of corresponding $R(X)_i^{ex}$ and f_i values.

Step 3

The $R(X)^{obs}$ estimate obtained in Step 1 is used to query the lookup table generated in Step 2. The value of $R(X)_i^{ex}$ closest to $R(X)^{obs}$ is identified and the corresponding penetrance value is taken (see Appendix A.1.2.1 and Table A-5 for comparison to a maximum-likelihood approach). This value is an uncorrected penetrance estimate, $f^{unadjusted}$, subject to a systematic bias within the approach and should not therefore be taken as the final estimate; this is determined in step 4. Note that $R(X)^{obs} \approx R(X)_i^{ex}$ unless $R(X)^{obs}$ exceeds or is less than the rate of state X expected between $f = 0, \ldots, 1$ at N and g.

Step 4

This step computes the final penetrance estimate to be returned by the method, $f^{adjusted}$. It corrects for systematic bias in the $f^{unadjusted}$ estimate from in Step 3, which diverges from the true penetrance value according to the combination of states modelled, the value of penetrance, and the structure of families sampled (see Figure A-1; Figure A-2).

In Step 4, firstly, a simulated dataset of 90,000 families is pseudo-randomly generated, where each simulated family is assigned a sibship size of the value $N_i^{(sim)}$. The population generated in this step aims to approximate the sibship structure of the real population sampled for penetrance estimation. To ensure replicability, all pseudo-randomisation in this step is performed using the R seed 24.

By default, simulated sibships follow a Poisson distribution with the lambda defined by the mean sibship size, N, specified for the real sample data. Example simulated Poisson sibship distributions are presented in Figure A-2 A.i-D.i. The Poisson distribution was selected as the default simulation distribution as it is a discrete probability distribution useful for estimating the number of events expected to occur within a given time frame. In this instance, an event is having a child (1 sib) and the time frame is the childbearing years for that family. We note that the Poisson distribution assumes the independence of events and that this assumption would not hold in the present instance (i.e., in real populations, the probability of having additional offspring will be influenced by having already had N_i offspring). However, Figure A-1 demonstrates that the degree of error made in Step 3 penetrance estimates is comparable between the Poisson distribution (Panel A.ii) and other hypothetical population structures (Panels B.ii-D.ii), including the distribution shown in C.i, which resembles that of a UK 1974 population birth cohort (Office for National Statistics, 2020). Therefore, a simulated population in which sib-sizes follow a Poisson distribution can be considered sufficient for approximating the expected error in unadjusted penetrance estimates made using data from randomly sampled populations. This is corroborated by the results of the simulations presented in Appendix A.1.2.3).

If the structure of sibships in the real sample is known, then the user can optionally supply the *adpenetrance* R function with either a vector containing all the sampled sibship sizes or a summary of the sibship distribution, declaring the sibship sizes contained in the sample and the proportion of the sample each sib-size represents. When sibship data are supplied, a simulated sibship distribution is generated based on these data, including only the sibship sizes represented and following its sibship distribution. This 'tailored' simulation population will give more precise $f^{adjusted}$ estimates than those obtained using the default Poisson

distribution (see Appendix A.1.2.3). However, the Poisson distribution is sufficient for adjustment when the sibship distribution of the real data are unknown, under the assumption that population sampling is random and does not exclude families of a particular sibship size (e.g., families of sibship size 0 are not excluded).

A sequence of 25 penetrance values between 0.01 and 1 is also defined, representing true penetrance values of a simulated variant, $f^{true\ (sim)}$. For each $f_i^{true\ (sim)}$, Equation 4-5, Equation 4-6, and Equation 4-7 are applied at each of the $N_i^{(sim)}$ sibship sizes within the simulated population to determine the probability of a family of sibship size $N_i^{(sim)}$ being familial, sporadic, or unaffected at that $f_i^{true\ (sim)}$. One of the familial, sporadic, and unaffected states is pseudo-randomly assigned to each simulated family, according to the probabilities expected at their sibship size and the given $f_i^{true\ (sim)}$. An unadjusted penetrance estimate is then made for the simulated population, $f_i^{unadjusted\ (sim)}$, according to the mean sibship size of the simulated population $N^{(sim)}$, and the $R(X)$ observed, $R(X)^{obs\ (sim)}$. Here, $N^{(sim)} \approx N$, with small variation between these values reflecting the pseudo-randomisation of population generation, and State X is defined as in Step 1, with $R(X)^{obs\ (sim)}$ being calculated as an unweighted proportion of the probabilities of X across the modelled states within the simulated dataset.

The difference between corresponding values of $f_i^{unadjusted\ (sim)}$ and $f_i^{true\ (sim)}$ is calculated: $f_i^{error\ (sim)} = f_i^{true\ (sim)} - f_i^{unadjusted\ (sim)}$. A positive $f_i^{error\ (sim)}$ indicates underestimation of penetrance, while negative values denote overestimation. The relationship between $f^{error\ (sim)}$ and $f^{unadjusted\ (sim)}$ is then established by fitting an nth degree polynomial regression model, which, by extension, also indicates the relationship between $f^{error\ (sim)}$ and $f^{true\ (sim)}$. Polynomial models between 1 and 5 degrees are tested, and the best fitting model is selected based on the Akaike Information Criterion. Figure A-1 (A.ii) and Figure A-2 (A.ii-D.ii) display examples of these error curves fitted for simulated populations where sibship sizes follow a Poisson distribution. Dynamic generation of these regression models is necessary to account for required changes in model fit according to population sibship structure (see Figure A-2).

The fitted polynomial regression model is then used to predict error in the penetrance estimate made for the real dataset in Step 3 based on the value of $f_i^{unadjusted}$, $f_i^{error\ (predicted)}$. The final penetrance estimate is then determined: $f_i^{adjusted} = f_i^{unadjusted} + f_i^{error\ (predicted)}$. The validity of these penetrance estimates is demonstrated in the simulation studies presented in Appendix A.1.2.3).

Optional step

Confidence intervals for the penetrance estimate can be derived through the calculus approach to error propagation (Hughes & Hase, 2010). For this, standard errors, $\sigma_{\overline{M_{X,Y,Z}}}$, of the variant frequency estimates given in Step 1 are required. Using these errors, we calculate the standard error in of $R(X)^{obs}$, $\sigma_{\overline{R(X)^{obs}}}$:

Equation A-3

$$\sigma_{\overline{R(X)^{obs}}} = \sqrt{\left(\frac{\partial R(X)^{obs}}{\partial M_X}\right)^2 \cdot \sigma_{\overline{M_X}}^2 + \left(\frac{\partial R(X)^{obs}}{\partial M_Y}\right)^2 \cdot \sigma_{\overline{M_Y}}^2 + \cdots}$$

Confidence intervals for $R(X)^{obs}$, $\mathrm{CI}_{\mathrm{R(X)}^{obs}}$, can then be obtained through z-score conversion ($CI_{R(X)^{obs}} = R(X)^{obs} \pm z \times \sigma_{\overline{R(X)^{obs}}}$). The lookup table is then queried as in operation 3 for upper and lower bounds of $\mathrm{CI}_{\mathrm{R(X)}^{obs}}$ to attain upper and lower bounds for the $f_i^{unadjusted}$ estimate obtained in Step 3. These values are then adjusted as in Step 4 according to the fitted polynomial regression model, giving the final penetrance estimates at the confidence interval bounds.

Appendix A.1.2. Approach validation and testing

The R scripts used for approach validation are available within our GitHub repository: https://github.com/ThomasPSpargo/adpenetrance/.

Appendix A.1.2.1. Lookup table validation: an alternative maximum-likelihood approach

The unadjusted penetrance estimates obtained in Step 3, $f^{unadjusted}$, can also be derived following a maximum likelihood approach. To validate the lookup table approach

implemented, we additionally derived $f^{unadjusted}$ estimates using Non-Linear Minimisation, leveraging *nlm* and *dbinom* functions available within the R *stats* package (v4.1.2) (R Core Team, 2021).

We constructed this validation approach by defining a negative likelihood function which determines, under a binomial distribution, the likelihood of the specified $R(X)^{obs}$ at a given $f^{unadjusted}$ and N. Within this function, values of $R(X)^{obs}$ are transformed into integers so that they represent a number of state X events across a certain number of trials (e.g., the rate 0.394 would be multiplied by three orders of magnitude, giving 394 events across 1000 trials). The probability function is defined using Equation 4-5, Equation 4-6, and Equation 4-7, and according to the states modelled in calculating $R(X)^{obs}$.

Non-Linear Minimisation was then applied to determine the most likely $f^{unadjusted}$ given $R(X)^{obs}$, N, and g. The starting value for minimisation was defined as the $f^{unadjusted}$ estimate previously determined via the Step-3 lookup approach.

This approach was applied to each of the case studies presented in Table 4-2 and we found negligible difference between the $f^{unadjusted}$ estimates generated within non-linear minimisation and via the lookup table method (see Table A-5). Thus, these findings confirm the validity of the lookup table approach. The alternative maximum-likelihood method was not adopted for penetrance calculation to avoid potential issues in model convergence if starting values are not appropriately defined.

Appendix A.1.2.2. Age-dependent penetrance: tolerance to age of sampling

The penetrance of variant M for an associated disease is determined within the present method according to $R(X)^{obs}$, N, and g. If age of disease onset varies across people harbouring the variant, then penetrance is also age-dependent. In a sample consisting only of families harbouring variant M, $R(X)^{obs}$ will inherently vary over time as people from sampled families age and become affected. Accordingly, penetrance estimates would be lower at an earlier time of sampling, and not accurately represent the true lifetime penetrance. This effect is demonstrated below within a simulation study (see Appendix

A.1.2.3, Figure A-9). Accordingly, a lifetime penetrance estimate is best obtained within this scenario when people sampled are beyond the typical age of onset for the studied trait.

Within a second sampling scenario, where $R(X)^{obs}$ is determined indirectly as a weighted proportion of a given disease state across variant frequency estimates (per Equation A-1 and Equation A-2) from samples of people with and without the variant across a valid combination of disease states, age-dependence will have a smaller effect upon estimation of lifetime penetrance.

This is true if the variability in the rate at which family disease states change over time are comparable between families affected by disease where a variant of interest does and does not occur. To illustrate this assumption with an example: If at a given time 100 of 1000 people with sporadic disease harbour the variant of interest, the variant frequency is 0.1. Suppose then that at a later time of sampling, 200 people of the original sample are now considered 'familial'. If the rate of family disease state change is comparable for people with and without the variant over time, then roughly 180 people without and 20 with the variant would have been reassigned as familial. This leaves 80 of 800 people harbouring the variant in the sporadic sample and the variant frequency remains 0.1. Accordingly, under this assumption, variant frequency estimates within a given disease state will be largely stable over time.

In practice, the rate of change over time is unlikely to correspond exactly between people with and without variant M. However, the assumption is reasonable for a disease with a heritable genetic basis when the tested variant is not thought to be indicative of an entirely distinct onset profile. Accordingly, whether the assumption is true will be influenced by two factors: (1) that variability in age of disease onset is comparable for people who will be affected in their lifetime with and without a given variant, and (2) that the number of disease occurrences (across the range of zero and two or more affected) within families is similar between the groups.

The first of these can be tested by comparing the age of disease onset profile for people with and without a given variant; if the groups have ‘equal onset variability’ over time, then

the assumption is more likely met. The important aspect of this test is that people with and without the variant progress from being unaffected to affected at a similar rate across age; absolute differences in age of onset between group (i.e., where a variant is associated with a younger/older disease phenotype) are tolerated. When equal onset variability is observed, change in $R(X)^{obs}$ over time will be determined by differences number of disease occurrences within families between groups; its estimation will be less affected by age-dependence than when sampling only from families within the variant group.

To facilitate testing of equal onset variability, we have made available an additional R function within the ADPenetrance GitHub repository: *checkOnsetVariability*. This function allows users to supply information regarding age of disease onset for two sample groups (with and without a given variant). The age of onset is then centred for each group by a chosen metric (e.g., mean or median), to enable (base R) plotting of either a density or cumulative density function which overlays onset variability for the two groups. In addition, the function calculates the relative difference in span of time between the first and third quartiles of disease onset in each group. (e.g., if there is an 8-year interval between the first and third quartile for onset among people with variant M, and a 10-year interquartile interval for people without M, then the relative difference is $10/8 = 1.25$, indicating that the variability in disease onset 1.25 is smaller among people with variant M, with less time taken to span the interquartile interval). This number is returned to users of *checkOnsetVariability* as a quantifiable indication of the scale of departure from the equal onset variability. Values of approximately 1 indicate equal onset variability, values >1 indicate that the onset interval is shorter for people in the variant group, values <1 indicate that the onset interval is protracted for people in the variant group. An example of plots returned using the *checkOnsetVariability* function is provided in Figure A-4, which presents testing of equal onset variability in the ALS case studies modelled versus a 'no variant' ALS population, characterised by absence of variants in *C9orf72* and *SOD1*.

The relative difference in onset variability returned by *checkOnsetVariability* can be supplied to a further function also available on GitHub, *simADPenetrance*, which enables users to perform a simulation study that returns a plot which visualises how much a given degree of departure from the assumption may affect penetrance estimates according to sampling age.

The *plyr* (v1.8.7), *ggplot2* (v3.4.0), and *reshape2* (v1.4.4) packages are dependencies for *simADPenetrance* (Wickham, 2007, 2011, 2016).

We present figures from simulation studies, performed using the *simADPenetrance* function, which demonstrate accuracy of lifetime penetrance estimation according to age of sampling and degree of departure from the test of equal onset variability. In these simulations, families containing the variant of interest are compared to a wider disease cohort of families without this variant and instead harbouring one of several other variants of varying penetrance. In Figure A-10, equal onset variability is observed, while Figure A-11 presents a 1.3 relative difference in onset variability, which can be compared to the relative difference of 0.77 (approximately the inverse of 1.3) presented in Figure A-12.

The simulations demonstrate reasonable accuracy in penetrance estimation across time of sampling when the assumption is met, and tolerable stability when the assumption violated by the tested degree of departure.

A full description of these simulation studies is provided subsequently (Section 1.2.3), and documentation for *checkOnsetVariability* and *simADPenetrance* is provided on GitHub.

Appendix A.1.2.3. Simulation studies

Here we present the results of simulation studies conducted to test the validity of the 4-step approach outlined in Appendix A.1.1. The studies described are split into 2 sets according to the methodology followed for generating simulated families. The simulated datasets used within all studies were generated pseudo-randomly in R with no set seed number and $g = 0$ except where stated.

Across both sets of simulation studies, families were pseudo-randomly generated based on sibship distributions previously reported in two distinct samples (see Figure A-3).

The first simulated population (henceforth: the UK population) resembles the sibship distribution across the UK population 1974 birth cohort at the end of their childbearing years (defined as 45 years of age) (Office for National Statistics, 2020). The families within

this simulated dataset were each pseudo-randomly assigned a sibship size between 0 and 4 according to the probabilities observed in this cohort (see Figure A-3) and the mean sibship size, N, is 1.84. The simulation population was modelled on these data because they describe the most recent birth cohort for which data is available at the completion of childbearing years and because the distribution is representative of a randomly sampled population. The distribution of sibship sizes across this cohort is comparable to other reported UK and USA birth cohorts (Kirmeyer & Hamilton, 2011; Office for National Statistics, 2020).

The second population (henceforth: the NS population) was simulated based on the distribution of sibship sizes reported for the Next Steps dataset, a longitudinal sample of children from England (Sheppard & Monden, 2020). Simulated families were pseudo-randomly assigned a sibship size between 1 and 7 according to the probabilities observed in the Next Steps sample (see Figure A-3) and $N = 3.006$. The simulation cohort was modelled on these data to illustrate the application of the method to a sample not fully representative of the population. In this case, the sample does not include families of sibship size 0.

Set 1

In the first set of simulation studies, the performance of the method was tested on simulated populations containing 90,000 simulated families.

A series of ground truth penetrance values, f_i^{true}, were generated for testing within each study. For each f_i^{true}, families from the two simulated populations were generated as described above and the familial, sporadic, and unaffected disease state probabilities expected at each of the occurring sibship sizes were calculated using Equation 4-5, Equation 4-6, and Equation 4-7. One of these three disease states was then pseudo-randomly assigned to each family with the probabilities expected in a family of that the sibship size. Penetrance estimates, $f_i^{adjusted}$, were then made for the population simulated under the specifications of that study. $f_i^{adjusted}$ estimates were made for each possible disease state combination, producing five $f_i^{adjusted}$ estimates for each value of f_i^{true}, across the

combinations of states modelled. The error in $f_i^{adjusted}$ was then determined: $f_i^{error} = f_i^{adjusted} - f_i^{true}$. Positive f_i^{error} values indicate overestimation of penetrance, while negative values indicate underestimation.

In each study, to test the two estimate adjustment approaches allowed in Step 4, we estimated f_i^{error} firstly when the method is supplied no information about the distribution of sibship sizes in the sample data and secondly when this information is supplied. As described in Step 4 (see Appendix A.1.1), the former condition adjusts $f_i^{unadjusted}$ by predicted error in the estimate under a polynomial regression model fitted to a population simulated within the method in which sibships follow a Poisson distribution. The latter condition 'tailors' adjustment of $f_i^{unadjusted}$, by fitting the regression model to a population simulated within the method which directly approximates the real sample data.

Validation under correct parameter specification

We first tested the approach by examining the accuracy of penetrance estimates made using correctly specified input parameters in simulated UK and NS populations harbouring hypothetical variants with known true penetrance values. A sequence of 20 ground truth penetrance values was first defined: $f_i^{true} = (0.05, 0.10, \dots, 1)$ and the populations were simulated as described above. To examine the influence of g, we simulated scenarios where $g = (0, 0.001, 0.1)$. Penetrance estimates, $f_i^{adjusted}$, were made for these populations, defining N according to the mean sibship size of that sample, approximately 1.84 for the UK and 3.01 for the NS populations, and with $R(X)^{obs}$ calculated across all possible disease state combinations. f_i^{error} was then determined. This simulation was repeated 5 times for each value of f_i^{true}, and the results are shown in Figure A-5, averaged across repetitions to determine the mean f_i^{error} observed at each value of f_i^{true}, across each of the disease state combinations. These findings evidence the validity and accuracy of penetrance estimates generated via this approach. They also demonstrate the benefit of supplying about the distribution of sibships in the sample data when this is known; this benefit is greater if sample data does not accurately represent sibship sizes across the population (e.g., where the NS dataset contains no families of sibship size 0).

Misspecification of sibship size

This simulation study examines the accuracy of penetrance estimates when the mean sibship size of sample populations is incorrectly defined. We simulate a wide range of misspecification for sibship size here, although it is likely that degree of misspecification in N would be relatively small for any population-representative sample.

Several values of true penetrance were defined: $f_i^{true} = (0.10,0.25,0.50,0.75,1.00)$. A sequence of values to represent the degree of misspecification in mean sibship size was also specified: $N_i^{modify} = (-1.5, -1.0, \ldots, 3.0)$. The simulated UK and NS populations were generated as before and estimates of $f_i^{adjusted}$ were made, calculating $R(X)^{obs}$ across all possible disease state combinations and defining N according to the mean sibship size of that sample, approximately 1.84 for the UK and 3.01 for the NS populations, adjusted by each value of N_i^{modify}. For instance, if $N = 1.84$ and $N_i^{modify} = -1.5$, penetrance would be estimated based on $N = 0.34$. f_i^{error} was then determined. This simulation was repeated 3 times for each value of f_i^{true}, and the results were averaged across these repetitions.

The results of these simulations are presented in Figure A-6. The increased impact of misspecifying N upon penetrance estimates in the UK compared to NS populations reflects that the difference in disease state rates between a family of 0 sibs compared to a family of 1 sibs is greater than between 1 and 2 or 2 and 3 sib families (etc.); this difference is illustrated in the original description of this disease model (Al-Chalabi & Lewis, 2011). Accordingly, misspecified, and particularly underestimated, N will be more impactful on penetrance estimation in the UK population, which has a lower mean sibship size than NS, since variation in disease state rates is greater between individual family sizes when there are fewer sibs.

Misspecification of disease state rates

This simulation study examines the accuracy of penetrance estimates when $R(X)^{obs}$ is incorrectly estimated. $R(X)^{obs}$ can be supplied directly to the tool or estimated from variant frequency estimates and weighting factors when supplying any valid disease state combination (see Table 4-1). Estimates of $R(X)^{obs}$, and subsequently penetrance, increase

alongside increases in M_X or W_X, and decrease alongside increases $M_{Y,Z}$ or $W_{Y,Z}$. Table A-6 summarises the direction of change in $R(X)^{obs}$ and associated penetrance estimates when values of each input parameter increase for each of the valid disease state combinations.

In this simulation study, several values of true penetrance were defined: $f_i^{true} = (0.10,0.25,0.50,0.75,1.00)$. A sequence of values to represent the degree of error in disease state rate estimates was also specified: $R(X)_i^{modify} = (-0.15, -0.10 \dots, 0.15)$. The UK and NS populations were simulated as before. $R(X)^{obs}$ was calculated for a given f_i^{true} across each of the five possible disease combinations, with the $R(X)^{obs}$ value to be defined in penetrance estimation being adjusted across each value of $R(X)_i^{modify}$; any adjusted $R(X)^{obs}$ values falling outside of the 0 to 1 interval were truncated to be 1x10^{-10} if below that interval or 1 if above. Penetrance estimates, $f_i^{adjusted}$, were made for the simulated UK and NS populations, defining N according to the mean sibship size of that sample, approximately 1.84 for the UK and 3.01 for the NS populations, and $R(X)^{obs}$ by the adjusted value obtained. For instance, if $R(X)^{obs} = 0.366$ and $R(X)_i^{modify} = 0.15$, then penetrance would be estimated based on $R(X)^{obs} = 0.516$. f_i^{error} was then determined for all estimates made. This simulation was repeated 3 times for each value of f_i^{true}, averaging the results across these repetitions. The results of this simulation study are presented in Figure A-7.

Testing the influence of g accuracy upon estimate accuracy

Here we examine how the importance of specifying residual disease risk g varies for penetrance estimation according to the prevalence of the disease, reflected in increased g. We estimate penetrance when g is correctly specified and when assumed that $g = 0$. This is tested for a series of values, where $g = (0,0.001,0.025,0.050,0.75, \dots, 0.2)$. Several values of true penetrance were defined: $f_i^{true} = (0.25,0.50,0.75,1.00)$. Penetrance estimates, $f_i^{adjusted}$, were made for the UK and NS populations, defining N according to the mean sibship size of each simulated sample, approximately 1.84 for UK and 3.01 for NS, and with $R(X)^{obs}$ calculated across all possible disease state combinations. f_i^{error} was then determined. This simulation was repeated 3 times for each value of f_i^{true}, and the results were averaged across each repetition.

The results of this simulation study are shown in Figure A-8. It illustrates that when the disease is rare in the population, and therefore g is small, accounting for g is less critical for attaining accurate penetrance estimates. However, for more common diseases, this is essential.

Set 2

This second set of simulation studies simulations aims to test the influence of age sampling upon the accuracy of penetrance estimation in phenotypes with age-dependent onset. Several simulation scenarios are presented.

In each simulation, several values of true penetrance were tested: $f_i^{true} = (0.25,0.50,0.75,1.00)$. As above, penetrance estimates, $f_i^{adjusted}$, were made for simulated representations of the UK and NS populations, defining N according to the mean sibship size of each simulated sample, approximately 1.84 for UK and 3.01 for NS, and with $R(X)^{obs}$ calculated across all possible disease state combinations. f_i^{error}, which in this simulation reflects difference between the estimate and lifetime penetrance at each time of sampling, was then determined. Each simulation was repeated 3 times for each value of f_i^{true} and the results were averaged across these triplicates.

As before, population structures were firstly generated by pseudo-randomly assigning each family a given sibship size, between 0 and 4 for the UK population and 1 and 7 for the NS sample according to the probabilities of each sibship size per population (see Figure A-3).

For a given family of sibship size N_i, individual family members are then generated, consisting of two parents and N_i siblings. Family members are each assigned relative ages at the time of first sampling, where 0 indicates the final age before the simulated disease becomes onsets in any person with or without the variant. The youngest of N_i siblings is assigned age 0, and the other siblings are, using the *rnorm* function, pseudo-randomly assigned age differences of mean 3 (SD=0.75) which are then summed relative to the age of the next youngest sibling and rounded to the nearest integer. This produces N_i siblings

separated by ~3 years of age. Each of the two parental ages are also assigned using *rnorm*. In a family with $N_i = 0\ or\ 1$, 'parental' ages are generated as mean age 25 (SD=3), rounded to the nearest integer. If $N_i > 1$, the mean age is adjusted in line with the age of the oldest sibling (e.g., if the oldest sibling is 9, then mean parental age is 34).

We simulate a disease which may onset across a 10-year period, where (as above) 0 represents the final age before disease could onset and 10 represents age by which all disease occurrences have onset. We optionally allow the onset window to scale separately within this 10-year window according to variant status (whether or not the variant with penetrance f_i^{true} is harboured). To give an example scenario: all disease occurrences will onset between ages 1 and 10, but onset for people with a variant of f_i^{true} onset may be from ages 1 to 7 versus 1 to 10 in people not harbouring f_i^{true}. Letting the onset scale to be distinct according to variant status enabled us to test the impact of deviation from equal onset variability (see Appendix A.1.2.2). Except where specified, these simulations let disease risk scale equally and onset between times 1 and 10 for people with and without f_i^{true}.

Accordingly, age-dependent disease risk is defined as a proportion of the lifetime risk to an individual according to their current age relative to the disease onset period and whether they harbour, do not harbour, or have 50% probability of inheriting the variant M which has lifetime penetrance f. Accordingly, the disease probability, $P(A)$, for an individual at relative age j is:

Equation A-4

$$P(A)_j^M = Q_j^f \times f\ ,$$

if they harbour M, and Q_j^f is the proportion of people with the variant of lifetime penetrance f affected by time point j; Then,

Equation A-5

$$P(A)_j^{M'} = Q_j^g \times g\ ,$$

if variant M is absent, denoted M', where Q_j^g is the proportion of people with residual risk g who are affected by time point j; Finally,

Equation A-6

$$P(A)_j^{M^{0.5}} = \frac{Q_j^f \times f}{2} + \frac{Q_j^g \times g}{2},$$

if they have 0.5 probability of inheriting M from a variant-harbouring parent, denoted $M^{0.5}$. Equation A-4, Equation A-5, and Equation A-6 respectively mirror Equation 3-2, Equation 3-3, and Equation 3-4 of the main text, with the integration of the Q term.

Let $t = (0, \dots, t, \dots, T)$ denote the time from the first sampling (at $t = 0$) until and including the time when the youngest family member reaches the final age for disease to onset, T. We simulate, using the *rbinom* function, whether each family member is affected at age j_t, according to the probability relevant to that person based on their variant status ($M, M', or\ M^{0.5}$) per Equation A-4, Equation A-5, and Equation A-6. We then sum the number of affected family members at each t, and define the family as 'unaffected' if no family member has disease at t, 'sporadic' if one family member has disease, or 'familial' if two or more family members have disease.

Families generated across the simulated population are then combined. When the number of sampling points until T varies between families, disease state assignments at $t = T$ are duplicated for those families with fewer sampling points until length of t is equal across the population. Penetrance is then estimated for each of the 5 possible disease state combinations at each time t.

Several simulation studies are now presented, demonstrate the effect of age across several scenarios.

Age-dependence when sampling only families harbouring interest variant

As described in Appendix A.1.2.2, $R(X)^{obs}$ will vary greatly in traits with age-dependent onset according to age of sampling when calculated directly from the observed proportions of disease states across a cohort consisting only of people harbouring the variant. We simulate this scenario by generating a cohort of 100,000 families per the above method where each family contains one variant-harbouring parent, one parent not harbouring the

variant, and N_i siblings who have a 50% chance of inheriting the variant. We estimate penetrance based on disease state proportions across the sample for each of the 5 possible disease state combinations at each of $t = 0, \dots, T$ representing the period across which the youngest sibling of each family could become affected.

Figure A-9 presents the results of this simulation. Penetrance estimates varied most when sampling includes the Familial state since most Familial state occurrences will emerge across this time period. Sampling the Sporadic or Affected relative to the Unaffected states has smaller degree of change since the elder generation already have the maximum lifetime risk of disease by $t = 0$. Should $R(X)^{obs}$ be estimated based on disease state proportions across a sample of only people harbouring the variant, we suggest that lifetime penetrance is best estimated based on people in the sample who have passed a typical age for disease onset and since family disease states can reasonably be expected not to change further.

Age-dependence when sampling across families with or without variant across a disease cohort

Age-dependence will affect lifetime penetrance estimation less substantially when $R(X)^{obs}$ is estimated from variant frequencies within each disease state and weighting factors defined by the general characteristics of the disease (see Table 4-1).

We simulate this scenario by generating a general disease cohort across which only certain families harbour the variant of interest, M, which has lifetime penetrance f_i^{true}. Variant M occurs within 100,000 of the generated families. A further 100,000 families are generated, where no family member harbours M, and instead occurs one of several other variants with autosomal dominant inheritance for the disease. Disease risks per age associated with these competing variants are generated as per Equation A-4, Equation A-5, and Equation A-6, but for further variants with lifetime penetrance $f_i^{competing} = (0.2, 0.4, 0.6, 0.8, 1.0)$; 20,000 families are generated for each of the 5 competing variants.

Accordingly, we simulated a total of 200,000 families. Each family contains one parent harbouring variant M or one of the variants with $f_i^{competing}$, one parent not harbouring the

variant of that family, and N_i siblings who have a 50% chance of inheriting the variant. We estimate penetrance, across each of the 5 possible disease state combinations for times $t = 0, \dots, T$ representing the time across which the youngest sibling of each family could become affected. $R(X)^{obs}$ is calculated in accordance with Table 4-1 and Equation A-1 and Equation A-2 as a weighted proportion of the relevant variant frequency estimates observed at each t and the appropriate weighting factors. At all times, weighting factors were defined according to their value at the final sampling time ($t = T$).

Figure A-10 displays the results of this simulation. After Step-4 error correction, and for $f_i = (0.25,0.5,0,75)$ penetrance estimated diverged from the true penetrance by no more than 5% at most sampling times and disease state combinations. When $f_i = 1.0$, error was somewhat greater when sampling the familial, sporadic, and unaffected, or the familial and sporadic states, but within a tolerable distance of true penetrance across all times of sampling. For all values of f_i penetrance was more accurately estimated as age approached the maximum lifetime risk.

In two further simulations, we modelled scenarios alike the previous simulation, but with unequal onset variability between groups. Thus, the onset window for disease differed among people with variant M and those with the competing variants (for an example of this, see Figure A-4). In the first simulation, we let the onset window for people with the variant be 1.3 times shorter than for those without the variant (This is comparable to the relative difference in time spanned by the interquartile interval in people with ALS harbouring the *C9orf72* variant compared to people with no *C9orf72* or *SOD1* variant; shown in Figure A-4). Accordingly, in this simulation all families in which variant M occurred reached their final disease state assignment by $t = 8$, as opposed to $t = 10$ for families where a competitor variant occurred. The results of this simulation are presented in Figure A-11. In the second simulation, we test the inverse of the previous analysis, with the relative onset variability of 0.77 (≈1/1.3), letting instead the onset window be shorter for people harbouring competitor variants (reaching final family disease states by $t = 8$). The results of this simulation are given in Figure A-12.

In both simulations where the variability of disease onset differed between people with and without the variant, penetrance was estimated with tolerable accuracy across all ages and values of f_i^{true}. However, further departure from equal onset variability would have greater impact upon penetrance estimation (see Appendix A.1.2.2).

Appendix A.1.3. ADPenetrance: a companion web tool

This method of penetrance calculation is additionally available as an open-access web tool accessible at https://adpenetrance.rosalind.kcl.ac.uk. This was coded in R (v4.1.2) and leverages the R *shiny* package (v1.7.3) (Chang et al., 2022). An example of the interface and output of this tool is shown in Figure 4-2, as applied to estimation of *SOD1* variant penetrance for ALS using data from a European sample as described in case study 3.

This tool can be used calculate penetrance for a given variant based on an estimate of $R(X)^{obs}$, a defined sibship size, and an estimate of g. State X is assigned to a particular state based on which disease states are included within input data, as indicated by the user. Those states represented can be any two or all three of the familial, sporadic, and unaffected states or the unaffected and affected states. If the familial state is represented within input data, then state X is familial. If only the sporadic and unaffected states are represented, then state X is sporadic. If the affected and unaffected states are represented, then state X is affected.

The user can derive $R(X)^{obs}$ independently, manually specifying the rate of the state requested by the tool. Alternatively, they can provide variant characteristics and weighting factors (see Table 4-1), in order to calculate $R(X)^{obs}$ as described in Step 1. These variant characteristics can be given in each disease state as either (1) variant counts and sample size among population-based samples or (2) directly as variant frequencies.

If data are given using variant counts and sample sizes for each disease state, then the error propagation step is included by default, deriving the standard error for each variant frequency from these values. If data are given using variant frequencies or if $R(X)^{obs}$ is provided directly, then the user can opt to provide error terms for those estimates specified

to enable error propagation. Error terms can be given either as standard errors or as confidence intervals from which standard errors are derived via z-score conversion. The user is asked to select which of these will be provided and, where confidence intervals are given, should indicate the level of confidence that these represent (95% confidence is assumed by default). Wherever error propagation is performed, the user will also need to specify the desired confidence level for the penetrance estimate output. This is to be selected from a series of options, where z-score conversion is used to transform the standard error of $R(X)^{obs}$ into the upper and lower confidence interval bounds of this estimate, which can then be used to estimate the bounds of the penetrance estimate.

The user must also indicate the average sibship size, N across the sample set. This can be specified either manually or by querying a repository of Total Fertility Rate estimates across many world regions which we have integrated within the tool (World Bank, 2020).

g is assumed to equal 0 by default and can optionally be specified to indicate residual disease risk for people within sampled families who do not harbour the tested variant. This term is important for more common phenotypes (e.g., where $g > 0.01$) but will have less influence upon penetrance estimation when $g \approx 0$, as would be the case for rare traits.

Once input data are specified, the tool can be operated and $R(X)_i^{ex}$ is calculated for all values of f_i between 0 and 1 at increasing increments of 0.0001. Penetrance is then estimated as in Steps 3 and 4 and a results table is produced.

The results table presents $R(X)^{obs}$ and the estimated $R(X)_i^{ex}$, $f_i^{unadjusted}$, and $f_i^{adjusted}$ values to which this corresponds, additionally noting which state X represents. $f^{adjusted}$ should be taken as the penetrance estimate. If error propagation is performed, upper and lower confidence intervals and the standard error of the $R(X)^{obs}$ will be provided, alongside corresponding confidence intervals for $R(X)^{ex}$, $f^{unadjusted}$, and $f^{adjusted}$.

Appendix A.2. Supplemental figures

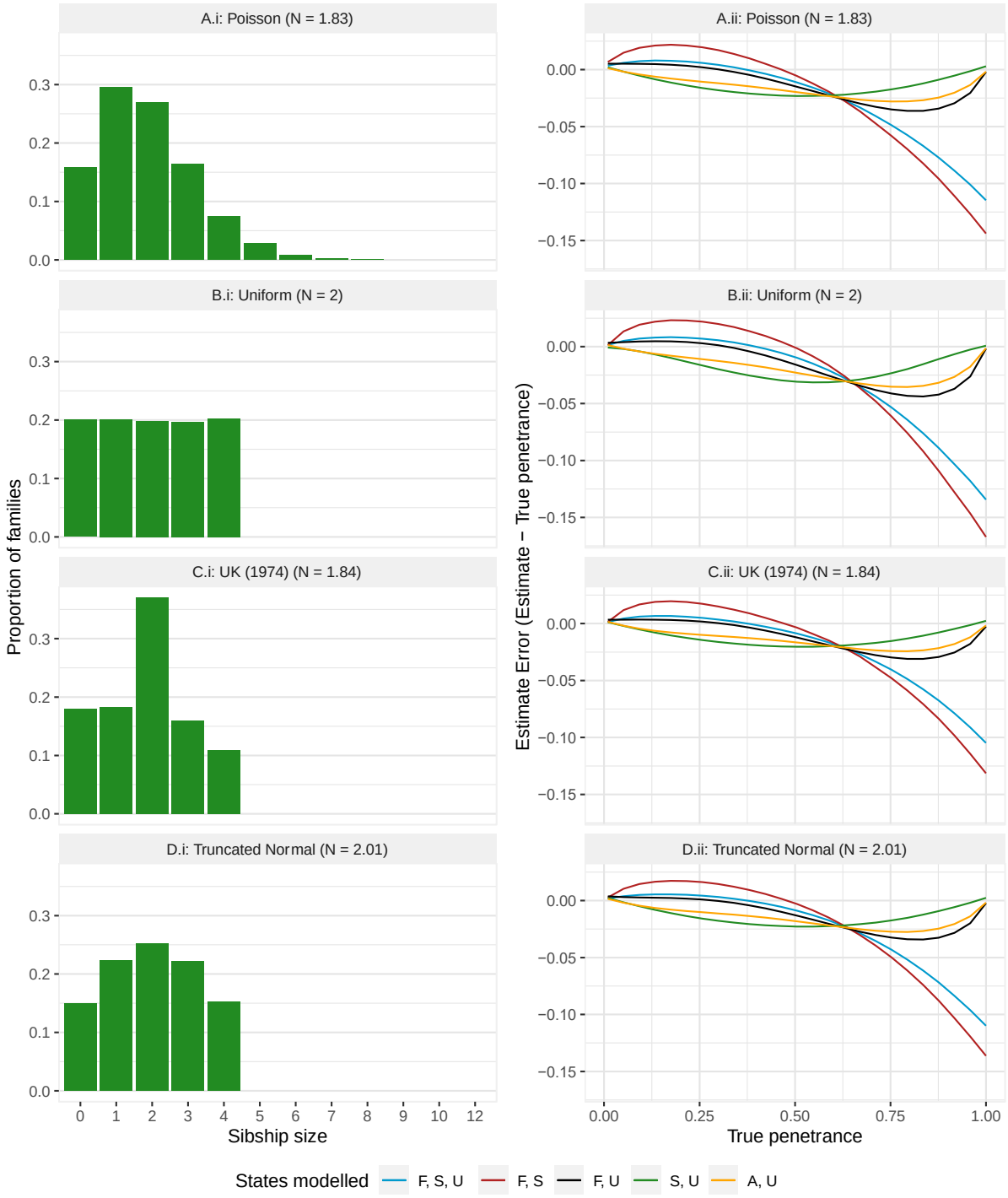


Figure A-1. Error in unadjusted penetrance estimates across true penetrance values and according to states modelled for a simulated population where sibship sizes follow a given distribution.

N = mean sibship size, F = familial, S = sporadic, U = unaffected, A = affected. Panels A.i-D.i show the distribution of sibship sizes across simulated families. Panels A.ii-D.ii display errors in penetrance estimates associated with the corresponding population structure - zero indicates a perfect penetrance estimate, positive values indicate overestimation and negative values underestimation; plotted points display raw error values calculated at each true penetrance value and plotted lines display error values predicted under a fitted polynomial regression model.

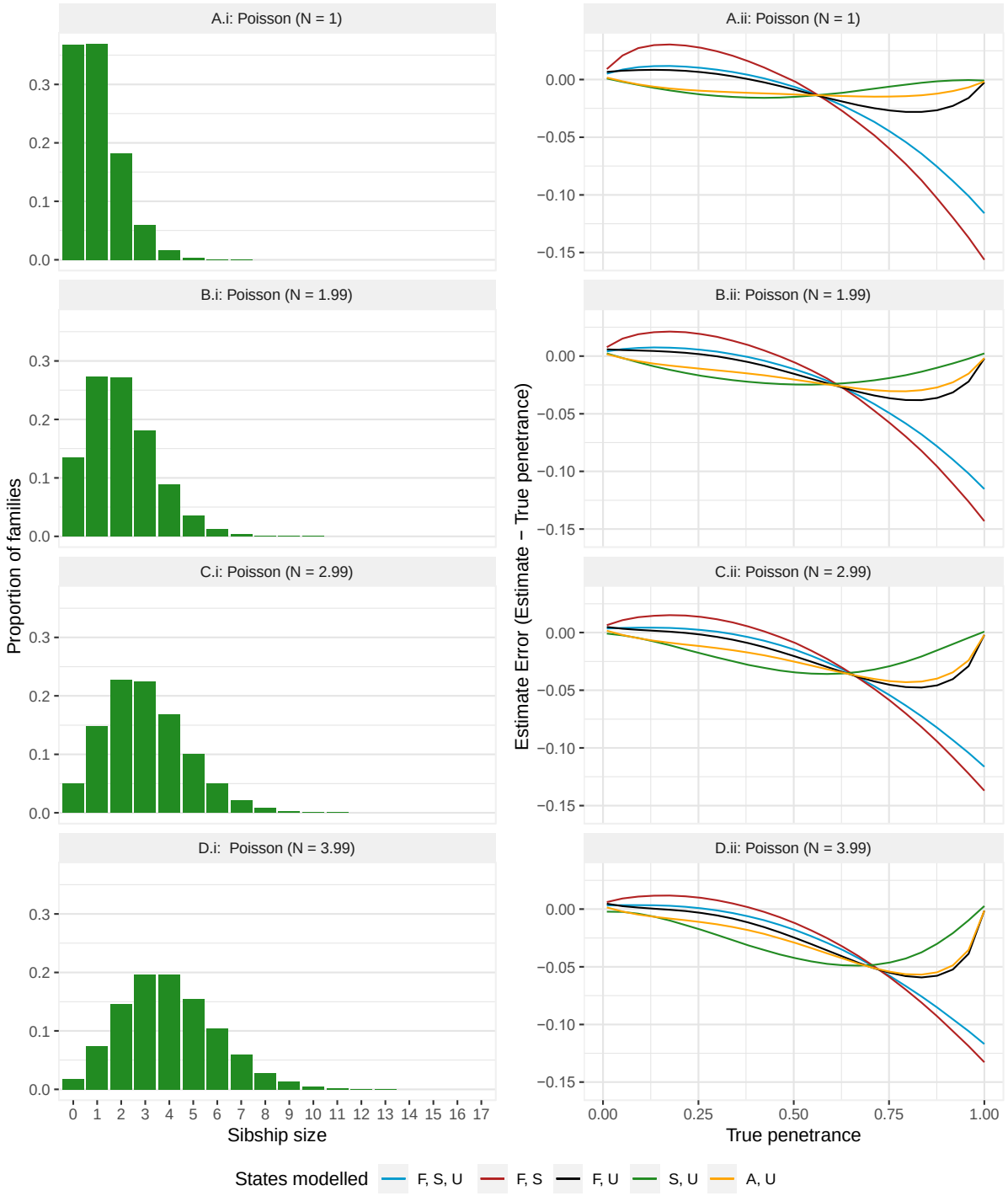


Figure A-2. Errors in unadjusted penetrance estimates across true penetrance values and according to states modelled for a simulated population.

Note: Sibship sizes in the simulated data follow a Poisson distribution varying by mean sibship size (lambda). N = mean sibship size, F = familial, S = sporadic, U = unaffected, A = affected. Panels A.i-D.i show the distribution of sibship sizes across simulated families. Panels A.ii-D.ii display errors in penetrance estimates associated with each corresponding population structure - zero indicates a perfect penetrance estimate, positive values indicate overestimation and negative values underestimation; plotted points display raw error values calculated at each true penetrance value and plotted lines display error values predicted under a fitted polynomial regression model.

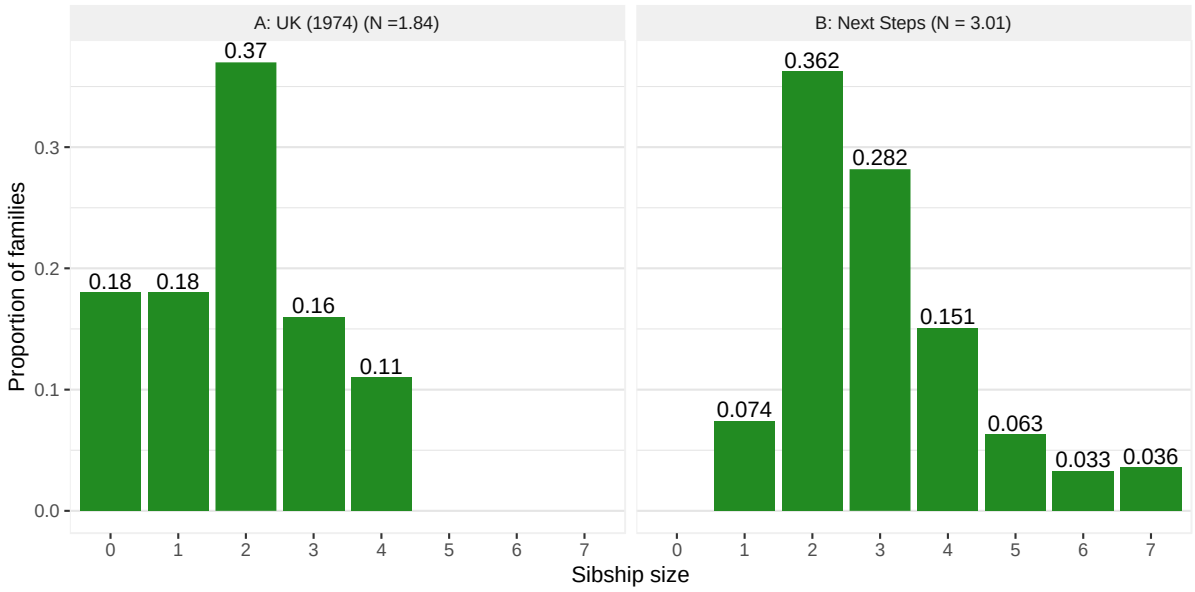


Figure A-3. Sibship distributions upon which simulated populations were modelled across simulation studies

Note: N = mean sibship size. Panel A presents the sibship distribution for the UK population 1974 birth cohort at the completion of their childbearing years; note that the original data reports sibships above size 4 within a collapsed '4 or more' category (Office for National Statistics, 2020). Panel B presents the sibship distribution across English families sampled in the Next Steps cohort study; note that the original data reports sibships above size 7 within a collapsed '7 or more' category (Sheppard & Monden, 2020).

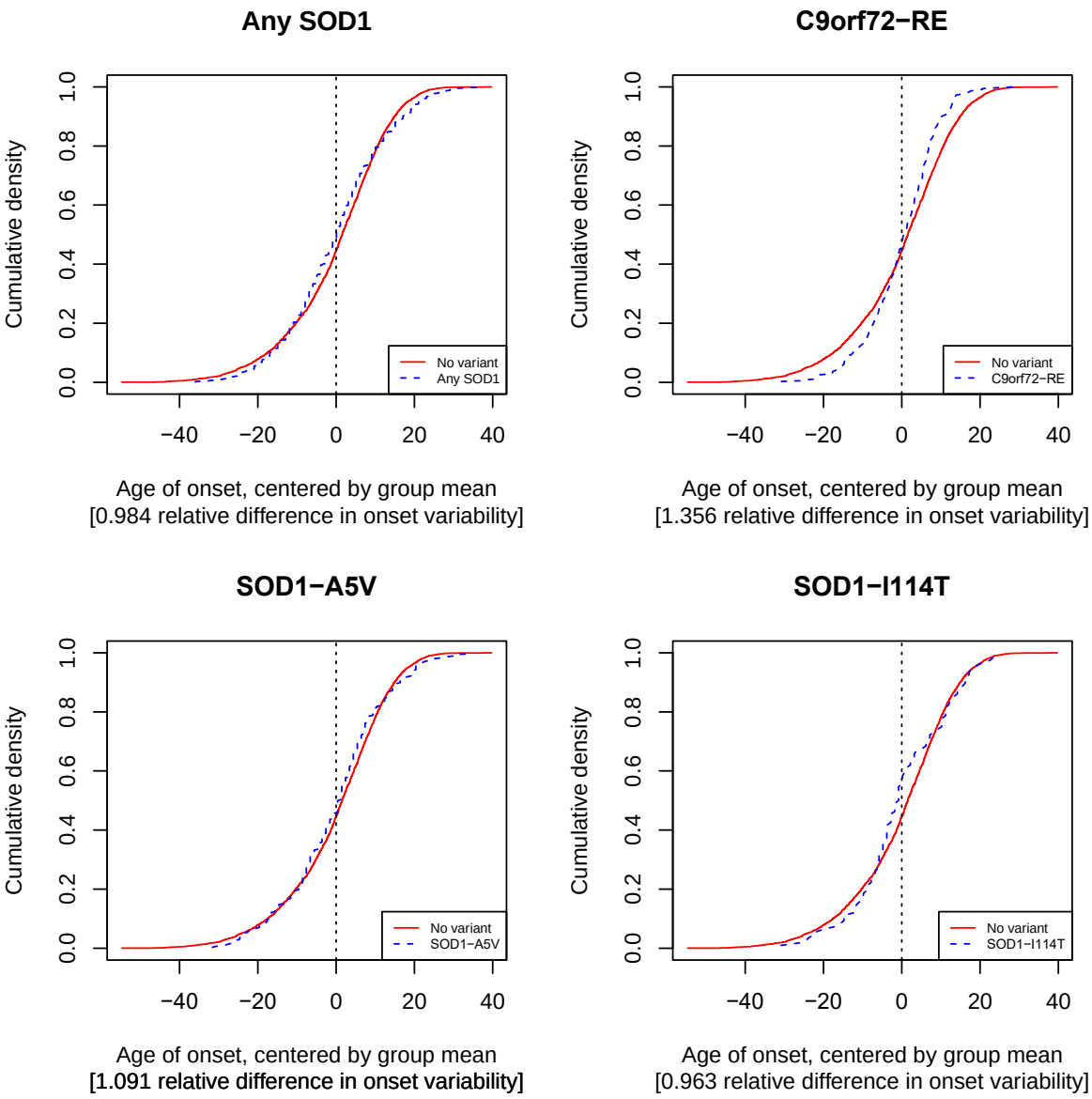


Figure A-4. Cumulative density plots comparing variability in age of ALS onset for people with and without SOD1 or C9orf72 gene variants

Onset distributions for the No variant (n = 5,568) and C9orf72-RE (n = 353) groups are derived from people with ALS from Project MinE (Project MinE ALS Sequencing Consortium, 2018; van der Spek et al., 2019). Those for Any SOD1 (n = 1,315), SOD1-A5V (n = 298), and SOD1-I114T (n = 108) are from a multicentre cohort of people with SOD1 variants (Opie-Martin et al., 2022). The indicated relative difference in onset variability indicates the relative difference in time between the first and third quartile of disease onset for the 'No variant' vs variant groups; values equal to 1 indicate the similar variability age of onset between groups, >1 indicate a shorter interquartile interval in the variant group, while <1 indicates longer interval for the variant group.

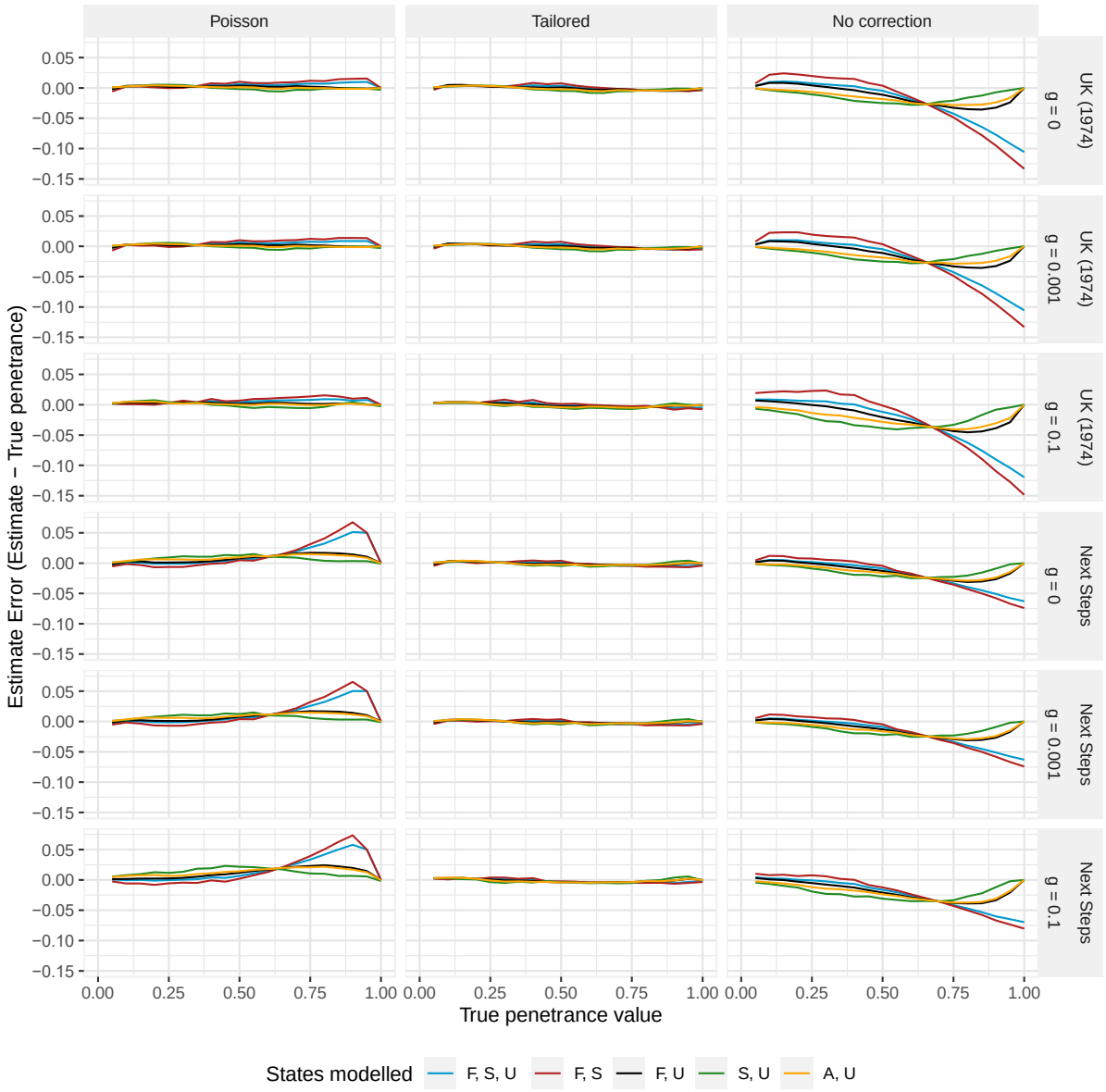


Figure A-5. Error in penetrance estimates across true penetrance values when $R(X)^{obs}$, N and g are specified correctly in the simulated UK (1974) and Next Steps populations

Zero indicates a perfect penetrance estimate, positive values indicate overestimation and negative values underestimation. Plot lines represent estimates made when $R(X)^{obs}$ is defined according to different disease state combinations; F = familial, S = sporadic, U = unaffected, and A = affected - state X is the first state named. The panel rows stratify firstly by population simulated (see Figure A-3) and second by the indicated value of residual disease risk g for people not harbouring the tested variant. The columns stratify by Step-4 estimate adjustment approach (see Appendix A.1.1), which follows either the default approach (denoted Poisson), is tailored to errors predicted under an internally-simulated sibship distribution directly approximating the sample data (denoted Tailored), or displays error with no adjustment made to penetrance estimates (denoted No correction).

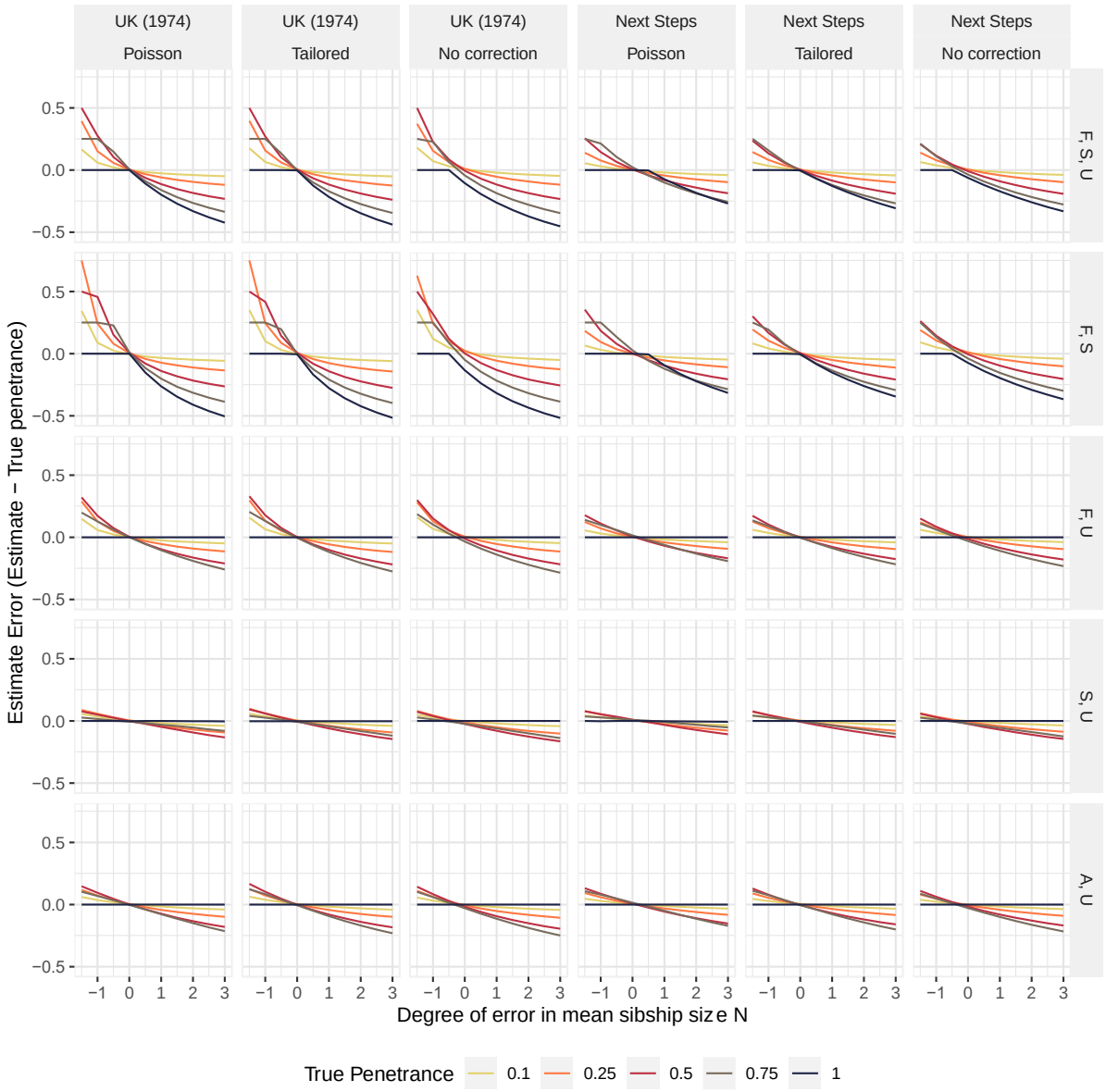


Figure A-6. Error in penetrance estimates according to degree of error in estimation of N

Zero indicates a perfect penetrance estimate, positive values indicate overestimation and negative values underestimation. Plot lines represent different true penetrance values. Panel columns stratify by population simulated (see Figure A-3) and by Step-4 estimate adjustment approach (see Appendix A.1.1), which follows either the default approach (denoted Poisson), is tailored to errors predicted under an internally-simulated sibship distribution directly approximating the sample data (denoted Tailored), or displays error with no adjustment made to penetrance estimates (denoted No correction). Panel rows stratify estimates according to the disease state combination from which $R(X)^{obs}$ is defined; F = familial, S = sporadic, U = unaffected, and A = affected - state X is the first state named.

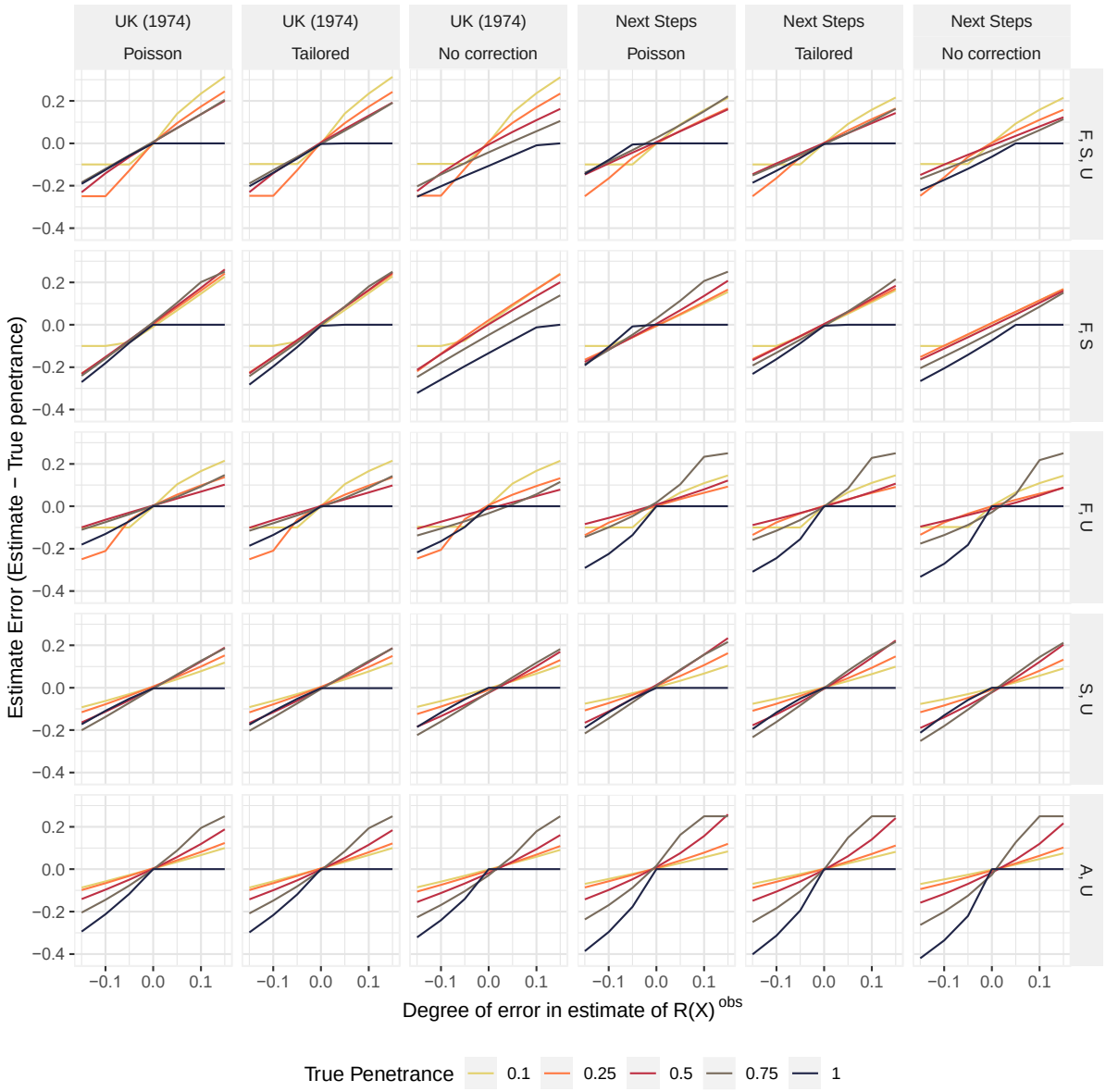


Figure A-7. Error in penetrance estimates according to degree of error in estimation of $R(X)^{obs}$

Zero indicates a perfect penetrance estimate, positive values indicate overestimation and negative values underestimation. Plot lines represent different true penetrance values. Panel columns stratify by population simulated (see Figure A-3) and by Step-4 estimate adjustment approach (see Appendix A.1.1), which follows either the default approach (denoted Poisson), is tailored to errors predicted under an internally-simulated sibship distribution directly approximating the sample data (denoted Tailored), or displays error with no adjustment made to penetrance estimates (denoted No correction). Panel rows stratify estimates according to the disease state combination from which $R(X)^{obs}$ is defined; F = familial, S = sporadic, U = unaffected, and A = affected - state X is the first state named.

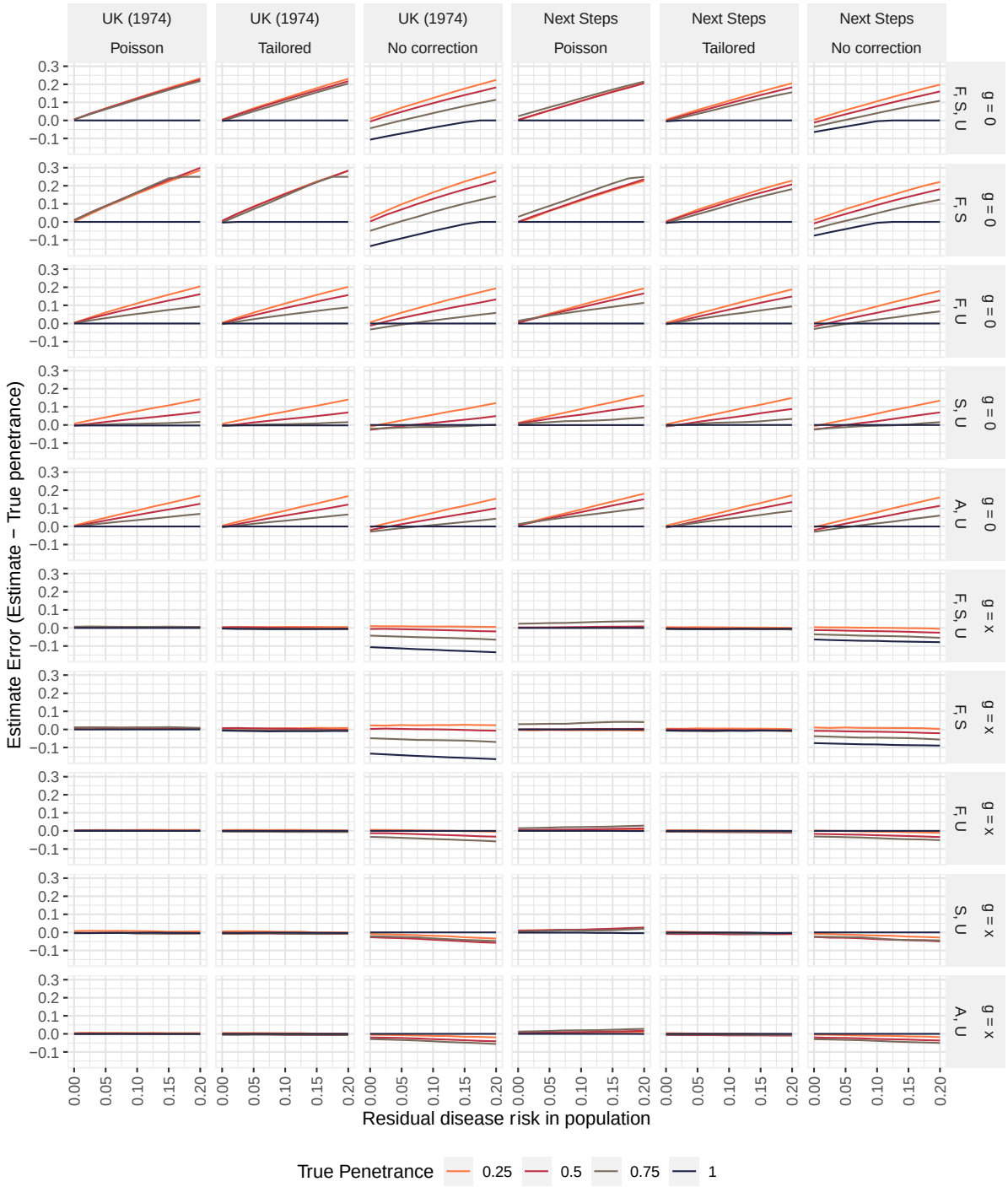


Figure A-8. Error in penetrance estimates according to magnitude of disease risk g for people not harbouring the variant.

Zero indicates a perfect penetrance estimate, positive values indicate overestimation and negative values underestimation. Plot lines represent true penetrance values. The x-axis indicates the probability of developing disease for people not harbouring the variant (g). Panel columns stratify by population simulated (see Figure A-3) and by Step-4 adjustment approach (see Appendix A.1.1), which follows either the default approach (denoted Poisson), is tailored to errors predicted under an internally-simulated sibship distribution directly approximating the sample data (denoted Tailored) or displays error with no adjustment made to penetrance estimates (denoted No correction). Panel rows stratify estimates firstly according to whether penetrance estimates account for g; $g = 0$ rows estimate penetrance under the assumption that people not harbouring the variant do not develop disease; $g = x$ rows make penetrance estimates when risk g is estimated

accurately according to x-axis values. Secondly, they stratify by the disease state combination from which $R(X)^{obs}$ is defined; F = familial, S = sporadic, U = unaffected, and A = affected - state X is the first state named.

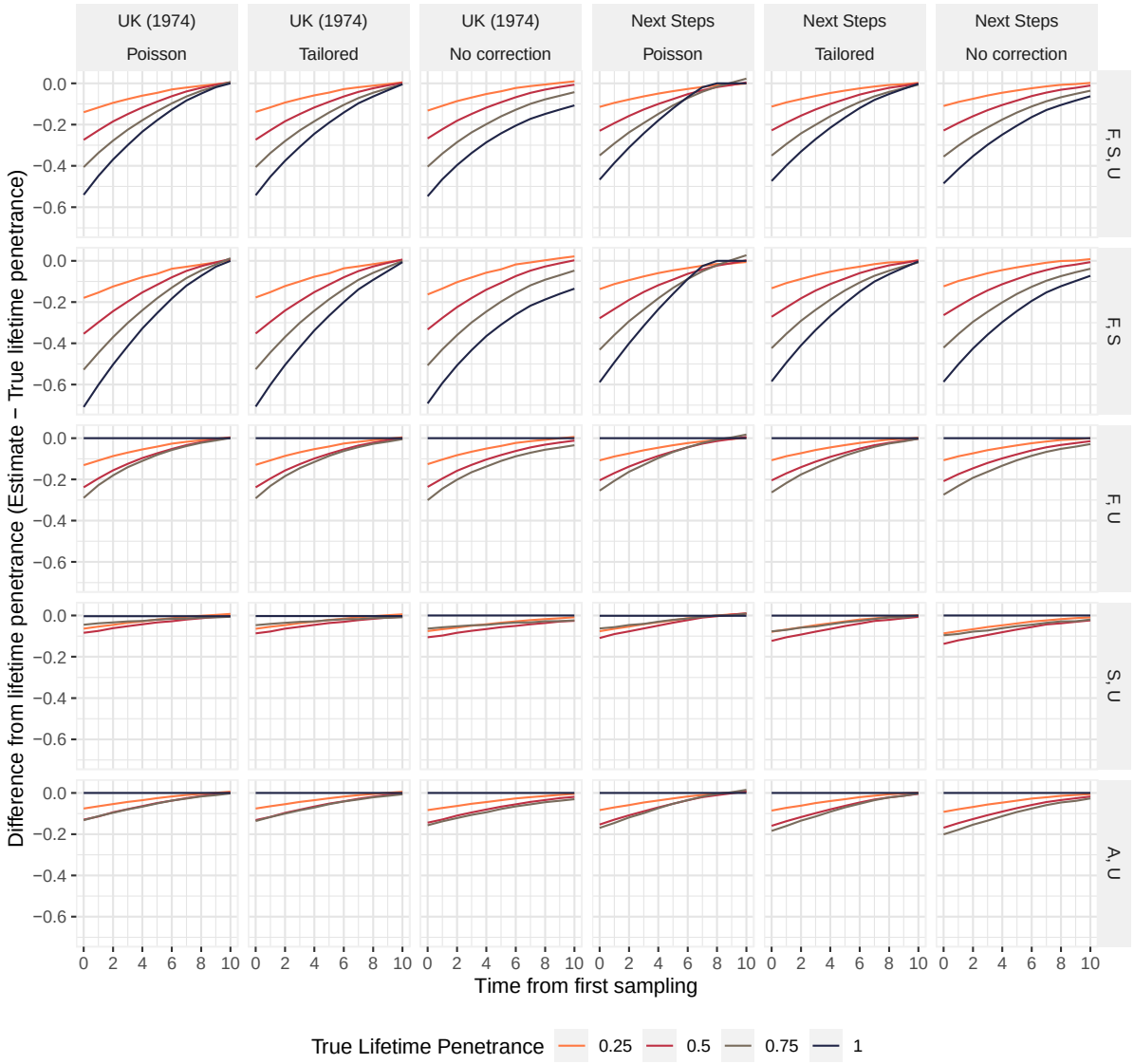


Figure A-9. Penetrance according to age of sampling across only families harbouring a variant of lifetime penetrance f

Zero indicates a perfect estimate of lifetime penetrance, positive values indicate overestimation and negative values underestimation. Plot lines represent true lifetime penetrance values. The x-axis indicates sampling across time from time 0 which is when the youngest sibling reaches the age at which disease may first onset and 10 is the point at which this sib (and therefore all family members) have reached the full lifetime penetrance of disease. Panel columns stratify by population simulated (see Figure A-3) and by Step-4 adjustment approach (see Appendix A.1.1), which follows either the default approach (denoted Poisson), is tailored to errors predicted under an internally-simulated sibship distribution directly approximating the sample data (denoted Tailored) or displays error with no adjustment made to penetrance estimates (denoted No correction). Panel rows stratify estimates according to the disease state combination from which $R(X)^{obs}$ is defined; F = familial, S = sporadic, U = unaffected, and A = affected - state X is the first state named.

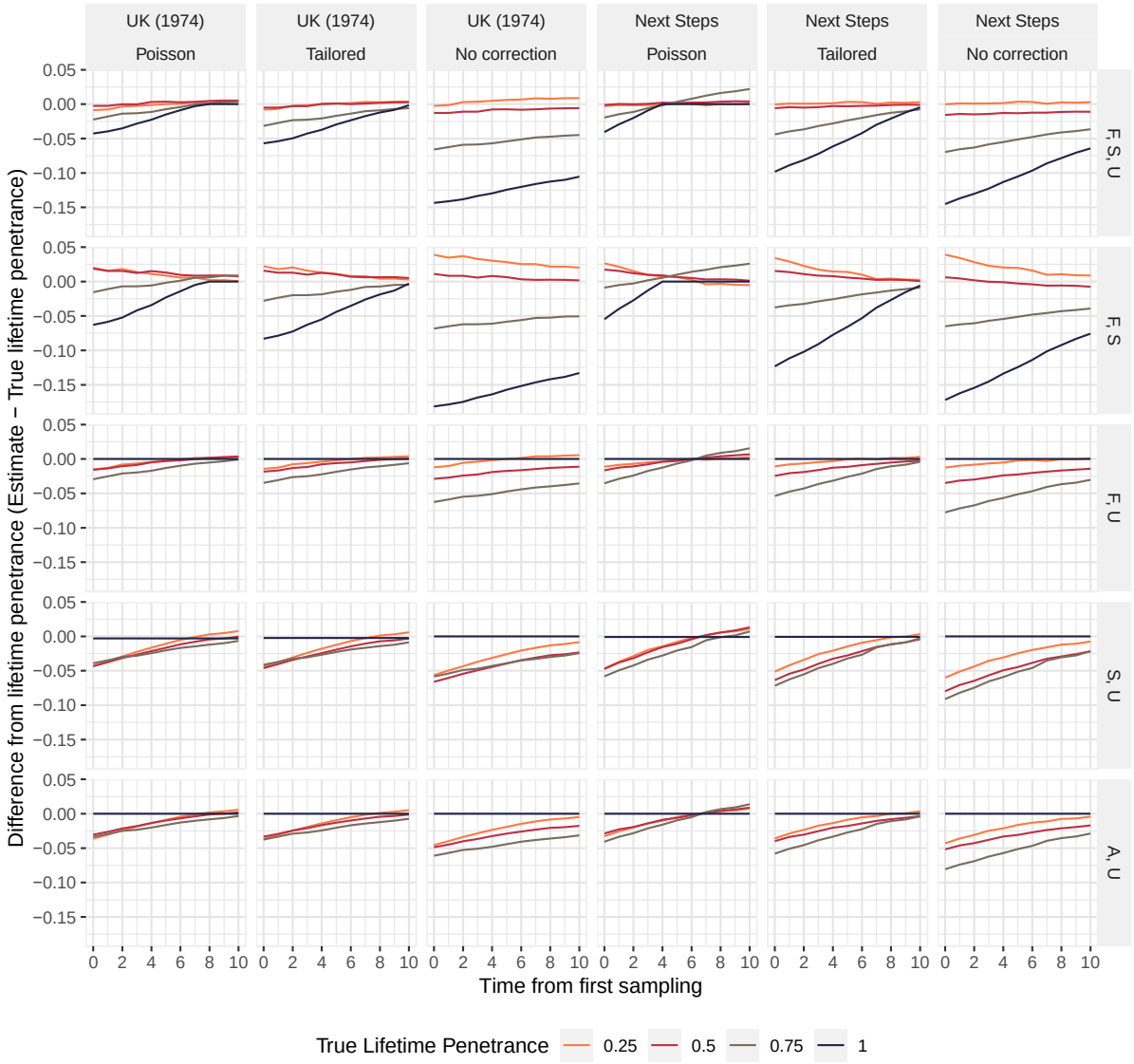


Figure A-10. Error in lifetime penetrance estimates according to age of sampling based on variant frequency estimates across a wider disease cohort with equal age of onset variability

In this simulation equal onset variability is observed. Zero indicates a perfect estimate of lifetime penetrance, positive values indicate overestimation and negative values underestimation. Plot lines represent true lifetime penetrance values. The x-axis indicates time of sampling t between $t = 0$, the final age before the youngest sibling an age where disease may first onset, and $t = 10$, the point at which all family members have reached the maximum age for disease onset. Panel columns stratify by population simulated (see Figure A-3) and by Step 4 adjustment approach (see Appendix A.1.1), which follows either the default approach (denoted Poisson), is tailored to errors predicted under an internally-simulated sibship distribution directly approximating the sample data (denoted Tailored) or displays error with no adjustment made to penetrance estimates (denoted No correction). Panel rows stratify estimates according to the disease state combination from which $R(X)^{obs}$ is defined; F = familial, S = sporadic, U = unaffected, and A = affected – state X is the first state named.

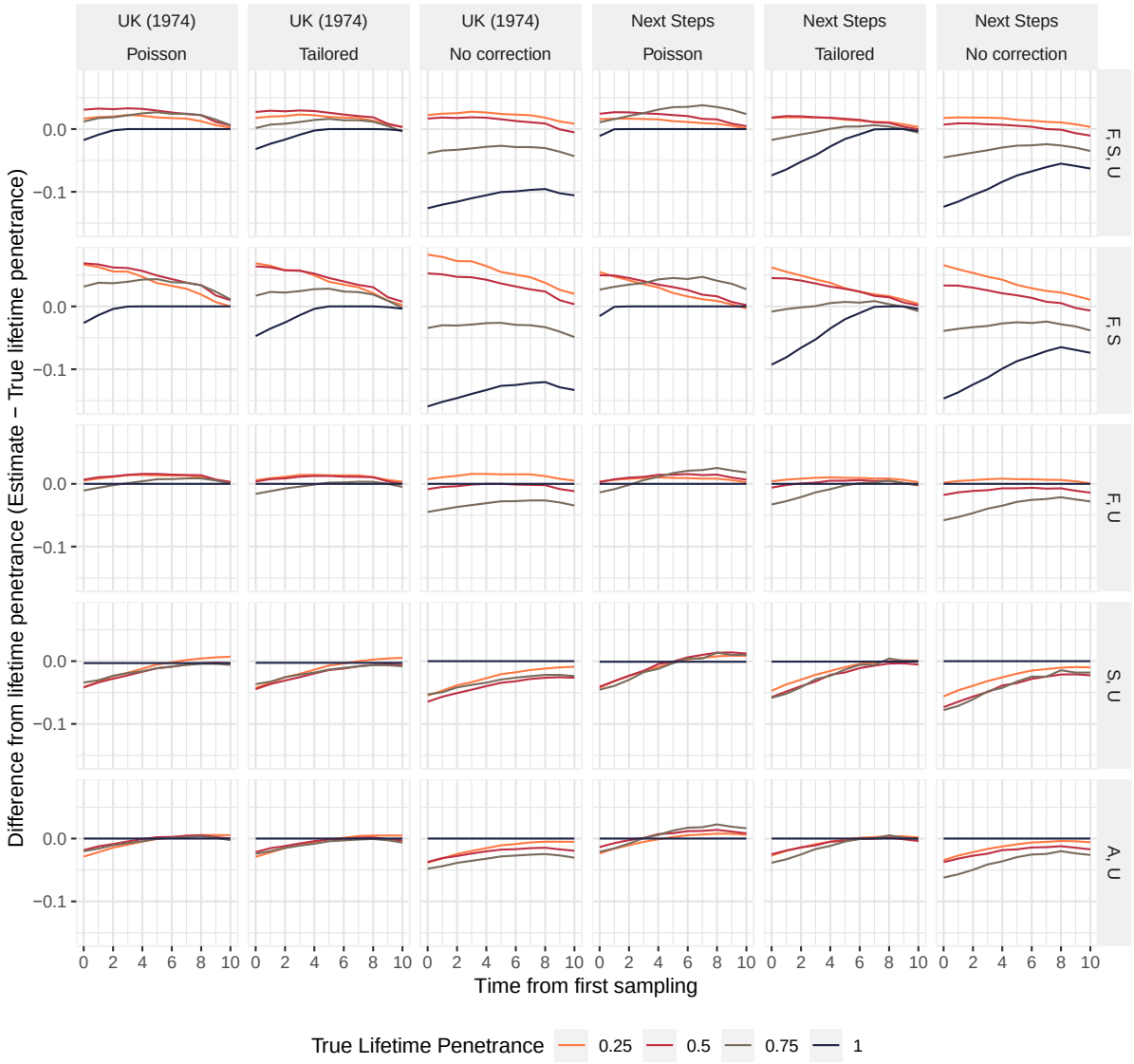


Figure A-11. Error in lifetime penetrance estimates according to age of sampling based on variant frequency estimates across a wider disease cohort when age of onset density is more compressed among people harbouring the tested variant

Equal onset variability (see Appendix A.1.2.2) is not observed; the disease onset window is 1.3 times shorter for people harbouring the tested variant than people without the variant. Zero indicates a perfect estimate of lifetime penetrance, positive values indicate overestimation and negative values underestimation. Plot lines represent true lifetime penetrance values. The x-axis indicates time of sampling t between $t = 0$, the final age before the youngest sibling an age where disease may first onset, and $t = 10$, the point at which all family members have reached the maximum age for disease onset. Panel columns stratify by population simulated (see Figure A-3) and by Step-4 adjustment approach (see Appendix A.1.1), which follows either the default approach (denoted Poisson), is tailored to errors predicted under an internally-simulated sibship distribution directly approximating the sample data (denoted Tailored) or displays error with no adjustment made to penetrance estimates (denoted No correction). Panel rows stratify estimates according to the disease state combination from which $R(X)^{obs}$ is defined; F = familial, S = sporadic, U = unaffected, and A = affected - state X is the first state named.

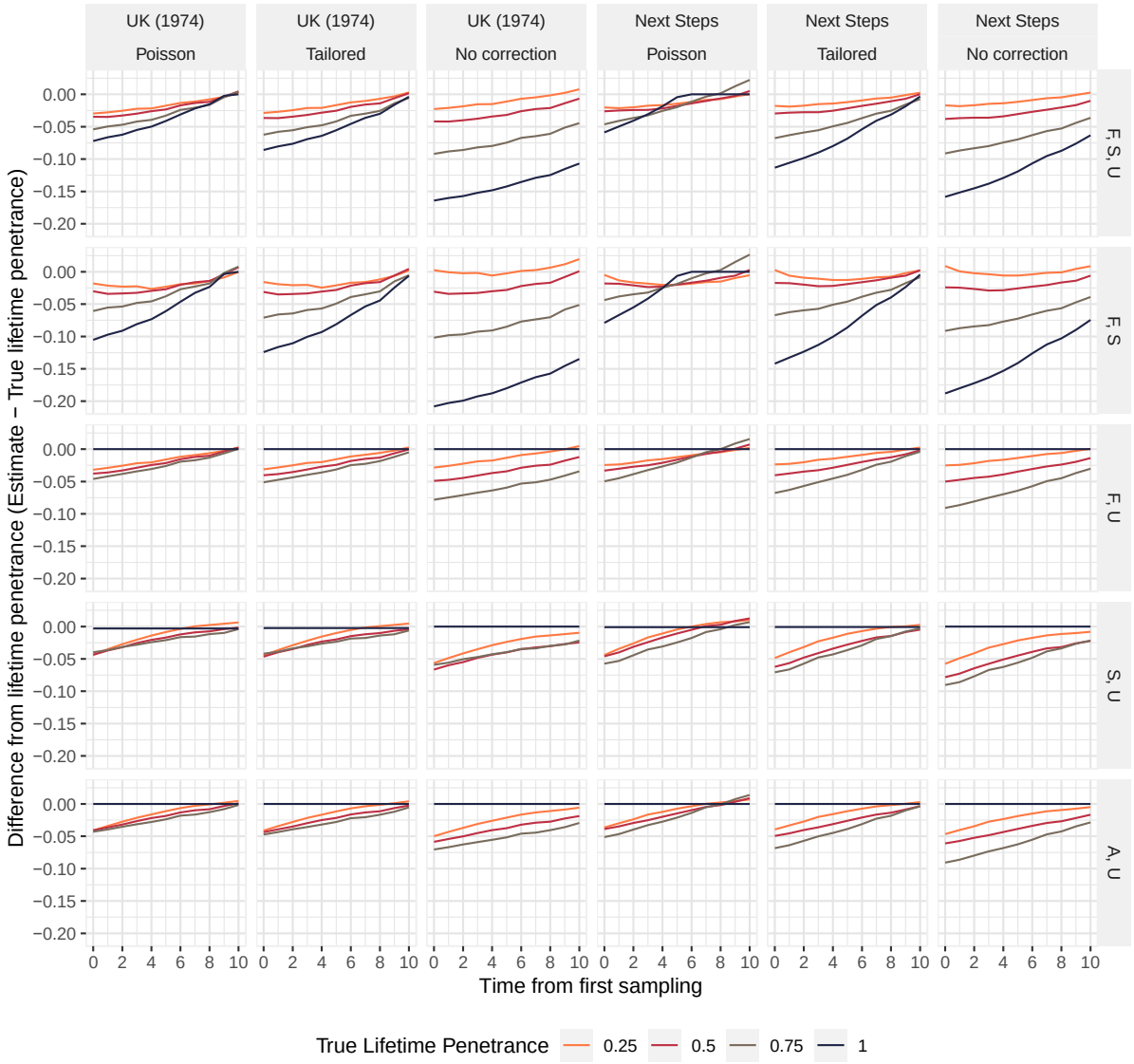


Figure A-12. Error in lifetime penetrance estimates according to age of sampling based on variant frequency estimates across a wider disease cohort when age of onset density is less compressed among people harbouring the tested variant

Equal onset variability (see Appendix A.1.2.2) is not observed; the disease onset density in people harbouring the tested variant is 0.77 times that of people without the variant. Zero indicates a perfect estimate of lifetime penetrance, positive values indicate overestimation and negative values underestimation. Plot lines represent true lifetime penetrance values. The x-axis indicates time of sampling t between $t = 0$, the final age before the youngest sibling an age where disease may first onset, and $t = 10$, the point at which all family members have reached the maximum age for disease onset. Panel columns stratify by population simulated (see Figure A-3) and by Step 4 adjustment approach (see Appendix A.1.1), which follows either the default approach (denoted Poisson), is tailored to errors predicted under an internally-simulated sibship distribution directly approximating the sample data (denoted Tailored) or displays error with no adjustment made to penetrance estimates (denoted No correction). Panel rows stratify estimates according to the disease state combination from which $R(X)^{obs}$ is defined; F = familial, S = sporadic, U = unaffected, and A = affected - state X is the first state named.

Appendix A.3. Supplemental tables

Table A-1. Sample characteristics and calculation of N for data applied in case study 1

[a]Populations sampled were assigned to joint ancestry regions based on the ancestry group most frequent among people from that population. [b]Total Fertility Rate (TFR) estimates for each individual region drawn from the World Bank database (World Bank, 2020): these estimates were assigned to each population using the region defined in the World Bank database that appeared most representative. Where a sampled population features more than one named country, TFR was defined by the country with the larger population. For the Ashkenazi Jewish and Basque populations, estimates were assigned based on the world regions for which their populations are largest. [c]Each weighted TFR value is calculated as: $TFR\ estimate\ for\ region \times percentage\ of\ joint\ population$*; [d]Marked TFR estimates were derived by summation of all weighted TFR estimates attributed to that region.*

Joint population[a]		Population sampled	Number of unrelated people sampled (Healy et al., 2008)				Percentage of joint population	Calculation of sibship size (N)		
			Sporadic	Familial	Unaffected	Total		World region[b]	Weighted TFR estimate[c]	TFR Estimate for region[d]
European ancestry sample	North European	North American (white)	2606	1450	4934	8990	50.20%	North America	0.856381708	1.706
		British	1145	192	1786	3123	17.44%	United Kingdom	0.292961081	1.68
		German and Austrian	803	231	436	1470	8.21%	Germany	0.128868167	1.57
		Norwegian	371	64	572	1007	5.62%	Norway	0.08771679	1.56
		Australian	578	252	0	830	4.63%	Australia	0.080641018	1.74
		French	300	174	348	822	4.59%	France	0.086289575	1.88
		Swedish	200	127	200	527	2.94%	Sweden	0.05179072	1.76
		Irish	236	35	212	483	2.70%	Ireland	0.04719694	1.75
		Polish	153	21	190	364	2.03%	Poland	0.029674465	1.46
		Russian	157	10	126	293	1.64%	Russian Federation	0.025685968	1.57
		Total	6549	2556	8804	17909	100%	-	-	1.687206433[d]

	South European	Italian and Sardinian	2516	633	1040	4189	52.45%	Italy	0.676660406	1.29
		Spanish	806	283	544	1633	20.45%	Spain	0.257648385	1.26
		Basque	117	41	425	583	7.30%	Spain	0.091983471	1.26
		Portuguese	317	85	100	502	6.29%	Portugal	0.089261207	1.42
		Cretan	174	92	0	266	3.33%	Greece	0.044966191	1.35
		Serbian	47	51	161	259	3.24%	Serbia	0.048323316	1.49
		Greek	235	0	0	235	2.94%	Greece	0.03972577	1.35
		Chilean	137	29	153	319	3.99%	Chile	0.065869146	1.649
		Total	4349	1214	2423	7986	100%	-	-	1.314437891[d]
	All European ancestry	North European	6549	2556	8804	17,909	69.16%	-	1.166873142	1.687206433[d]
		South European	4349	1214	2423	7986	30.84%	-	0.405371732	1.314437891[d]
		Total	10,898	3770	11,227	25,895	100%	-	-	1.572244874[d]
Global ancestries sample		Chinese	1360	973	938	3271	9.55%	China	0.161344638	1.69
		Japanese	526	60	372	958	2.80%	Japan	0.039704629	1.42
		Korean	436	17	0	453	1.32%	Korean Republic	0.012917547	0.977
		Indian	718	82	1200	2000	5.84%	India	0.12970638	2.222
		North African Arabs	56	143	739	938	2.74%	Middle East and North Africa	0.076902749	2.809
		Ashkenazi Jews	259	78	410	747	2.18%	North America	0.037195202	1.706
		All European ancestry	10,898	3770	11,227	25,895	75.58%	-	1.188292598	1.572244874[d]
		Total	14,253	5,123	14,886	34,262	100%	-	-	1.64606374[d]

Table A-2. Penetrance estimation of the LRRK2 p.Gly2019Ser variant for Parkinson's Disease across populations sampled in case study 1

Population specific penetrance estimates were made only for those with at least 5 people harbouring LRRK2 p.Gly2019Ser in both the familial and sporadic states. Lifetime disease risk, $P(A)^{pop}$, was 1/37 (0.027) in all calculations; the proportions familial, $P(F|A)$, and sporadic, $P(S|A)$, were respectively 0.105 and 0.895. The familial and sporadic states were modelled in all included populations. Penetrance was also modelled using the unaffected state in only the North African Arabic and Ashkenazi Jewish populations because the variant was sparse in all control samples – occurring predominantly in these two groups. As the variant count was low in the unaffected state for all joint population estimates, we conducted estimates using all states modelled for the All European ancestry and Total Worldwide samples only; these analyses were conducted using the main dataset (Healy et al., 2008) and repeated with variant frequency for the unaffected state estimated using the larger sample of the gnomAD v2.1.1 (controls) database (Karczewski et al., 2020).

[a]Estimated using Total Fertility Rates described for each population in Table A-1; [b] Derived per Equation 4-9; [c]F=familial, S=sporadic, U=unaffected (controls); [d]Step 4 penetrance estimates are shown; [e]Rate of sporadic disease has been calculated here because the familial state is not represented; [f]Unaffected variant frequency estimated from the gnomAD v2.1.1 (controls) sample, using the full sample for total worldwide and the European (non-Finnish) sample for all European ancestry; 95% CI = confidence interval.

	Sample	Variant frequency in state (Healy et al., 2008) (variant count / sample size)			Ave. sibship size[a]	Residual disease risk[b]	States modelled [c]	Familial disease rate among those harbouring the variant across states modelled (95% CI)	Penetrance (95% CI)[d]	
		familial	sporadic	unaffected					Assuming no residual disease risk	Accounting for residual disease risk
Individual population estimates	North American (white)	0.0310 (45/1450)	0.00998 (26/2606)	2.027×10^{-4} (1/4934)	1.706	0.0267	F, S	0.267 (0.174, 0.361)	0.436 (0.277, 0.596)	0.389 (0.23, 0.551)
	Italian and Sardinian	0.0411 (26/633)	0.0147 (37/2516)	9.615×10^{-4} (1/1040)	1.29	0.0266	F, S	0.247 (0.155, 0.339)	0.502 (0.311, 0.694)	0.447 (0.255, 0.641)
	Spanish	0.0495 (14/283)	0.0273 (22/806)	0 (0/544)	1.26	0.0262	F, S	0.175 (0.080, 0.270)	0.361 (0.159, 0.562)	0.304 (0.107, 0.506)
	Portuguese	0.141 (12/85)	0.0410 (13/317)	0 (0/100)	1.420	0.0257	F, S	0.288 (0.135, 0.441)	0.544 (0.246, 0.852)	0.494 (0.197, 0.804)
	North African Arabs	0.3566 (51/143)	0.3929 (22/56)	0.00541 (4/739)	2.809	0.0168	F,S,U	0.064 (0.036, 0.092)	0.185 (0.135, 0.227)	0.166 (0.116, 0.208)
							F, S	0.096 (0.062, 0.130)	0.097 (0.060, 0.135)	0.075 (0.037, 0.113)
							F,U	0.161 (0.026, 0.297)	0.244 (0.101, 0.333)	0.226 (0.081, 0.316)
							S,U	0.643 (0.407, 0.880)[d]	0.513 (0.254, 0.857)	0.506 (0.241, 0.856)

	Ashkenazi Jews	0.282 (22/78)	0.0965 (25/259)	0.00976 (4/410)	1.706	0.0242	F,S,U	0.063 (0.012, 0.115)	0.254 (0.103, 0.355)	0.223 (0.072, 0.323)
							F, S	0.255 (0.158, 0.353)	0.416 (0.249, 0.582)	0.373 (0.207, 0.541)
							F,U	0.078 (0.003, 0.152)	0.232 (0.050, 0.317)	0.203 (0.019, 0.289)
							S,U	0.197 (0.032, 0.363) [d]	0.127 (0.018, 0.263)	0.106 (0.001, 0.246)
Joint population estimates	North European ancestry	0.0274 (70/2556)	0.00809 (53/6549)	$1.136x10^{-4}$ (1/8804)	1.687	0.0268	F, S	0.284 (0.212, 0.356)	0.469 (0.346, 0.592)	0.422 (0.298, 0.546)
	South European ancestry	0.0461 (56/1214)	0.0177 (77/4349)	$4.127x10^{-4}$ (1/2423)	1.314	0.0265	F, S	0.234 (0.173, 0.295)	0.468 (0.343, 0.593)	0.412 (0.288, 0.539)
	All European ancestry	0.0334 (126/3770)	0.0119 (130/10898)	$1.781x10^{-4}$ (2/11227)	1.572	0.0267	F, S, U	0.170 (0.092, 0.249)	0.468 (0.328, 0.587)	0.432 (0.293, 0.552)
							F, S	0.247 (0.202, 0.292)	0.429 (0.348, 0.509)	0.379 (0.299, 0.461)
							F, U	0.354 (0.034, 0.673)	0.494 (0.167, 0.709)	0.466 (0.133, 0.69)
							S, U	0.625 (0.297, 0.952) [e]	0.547 (0.212, 0.950)	0.538 (0.191, 0.95)
				$4.677x10^{-4}$ (10/21383) Ref. (Karczewski et al., 2020)	1.572	0.0267	F, S, U [f]	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 [f]	0.172 (0.081, 0.264)	0.35 (0.247, 0.428)	0.32 (0.215, 0.399)
							S, U [f]	0.388 (0.235, 0.541) [e]	0.293 (0.161, 0.45)	0.275 (0.138, 0.438)
	Total Worldwide	0.03923 (201/5123)	0.01255 (179/14253)	$7.389x10^{-4}$ (11/14886)	1.646	0.0266	F, S, U	0.098 (0.059, 0.137)	0.332 (0.251, 0.402)	0.297 (0.216, 0.366)
							F, S	0.268 (0.229, 0.307)	0.450 (0.382, 0.517)	0.402 (0.334, 0.47)
							F, U	0.134 (0.064, 0.204)	0.304 (0.216, 0.370)	0.273 (0.183, 0.341)
							S, U	0.297 (0.170, 0.424) [e]	0.209 (0.110, 0.324)	0.188 (0.086, 0.307)
				$5.485x10^{-4}$ (30/54699) Ref. (Karczewski et al., 2020)	1.646	0.0266	F, S, U [f]	0.117 (0.089, 0.145)	0.368 (0.315, 0.416)	0.332 (0.28, 0.38)
							F, S	0.268 (0.229, 0.307)	0.45 (0.382, 0.517)	0.402 (0.334, 0.47)
							F, U [f]	0.173 (0.118, 0.227)	0.342 (0.287, 0.389)	0.312 (0.256, 0.36)
							S, U [f]	0.363 (0.273, 0.452) [e]	0.266 (0.189, 0.351)	0.248 (0.168, 0.336)

Table A-3. Penetrance estimation for heterozygous inheritance of widely-described SOD1 variants

Variant frequencies are estimated using the ALS Variant Server (ALS Variant Server) for the familial state, the Project MinE database (van der Spek et al., 2019) for the sporadic state and the European (non-Finnish) population of the gnomAD v2.1.1 (control) database (Karczewski et al., 2020) for the unaffected state.

[a]The number of people heterozygous for the tested variants; sample size = the number of people sequenced for variants at this locus; [b]Proportion sporadic is defined as 1 – proportion familial ($P(S|A) = 1 - P(F|A)$); [c]Estimated based on Total Fertility Rates for the European Union region in 2018 (World Bank, 2020); [d]Derived per Equation 4-9, letting unaffected variant frequency equal 0 for p.Ala5Val and p.Ile114Thr, and familial variant frequency equal 0 for p.Asp91Ala; [e]F=familial, S=sporadic, U=unaffected; [f]The familial disease rate estimate is truncated to 1 as the upper 95% confidence interval bound exceeds the highest possible frequency; [g]p.Asp91Ala is most frequently associated with autosomal recessive ALS presentations, we have modelled the penetrance of its autosomal dominant form only; [h]Rate of sporadic disease has been calculated here because the familial state is not represented – no occurrences of the SOD1 p.Asp91Ala variant are reported in the ALS Variant Server. [h]Step 4 penetrance estimates are shown.

SOD1 variant	Variant frequency in state (variant count[a] / sample size)			Lifetime risk of disease (Alonso et al., 2009)	Proportion familial[b]	Average sibship size[c]	Residual disease risk[d]	States modelled[e]	Familial disease rate among those harbouring the variant across states modelled (95% confidence interval)	Penetrance (95% Confidence interval)[h]	
	familial (ALS Variant Server)	sporadic (van der Spek et al., 2019)	unaffected (Karczewski et al., 2020)							Assuming no residual disease risk	Accounting for residual disease risk
p.Ala5Val	0.006222 (7/1125)	0.000229 (1/4366)	-	0.0025	0.050	1.543	0.0025	F, S	0.588 (0.081, 1[f])	1 (0.133, 1)	1 (0.128, 1)
p.Asp91Ala[g]	-	0.000916 (4/4366)	0.00137 (33/24,143)	0.0025	0.050	1.543	0.0025	S, U	1.59×10^{-3} (0.000, 0.003)[h]	8.98×10^{-5} (0, 0.001)	0.000 (0, 0)
p.Ile114Thr	0.01491 (17/1140)	0.001374 (6/4366)	-	0.0025	0.050	1.543	0.0025	F, S	0.364 (0.149, 0.578)	0.648 (0.255, 1)	0.644 (0.25, 1)

Table A-4. Estimation of the incidence of amyotrophic lateral sclerosis relative to frontotemporal dementia among people of European ancestry who harbour the pathogenic hexanucleotide GGGGCC repeat expansion of the C9orf72 gene (C9orf72^{RE})

Calculations are shown with respect to mathematical notation assigned to each row:

$^{a}E = (C \times B) + (D \times (1 - B));$

$^{b}F = E \times A;$

$^{c}G = F_{ALS}/F_{FTD}.$

		Phenotype		Mathematical notation
		ALS	FTD	
Published data	Lifetime risk (1/N)	1/400 (Alonso et al., 2009)	1/742 (Coyle-Gilchrist et al., 2016)	A
	Familial disease rate (freq.)	0.05 (Byrne et al., 2011; Turner et al., 2017)	0.30 (Turner et al., 2017)	B
	C9orf72^{RE} rate in familial state (freq.)	0.32 (Marogianni et al., 2019)	0.248 (Majounie et al., 2012)	C
	C9orf72^{RE} rate in sporadic state (freq.)	0.05 (Marogianni et al., 2019)	0.060 (Majounie et al., 2012)	D
Estimated value	Overall *C9orf72*^{RE} rate (freq.)	0.064	0.116	E^{a}
	Rate of *C9orf72*^{RE} and phenotype in population (freq.)	1.588x10^{-4}	1.568x10^{-4}	F^{b}
	Incidence relative to FTD among people harbouring *C9orf72*^{RE}	1.012	-	G^{c}

Table A-5. Comparison of unadjusted penetrance estimates derived for the case studies presented in Table 4-2 between the lookup table and maximum-likelihood approaches

[a]All penetrance estimates take into account residual risk g, calculated in accordance with Equation 4-9; [b]F=familial, S=sporadic, U=unaffected (controls); [c]Approach is described in Appendix A.1.2.1; [d]Adjusted penetrance estimates were derived from unadjusted lookup approach estimate, but are representative of both methods. PD = Parkinson's disease, PAH = pulmonary arterial hypertension, ALS = amyotrophic lateral sclerosis, C9orf72^{RE} = the pathogenic C9orf72 GGGGCC hexanucleotide repeat expansion.

Case study	Data subset	Residual disease risk[a]	States modelled[b]	Unadjusted penetrance estimates (95% Confidence interval) [a]		Adjusted penetrance (95% Confidence interval) [a,d]
				Lookup approach	Non-Linear Minimisation[c]	
LRRK2 p.G2019S for PD (Healy et al., 2008)	European ancestry	0.0267	F, S, U	0.338 (0.256, 0.407)	0.338 (0.256, 0.407)	0.334 (0.249, 0.408)
			F, S	0.393 (0.319, 0.464)	0.393 (0.319, 0.464)	0.379 (0.299, 0.461)
			F, U	0.319 (0.218, 0.393)	0.319 (0.218, 0.393)	0.32 (0.215, 0.399)
			S, U	0.257 (0.129, 0.414)	0.257 (0.129, 0.414)	0.275 (0.138, 0.438)
BMPR2 variants for PAH	All variants (Evans et al., 2016)	0.0401	F, S	0.33 (0.289, 0.369)	0.33 (0.289, 0.369)	0.309 (0.267, 0.352)
	All variants (Aldred et al., 2006)	0.0388	F, S	0.237 (0.144, 0.326)	0.236 (0.144, 0.326)	0.212 (0.121, 0.305)
	Small variants (Aldred et al., 2006)	0.0413	F, S	0.25 (0.133, 0.361)	0.249 (0.133, 0.361)	0.225 (0.11, 0.343)
	Large variants (Aldred et al., 2006)	0.0475	F, S	0.162 (0, 0.363)	0.162 (0, 0.363)	0.138 (0, 0.345)
SOD1 variants for ALS (Z.-Y. Zou et al., 2017)	Asian	0.00243	F, S	0.746 (0.626, 0.861)	0.746 (0.626, 0.861)	0.826 (0.661, 1)
	European	0.00245	F, S	0.656 (0.49, 0.809)	0.656 (0.49, 0.809)	0.701 (0.491, 0.926)
C9orf72^{RE} for ALS (Marogiannni et al., 2019)	Asian	0.00247	F, S	0.278 (0.019, 0.511)	0.278 (0.019, 0.511)	0.258 (0.011, 0.518)
	European	0.00234	F, S	0.445 (0.373, 0.514)	0.445 (0.373, 0.514)	0.439 (0.358, 0.52)

Table A-6. Direction of change in $R(X)^{obs}$ and penetrance estimates according to increases in variant frequency and weighting factor inputs

[a]F=familial, S=sporadic, U=unaffected (controls), A= affected.

Parameter (Notation)	States modelled[a]				
	F, S, U	F, S	F, U	S, U	A, U
Variant frequency in familial state (M_F)	↑	↑	↑	-	-
Variant frequency in sporadic state (M_S)	↓	↓	-	↑	-
Variant frequency in unaffected state (M_U)	↓	-	↓	↓	↓
Variant frequency in affected state (M_A)	-	-	-	-	↑
Familial disease rate ($P(F\|A)$)	↑	↑	↑	↓	-
Probability of a person in population being affected ($P(A)^{pop}$)	↑	-	↑	↑	↑

Appendix B. Chapter 5 supplementary materials

Appendix B.1. Supplemental methods

Appendix B.1.1. Estimating analytical validity: sensitivity and specificity

Test performance parameters were defined based on the benchmarking estimates of state-of-the-art tools for analysis of next generation sequencing data. We estimated the probability of a positive test result given the presence of a genetic marker, *P(T|M)* (a.k.a. sensitivity, true positive rate, recall), and the probability of a negative test result given the absence of a genetic marker, *P(T'|M')* (a.k.a. specificity, true negative rate, selectivity). Benchmarking papers we identified provided two performance estimates directly, *P(T|M)* and the probability of mutation given a positive test, *P(M|T)* (a.k.a. precision, positive predictive value). Therefore, to derive specificity, we first calculated the false positive rate, *P(T|M')*, exploiting that

Equation B-1

$$P(M|T) = \frac{P(T|M)}{P(T|M) + P(T|M')} ,$$

which can be rearranged as

Equation B-2

$$P(T|M') = \frac{P(T|M) - P(M|T) \times P(T|M)}{P(M|T)} = \frac{P(T|M)}{P(M|T)} - P(T|M).$$

P(T|M') can then be used to determine specificity:

Equation B-3

$$P(T'|M') = 1 - P(T|M') .$$

Table B-1 presents performance estimates for tools specialised for genotyping various types of genetic variation in sequence data.

Table B-1. Performance benchmarks of next generation sequencing tools specialised for genotyping different types of variants

**Derived per Equation B-2 and Equation B-3.*

Tool	Variant type	Sensitivity	Precision	Specificity*
-	-	P(T|M)	P(M|T)	P(T'|M')
Dragen Pipeline v3 (Illumina, 2019)	Single nucleotide variant	99.96%	99.95%	99.95%
Dragen Pipeline v3 (Illumina, 2019)	Insertion or deletion (small)	99.62%	99.71%	99.71%
ExpansionHunter (Dolzhenko et al., 2019)	Short tandem repeat expansion	99%	91%	90%
GRIDSS (Cameron et al., 2017; Kosugi et al., 2019)	Copy number variant - gene deletion	28.9%	87.6%	95.9%
Wham (Kosugi et al., 2019)	Copy number variant - gene duplication	10.20%	57.1%	92.33%

Appendix B.1.2. Parameter estimates by case study

Several input parameters were defined for each modelled case study scenario:

- *P(D)*, probability of a person having or later manifesting disease *D* prior to testing
- *P(M|D)*, frequency of marker *M* among those affected by *D*
- *P(D|M)*, penetrance, probability of having or later manifesting *D* for people harbouring *M*
- *P(T|M)*, sensitivity (true positive rate) of the testing procedure for detecting *M*
- *P(T'|M')*, specificity (true negative rate) of the testing procedure for identifying the absence of *M*

These were estimated with data drawn from published literature and suitable online genetic databases.

Table B-2 overviews assumptions made across the present case studies and the realities to which they correspond. Estimates of *P(T|M)* and *P(T'|M')* were specified for each scenario of the diseases examined according to the variant type in the assessed gene which is most frequently associated with the considered disease and based on the performances reported in Table B-1. Table 5-1 overviews the final parameter assignments. Below follows a description of their ascertainment.

Table B-2. Case study assumptions

Assumption	Reality
The person undergoing genetic screening will live a normal lifespan	There is no guarantee that a person will live to the age at which a phenotype would onset
Analytical validity is only imperfect at the point of variant calling	Errors can be introduced at any stage of sequencing and data processing, including clerical errors, poor read quality, and incorrect alignment
In recessive diseases, only biallelic mutations are pathogenic and both homozygosity and compound heterozygosity result in equivalent phenotypes	Heterozygous inheritance of variants pathogenic for recessively inherited phenotypes will likely bear some consequence and compound heterozygosity may modify disease presentations
Variant penetrance is defined for the state of disease manifesting	A pathogenic variant may produce clinicomolecular evidence of disease in the absence of a phenotypic disease manifestation
Penetrance is measured only as applied to the disease named	Penetrance of variants with pleiotropic effects can be considered according to pathogenicity for any number of implicated traits

Appendix B.1.2.1. Huntington's disease

For the blind screening scenario of the Huntington's disease (HD) case study, we estimated that $P(D) = 0.00041$, 1 in 2439, representing the frequency of a pathogenic *HTT* CAG short tandem repeat expansion (STRE) of ≥40 repeat units across people sampled from Scotland, the United States of America, and British Columbia (Kay et al., 2016). We deemed this a suitable estimate of *P(D)* because the *HTT* CAG expansion at >40 repeat units is fully penetrant within a normal lifespan and accounts for the vast majority of observed HD cases (Dorsey & Huntington Study Group, 2012; Langbehn et al., 2004; The U.S.–Venezuela Collaborative Research Project & Wexler, 2004). The estimate is also comparable to the frequency of HD cases recorded between 1986-2015 in two Norwegian death registries (Solberg, Filkuková, Frich, & Feragen, 2018). It is sufficiently precise for the purposes of our study.

We specified that $P(M|D) = 1$, letting *M* represent harbouring an *HTT* STRE of ≥40 repeat units. In reality, a small percentage of people who develop HD harbour expansions of fewer repeat units, however, as *P(D)* is defined according to population frequency of ≥40 repeat unit *HTT* CAG expansions it would be inappropriate to define M as less than 1. We similarly defined that $P(D|M) = 1$, in line with the definition of *P(D)* used in this scenario.

In the targeted testing scenario of this case study, we modelled risk for a person whose parent harbours the fully penetrant form of this HTT STRE and who has a 0.5 probability of inheriting an identical variant. Therefore, we adjusted the probability of disease parameter to $P(D) = 0.5$.

Sensitivity and specificity for sequencing HTT were based on the performance of ExpansionHunter (Dolzhenko et al., 2019) for sequencing STREs, $P(T|M) = 0.99$, $P(T'|M') = 0.90$.

Appendix B.1.2.2. Amyotrophic lateral sclerosis

In screening for amyotrophic lateral sclerosis (ALS), $P(D) = 0.0033$, 1 in 300, representing the upper-bound of estimated lifetime cumulative risk of ALS (Alonso et al., 2009; Johnston et al., 2006).

Several scenarios of M were modelled in this case study:

- For the *SOD1* (all) scenario, *M* represents harbouring any *SOD1* variant reported across the familial and sporadic ALS European population sample sets of a large meta-analysis (Z.-Y. Zou et al., 2017)
- For *SOD1* (A5V), *M* represents harbouring the widely described *SOD1* p.A5V single nucleotide variant (SNV)
- For *FUS* (all), *M* represents harbouring any *FUS* variant reported across the familial and sporadic ALS European population sample sets of the previous meta-analysis (Z.-Y. Zou et al., 2017)
- For *FUS* (ClinVar), *M* includes any of 21 *FUS* variants reported as pathogenic or likely pathogenic for ALS within ClinVar and present within databases of familial and sporadic ALS (see Table B-3) (ALS Variant Server; Landrum et al., 2018; van der Spek et al., 2019)
- For *C9orf72, M* represents harbouring a pathogenic hexanucleotide, GGGGCC, STRE of ≥30 repeat units within the first intron of the *C9orf72* gene

Table B-3. FUS gene variants recorded in ClinVar (Landrum et al., 2018) as "pathogenic" or "likely pathogenic" for amyotrophic lateral sclerosis (ALS) and their prevalence in databases of people with familial and sporadic disease

ClinVar variant search performed 24/05/2021. †Occurrence of these variants in people with ALS was estimated based on the ALS Variant Server (ALS Variant Server), a database of familial ALS, and the Project MinE Data Browser (van der Spek et al., 2019), a database of sporadic ALS.

***FUS* gene variant (Protein consequence)**	**ClinVar**		**Occurrence in ALS databases (n/N)†**	
	Classification	**Accession**	**Familial**	**Sporadic**
c.412_429GGACAGCAGCAAAGCTAT[1] (p.138_143GQQQSY[1])	Likely pathogenic	VCV000873229.1	-	-
c.616G>A (p.Gly206Ser)	Pathogenic	VCV000029708.1	-	-
c.646C>T (p.Arg216Cys)	Pathogenic	VCV000016227.1	1/1005	-
c.1394-2del (-)	Pathogenic	VCV000447355.3	-	-
c.1394-1G>T (-)	Pathogenic	VCV000873230.1	-	-
c.1483C>T (p.Arg495Ter)	Pathogenic	VCV000029707.2	1/1012	-
c.1504_1505AG[3] (p.Gly503fs)	Pathogenic	VCV000665141.1	-	-
c.1509dup (p.Gly504fs)	Pathogenic	VCV000933229.1	-	-
c.1520G>A (p.Gly507Asp)	Pathogenic	VCV000016226.1	-	1/4366
c.1540A>T (p.Arg514Trp)	Likely pathogenic	VCV000803253.1	-	-
c.1551C>G (p.His517Gln)	Pathogenic	VCV000016221.1	-	-
c.1553G>A (p.Arg518Lys	Pathogenic	VCV000016223.1	-	-
c.1554_1557del (p.Gln518fs)	Pathogenic	VCV001073222.1	-	-
c.1555C>T (p.Gln519Ter)	Pathogenic	VCV000873231.1	-	-
c.1561C>T (p.Arg521Cys)	Pathogenic	VCV000016224.1	8/1110	-
c.1561C>G (p.Arg521Gly)	Pathogenic	VCV000016222.3	-	-
c.1562G>T (p.Arg521Leu)	Pathogenic	VCV000873232.2	-	2/4366
c.1562G>A (p.Arg521His)	Pathogenic	VCV000016225.1	2/1106	3/4366
c.1571G>T (p.Arg524Met)	Likely pathogenic	VCV000873233.1	1/1123	-
c.1574C>T (p.Pro525Leu)	Pathogenic	VCV000280110.9	4/1126	2/4366
c.1577A>G (p.Tyr526Cys)	Pathogenic	VCV000873234.1	-	-

Table B-4 presents estimates of variant frequency, $P(M|D)$, across the ALS case study scenarios. These estimates were derived as a weighted sum across variant frequency estimates for people with familial and sporadic ALS (i.e., with and without family disease history). $P(M|D)$ was estimated as:

Equation B-4

$$P(M|D) = \left(P(M|D)_{fam} \times 0.05\right) + \left(P(M|D)_{spor} \times 0.95\right),$$

letting $P(M|D)_{fam}$ be the familial ALS variant frequency and $P(M|D)_{spor}$ be for sporadic, and setting 0.05 and 0.95 as weighting factors reflecting that approximately 5% of people with ALS have a family disease history (Byrne et al., 2011).

Table B-4. Estimates of variant frequency among people with ALS, P(M|D), across ALS case study scenarios

*Estimates are based on European ancestry cohorts. †SOD1 (all), FUS (all), and C9orf72: variant frequency (95% CI) from meta-analyses of ALS-implicated genetic variation (Marogianni et al., 2019; Z.-Y. Zou et al., 2017); SOD1 (A5V) and FUS (ClinVar): variant frequency (n/N) based on number of variants (n) across people sampled (N) in familial and sporadic ALS databases (ALS Variant Server; van der Spek et al., 2019). For FUS (ClinVar) the estimate aggregates across variants indicated in ClinVar as pathogenic or likely pathogenic for ALS (see Table B-3) - 'n' is the sum of these variant occurrences respective to each database and 'N' is the median number of people sampled across variants. For SOD1 (A5V) 'n' refers to people harbouring the p.A5V variant of SOD1. *Derived in accordance with Equation B-4, Equation B-5, Equation B-6, Equation B-7, and Equation B-8. CI = confidence interval. P(M|D) = variant frequency in people with ALS; '$P(M|D)_{fam}$' indicates estimates for people with family disease history and '$P(M|D)_{spor}$' is for people without.*

Case study scenario	$P(M\|D)_{fam}$ †	$P(M\|D)_{spor}$ †	$P(M\|D)$* (95% CI)
SOD1 (all)	0.148 (0.115, 0.185)	0.012 (0.007, 0.019)	0.0188 (0.0138, 0.0238)
SOD1 (A5V)	0.00622 (7/1125)	0.000229 (1/4366)	0.000529 (4.43×10^{-5}, 0.00101)
FUS (all)	0.028 (0.021, 0.035)	0.003 (0.001, 0.005)	0.00425 (0.0023, 0.0061)
FUS (ClinVar)	0.0153 (17/1108)	0.00183 (8/4366)	0.00251 (0.00125, 0.00377)
C9orf72 (≥30 repeat units)	0.32 (0.28, 0.37)	0.05 (0.04, 0.06)	0.0635 (0.0538, 0.0732)

We additionally calculated 95% confidence intervals for each *P(M|D)*, $P(M|D)^{95\%CI}$, by propagating the uncertainty in $P(M|D)_{fam,spor}$ (Hughes & Hase, 2010). For this, we first calculated the 95% margin of error, E, for $P(M|D)_{fam,spor}$. When estimating *P(M|D)* based on existing estimates of $P(M|D)_{fam,spor}$ with accompanying 95% confidence intervals, this is:

Equation B-5

$$E_{P(M|D)_i} = P(M|D)_i - P(M|D)_i^{95\%lower},$$

letting i represent either the familial or sporadic states and $P(M|D)_i^{95\%lower}$ denote the lower bound 95% confidence interval of $P(M|D)_i$. When estimating *P(M|D)* based on the number of variants in a sample of size *N*, then this was derived based on the standard error for a proportion *p*,

Equation B-6

$$E_{P(M|D)_i} = \sqrt{\frac{p_i \times (1 - p_i)}{N_i}} \times 1.96\,,$$

multiplying by 1.96 to convert from standard error to the 95% margin of error. $E_{P(M|D)_{fam,spor}}$ can then be summed in quadrature, weighted by the constants from Equation B-4, to obtain E in *P(M|D)*:

Equation B-7

$$E_{P(M|D)} = \sqrt{\left(E_{P(M|D)_{fam}} \times 0.05\right)^2 + \left(E_{P(M|D)_{spor}} \times 0.95\right)^2},$$

from which confidence intervals for *P(M|D)* can be obtained:

Equation B-8

$$P(M|D)^{95\%CI} = P(M|D) \pm E_{P(M|D)}.$$

Penetrance, $P(D|M)$, was estimated for the *SOD1* (all), *FUS* (all), *FUS* (ClinVar), and *C9orf72* ALS case study scenarios with the adpenetrance approach described in Chapter 4. This was selected because the modelled markers are all rare in the population and the approach can provide population-based penetrance estimates for rare variants while avoiding the ascertainment biases inherent when examining the distribution of a variant between people affected and healthy controls. In the original publication describing this method, we previously estimated penetrance for some of the present case study scenarios: $P(D|M) = 0.701\ (95\%\ CI\colon 0.491, 0.926)$ for *SOD1* (all), and $P(D|M) = 0.439\ (95\%\ CI\colon 0.358, 0.520)$ for *C9orf72*. Table B-5 presents derivation of new penetrance estimates for this study: For FUS (all) we estimated that $P(D|M) = 0.579\ (95\%\ CI\colon 0.291, 0.884)$ and, for FUS (ClinVar) $P(D|M) = 0.536\ (95\%\ CI\colon 0.211, 0.877)$.

Table B-5. Estimation of penetrance for FUS variants within the adpenetrance approach based on variant frequencies in people with familial and sporadic ALS

The approach of adpenetrance is described in Chapter 4. FUS (all) denotes any FUS variant identified in people with ALS from a European population (Z.-Y. Zou et al., 2017); FUS (ClinVar) indicates FUS variants identified as pathogenic or likely pathogenic for ALS within the ClinVar database and occurring in familial and sporadic ALS databases (see Table B-3; Table B-4) (ALS Variant Server; Landrum et al., 2018; van der Spek et al., 2019). †FUS (all): variant frequency (95% CI) from previous meta-analysis; FUS (ClinVar): variant frequency (n/N) based on number of variants (n) across people sampled (N) in familial and sporadic ALS databases - 'n' is the sum of these variant occurrences respective to each of the familial and sporadic ALS databases and 'N' is the median number of people sampled across variants. adpenetrance parameter settings: Estimates account for approximately 0.0033 probability of developing ALS among people not harbouring the variant (denoted g). The rate of first-degree family history for ALS is set at 0.050 (Byrne et al., 2011), and average sibship size is estimated at 1.543, the Total Fertility Rates reported for the European Union region in 2018 (World Bank, 2020). CI = confidence interval.

ALS case study scenario	**Variant frequency in familial state†**	**Variant frequency in sporadic state†**	**Familial disease rate among people harbouring the variant across familial and sporadic states (95% CI)**	**Penetrance (95% CI) §**
-	$P(M\|D)_{fam}$	$P(M\|D)_{spor}$	-	$P(D\|M)$
FUS (all)	0.028 (0.021, 0.035)	0.003 (0.001, 0.005)	0.329 (0.172, 0.487)	0.579 (0.291, 0.884)
FUS (ClinVar)	0.0153 (17/1108)	0.00183 (8/4366)	0.306 (0.128, 0.484)	0.536 (0.211, 0.877)

P(D|M) was estimated as 0.91 for *SOD1* (A5V) based on previous figures (Cudkowicz et al., 1997). This estimate was taken in preference to one obtained via adpenetrance because of high uncertainty in the *SOD1* p.A5V penetrance estimate (1 (95% CI: 0.128, 1)), reflecting its low frequency among the sporadic ALS sample; occurring in only 1 person. The two estimates do, however, correspond.

Sensitivity and specificity were defined for the *SOD1* (all), *SOD1* (A5V), *FUS* (all), and *FUS* (ClinVar) scenarios according to the performance of the Dragen Pipeline v3 (Illumina, 2019) for sequencing SNVs: $P(T|M) = 0.9996$, and $P(T'|M') = 0.9995$. This reflects that the ALS-associated risk variants represented in these genes are predominantly SNVs (Abel et al., 2012; Lattante, Rouleau, & Kabashi, 2013).

For the *C9orf72* marker, we modelled two testing scenarios: (1) genetic screening with sensitivity and specificity defined by performance of ExpansionHunter (Dolzhenko et al., 2019) for genotyping STREs, $P(T|M) = 0.99$, $P(T'|M') = 0.90$. (2) using repeat-primed polymerase chain reaction with amplicon-length analysis (Akimoto et al., 2014) as a secondary test to validate a positive NGS screening result from scenario 1. In the second scenario $P(D) = 0.0052$, which is the probability of disease given a positive test result from scenario 1, and sensitivity and specificity are determined by performance of the secondary testing protocol (Akimoto et al., 2014): $P(T|M) = 0.95$, $P(T'|M') = 0.98$.

Appendix B.1.2.3. Phenylketonuria

In screening for phenylketonuria (PKU), $P(D) = 0.0001$, 1 in 10,000, representing the approximate birth prevalence of PKU observed in the United States of America and United Kingdom (Hillert et al., 2020). Established metabolic screening protocols use Tandem Mass Spectrometry to test for elevated phenylalanine concentration (>150μmol/L) in the dried bloodspots of neonates, with confirmatory diagnosis given at phe >600μmol/L (Schulze et al., 2003). This metabolic protocol has an estimated sensitivity, $P(T_{met}|D)$, of 1 and specificity, $P(T'_{met}|D')$, of 0.9995 for detecting PKU. Letting $P(D) = 0.0001$, the probability of disease given a positive metabolic test result, $P(D|T_{met})$, can therefore be estimated:

Equation SB-9

$$P(D|T_{met}) = \frac{P(D) \times P(T_{met}|D)}{P(T_{met}|D) \times P(D) + \left(1 - P(T'_{met}|D')\right) \times \left(1 - P(D)\right)} = 0.167.$$

PKU is caused by variants in the *PAH* gene and has an autosomal-recessive inheritance pattern. Over 50% of *PAH* genotypes associated with PKU are unique to a particular person (Hillert et al., 2020) and the pathogenicity of such variants would be impossible to establish if identified within a genetic screening without further data. Accordingly, we defined *M* as the state of being homozygous or compound heterozygous for any of the three most common *PAH* variants in European ancestry populations of people with PKU, each of which is classified as pathogenic within ClinVar (Landrum et al., 2018). These *PAH* variants and their respective allele frequencies, AF, among people with PKU are: p.Arg408Trp (AF = 0.637), c.1066-11G>A (AF = 0.11), and p.Arg261Gln (AF = 0.11). Their summed allele

frequency is 0.857. Per the Hardy-Weinberg equilibrium, if $q = 0.857$, then $q^2 = 0.734449$, which was taken as *P(M|D)*.

P(D|M) was calculated for this *PAH* marker using a Bayesian approach (Benn et al., 2018; Kirov et al., 2014; Minikel et al., 2016), where:

Equation B-10

$$P(D|M) = \frac{P(D) \times P(M|D)}{P(M)} = \frac{P(D) \times P(M|D)}{P(D) \times P(M|D) + \big(1 - P(D)\big) \times P(M|D')},$$

letting *P(M)* represent the total probability of a person harbouring marker *M* and *P(M|D')* be the probability of M among people without the disease; in rare diseases, *P(M|D')* ≈ *P(M)*. This method was selected because the required input parameters can be readily derived and the approach used for the ALS case study is only suitable in autosomal dominant traits.

The probability of having the marker given no disease, *P(M|D'),* was determined from the allele frequencies of the three variants in the European (non-Finnish) population of the gnomAD v2.1.1. (control) database (Karczewski et al., 2020): p.Arg408Trp (AF = 0.002071), c.1066-11G>A (AF = 0.0004557), and p.Arg261Gln (AF = 0.0004556). Their summed AF is 0.0029823. Therefore, if $q = 0.0029823$, then $q^2 = 0.00000889411329$, which was taken as *P(M|D')*. Per Equation B-10, this was applied alongside the previously determined estimates of *P(D)* and *P(M|D)* to calculate $P(D|M) = 0.8919914195$.

Sensitivity and specificity were defined in this case study according to the performance of the Dragen Pipeline v3 (Illumina, 2019) for sequencing SNVs: $P(T|M) = 0.9996$, and $P(T'|M') = 0.9995$.

Appendix C. Chapter 6 supplementary materials

Appendix C.1. Supplemental methods

Appendix C.1.1. Analysis of gene expression and methylation data

Gene expression analysis

The processing pipeline used to generate the raw counts expression matrix is available at https://github.com/rkabiljo/RNASeq_Genes_ERVs. Briefly, paired FASTQ files were interleaved using BBMap reformat v38.18.0 under default options before adapters were right-clipped and both sides of each read were quality-trimmed with BBMap bbduk v38.18.0. Interleaved files were aligned to hg38 using STAR v2.7.10a (Dobin et al., 2012) before transcripts were quantified using HTSeq (Anders, Pyl, & Huber, 2014). Differential expression between assigned classes was performed using DESeq2 (v4.1.1) (Love, Huber, & Anders, 2014), controlling for sex, age at death, post-mortem delay, RNA integrity number, and surrogate variables. The scripts used to generate this is available at https://github.com/rkabiljo/DifferentialExpression_Genes. Multiple testing correction of differential expression results was performed using independent hypothesis weighting, with an adjusted p-value of <0.05 denoting significance.

Gene enrichment analysis

Gene enrichment analysis was performed using *gprofiler2* (v0.2.1) (Kolberg, Raudvere, Kuzmin, Vilo, & Peterson, 2020) and the following databases: Gene Ontology (Biological Process (GO:BP), Molecular Function (GO:MF) and Cellular Component (GO:CC)), Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, CORUM, TRANSFAC and miRTarBase. The default g:SCS algorithm was used to assess significant enrichments, with the brain expressed gene expression matrix (~34,000 genes) used as a custom gene background.

Cell composition analysis

Cell composition analysis was performed using BRETIGEA (v4.1.3) (McKenzie et al., 2018), under default options. Differences in composition between clusters, based on singular decomposition values for each cell type, was compared using the Wilcoxon rank-sum test, with p-values <0.05 denoting significance.

Biological age analysis

Transcriptional age was estimated from the expression data using RNAAgeCalc (Ren & Kuan, 2020). Biological age was estimated from the methylation beta-value matrix using CorticalClock (Shireby et al., 2020). Methylation- and transcriptional-based age were both subtracted from the chronological age of each sample, which corresponds to the age of death recorded in the KCL Brain Bank. Differences in omics-based age between clusters was assessed with independent samples t-test, with a p-values <0.05 denoting significance.

Appendix C.1.2. Prediction of cluster membership using baseline data

Random forest and eXtreme Gradient Boosting classification algorithms were trained to predict class membership. Feature importance was evaluated across each trained algorithm based on SHapley Additive exPlanations (SHAP) (Covert & Lee, 2021; Lundberg & Lee, 2017). In a multi-class algorithm, SHAP values can be determined for each level of the multiclass outcome variable and indicate feature importance for predicting a certain group relative to all other groups. A binary classification objective must be used to evaluate feature importance for predictions distinguishing between two specific groups.

Accordingly, a total of 12 machine-learning algorithms were trained. Six algorithms were trained with a multiclass objective, across all classes with sufficient data available in the 3 defined data configurations and two machine-learning approaches applied. A further 6 were trained with a binary classification objective, restricting to people in Classes 1 and 2 only for the 3 data configurations and two machine-learning approaches.

We evaluated the performance of the multiclass objective algorithms only as the primary objective was the classification across all clusters. Algorithms trained with both the multiclass and binary objectives were evaluated for the secondary investigation of feature importance.

Appendix C.1.2.1. Machine-learning algorithm training procedure

Only people with no missingness on included features were used when training the algorithms. Therefore, the total sample size under the multiclass (binary) objective was 12,508 (11,109) for data configuration 1, and 3,226 (2,990) for data configurations 2 and 3. For data configurations 2 and 3, Class 5 was excluded from the multiclass objectives as fewer than 20 people in the class remained in the sample.

The clinical diagnosis variable was entered into algorithms using one-hot encoding to represent each level across 3 binary features. Diagnostic delay was, as for other analyses, standardised per-country. No other variables were recoded.

Classification algorithms were trained using *caret* (v6.0.93) (Kuhn, 2022), *randomForest* (v4.7.1.1) (Liaw & Wiener, 2002), and *xgboost* (v1.7.1.1) (T. Chen et al., 2022) R packages. Training was primarily performed within the caret *train* function, to maximise the area under the receiver operating characteristic curve (AUC), as calculated within the *multiClassSummary* function for the multiclass objective and within the *twoClassSummary* function for the binary objective (where the metric is labelled 'ROC'). Class weights in each set of training data were passed to the algorithm to account for class imbalance. Algorithms were trained with 10-fold cross-validation, repeated 10 times.

Repeated cross-validation folds were generated via the *caret createMultiFolds* function which uses stratified subsampling across the groups. The cross-validation resamples were regenerated for each stage of parameter tuning with a different fixed seed to ensure pseudo-randomisation replicability.

Random forest tuning

Random forests were tuned via grid-search (see Table C-12 for the best hyperparameter configuration), using the hyperparameters *ntrees*, *nodesize,* and *mtry* (Probst, Wright, & Boulesteix, 2019).

The *ntrees* hyperparameter was tested at two values, 501 and 1001. This was done to ensure that setting 501 trees was sufficient for classification such that further increases to

the number of trees did not substantially improve performance. *nodesize* indicates the minimum number of observations in the terminal node and was tested at all integers between 1 (the default) and 40. The *mtry* values tested varied by model, ranging between $mtry = 1$ and $mtry = N_{features}$. Therefore, in the clinical data only forests, mtry was tested at every integer between 1 and 7, and in the forests combining clinical and genetic features, every integer between 1 and 17.

eXtreme Gradient Boosting tuning

eXtreme Gradient Boosting algorithms were tuned with the *xgbTree* classifier via a multi-step grid-search (see Table C-12 for the best hyperparameter configuration), tuning the hyperparameters *max_depth*, *min_child_weight*, *gamma*, *subsample*, *colsample_bytree*, *eta*, and *nrounds*. Multiclass objectives were trained with the *multi:softprob setting,* and *binary:logistic* was used for binary objectives.

In eXtreme Gradient Boosting, it is important to control the number of boosting iterations, the *nrounds* parameter, performed by the model at a given learning rate, *eta*, to avoid overfitting to the training sample. Therefore, before the first grid search step, we determined an appropriate *nrounds* for the given data at a reasonably high learning rate (*eta* = 0.3). This was performed using the *xgboost* package *xgb.cv* function and the option for early stopping after further iterations no longer improve model performance in the out-of-fold test data. We set this to stop after 50 rounds with no improvement in AUC as calculated by *xgb.cv* across the 10-fold cross validation dataset and compared the optimum *nrounds* across the 10 resamples, using the 3rd Quartile *nrounds,* rounded up to the nearest integer, across resamples for the subsequent grid search steps: $nrounds_{init}$. The other parameters were set to their defaults: *max_depth* = 6, *min_child_weight* = 1, *gamma* = 0.0, *subsample* = 1, *colsample_bytree* = 1.

We next tuned the tree-based parameters (*max_depth*, *min_child_weight*, and *gamma*) via grid search within the *caret* package *train* functionality, holding *subsample* and *colsample_bytree* at their defaults and using *eta* = 0.3 and *nrounds* = $nrounds_{init}$. Initial grid search was performed across the values: *max_depth* = [5, 6, 7, 8, 9, 10], *min_child_weight* = [1, 2, 3, 4, 5, 6], *gamma* = [0.0, 0.1, 0.2, 0.3, 0.4]. If an optimum parameter value was

identified at the edge of the grid search (and not at the limit for that parameter) a smaller subsequent search was performed, recursively extending any applicable parameters by several additional steps outside the values tested until the optimum was found.

The parameters *subsample* and *colsample_bytree* were next trained using the *train* function, taking the initial values of [0.6, 0.65, 0.70, …, 1] for each, setting *eta* = 0.3 and *nrounds* = $nrounds_{init}$, and lastly *max_depth*, *min_child_weight*, and *gamma* to the optimum values identified prior. As before, the grid search was recursively adjusted when optimum parameters were found at the edge of the grid search.

Finally, the parameters *eta* and *nrounds* were tuned together in a 2-step process, holding all other parameters at the optimum determined prior. First, using *xgb.cv* and the early stopping functionality as before, we determined the optimum number of *nrounds* for each of *eta* = [0.01, 0.02, 0.04, 0.06, 0.08, …, 0.3] as the median *nrounds* across the 10 cross-validation resamples rounded up to the nearest integer. Second, for each pairing of *nrounds* and *eta* we assessed in the *train* function which provided the optimum performance; recursive grid search adjustment was not performed here. The best tuning parameters at this stage were accepted as the optimum tuning for eXtreme Gradient Boosting.

Appendix C.1.2.2. Assessment of model performance and feature importance

Metrics of sensitivity, specificity, precision, and balanced accuracy were calculated via the *caret* package *confusionMatrix* function. Receiver operator characteristic curves were generated and the area under the curve calculated using *pROC (Robin et al., 2011).* In multiclass objectives, curves were generated for each group vs all other groups and for all pairwise group combinations. In binary objectives, a single curve was generated using Class 1 as the reference group and Class 2 as the 'positive' value; the other performance metrics were also encoded in this direction.

Feature importance was determined using SHAP values calculated via the R package *kernelshap* (v0.3.3) (Mayer, 2023). In multiclass objective algorithms, absolute mean SHAP values were estimated for prediction of each outcome category based on classification probabilities across the full training sample. In binary objectives, absolute mean SHAP values

were estimated for classification probability across the overall algorithm. A pseudorandomised subset of approximately 500 records were extracted from the dataset using stratified sampling via the *caret createDataPartition* function and supplied as background data required by *kernelshap*; class imbalance was accounted for in *kernelshap* by weighting these data according to class. SHAP values were visualised using *ggplot2*.

Appendix C.2. Supplemental figures

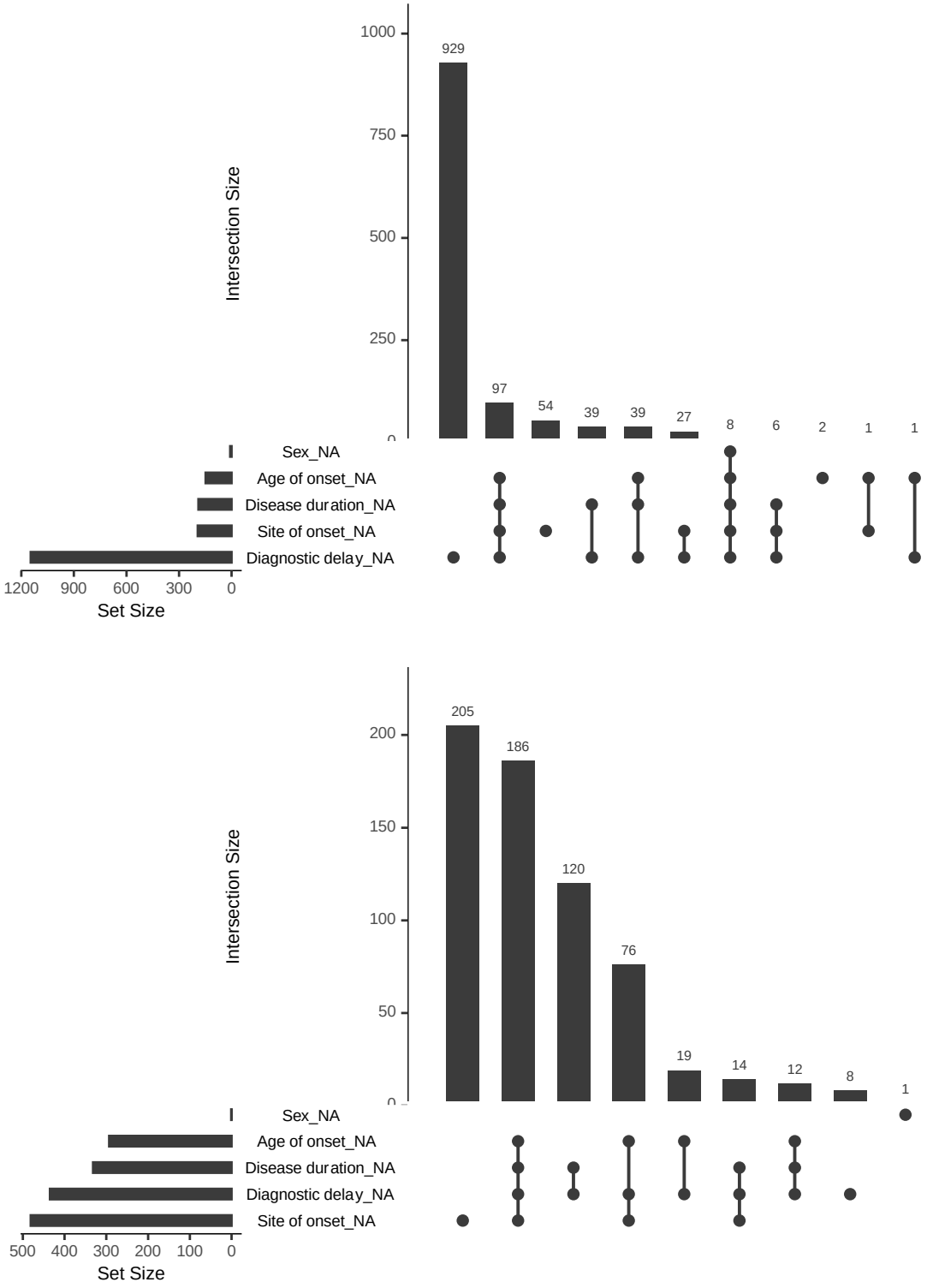


Figure C-1. Upset plots for missingness across clinical features used in LCA for the Project MinE (top) and STRENGTH (bottom) cohorts

Plotting was performed with the R naniar package (v0.6.1) (Tierney, Cook, McBain, & Fay, 2021).

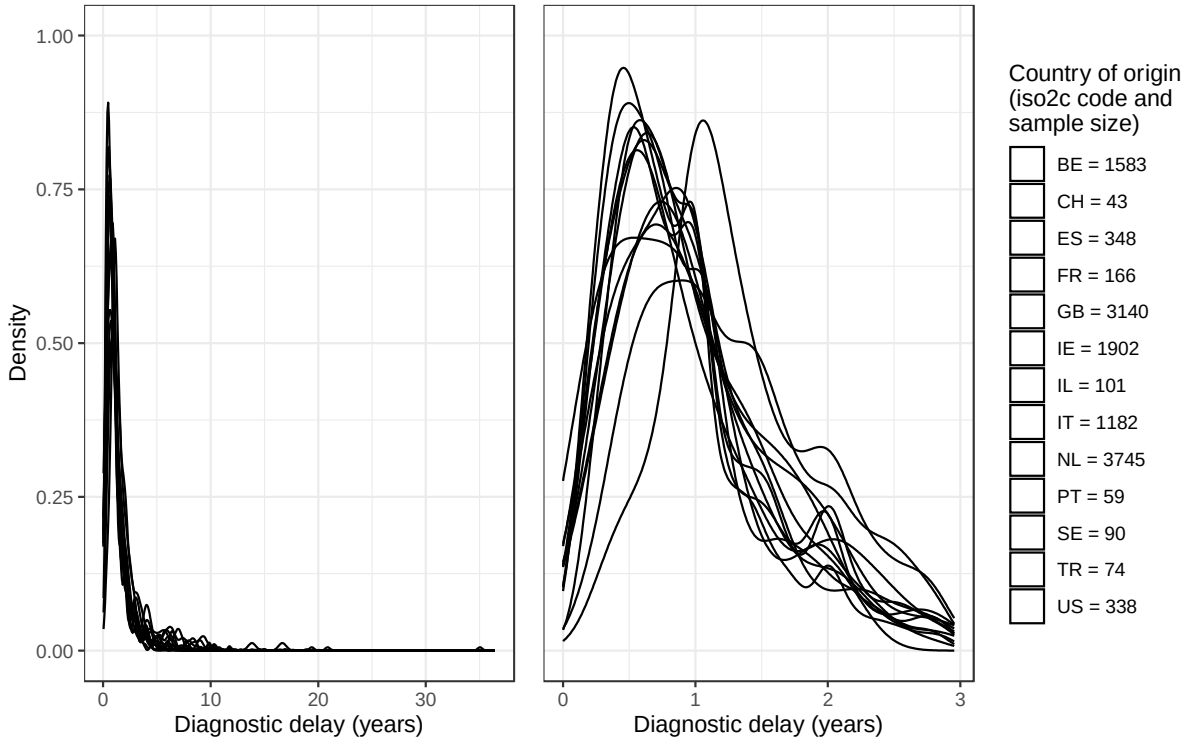


Figure C-2. Density plot for diagnostic delay across countries for the combined Project MinE and STRENGTH datasets

The two panels display the same data; the left panel visualises the full distribution of diagnostic delay, and the right panel x-axis is truncated to exclude the final decile of records. Table C-1 presents mean and standard deviations for each country.

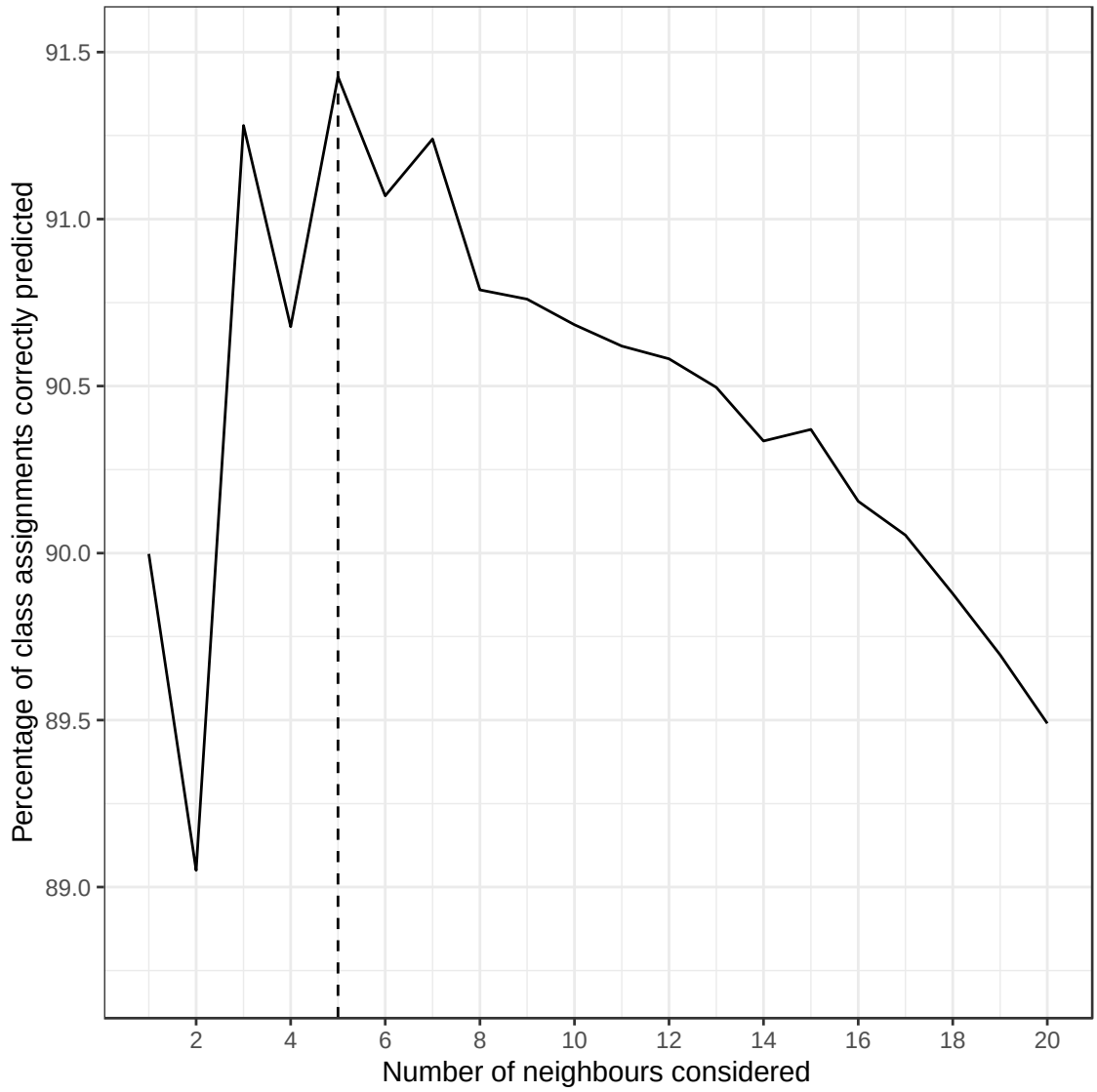


Figure C-3. Accuracy of K-Nearest Neighbours algorithm for predicting class membership in STRENGTH according to number of neighbours

Points indicate performance for each of 20 runs of the KNN algorithm for each number of neighbours using pseudorandomised seeds. The line indicates mean performance for each number of neighbours. The hatched vertical line indicates the configuration with the highest mean predictive accuracy.

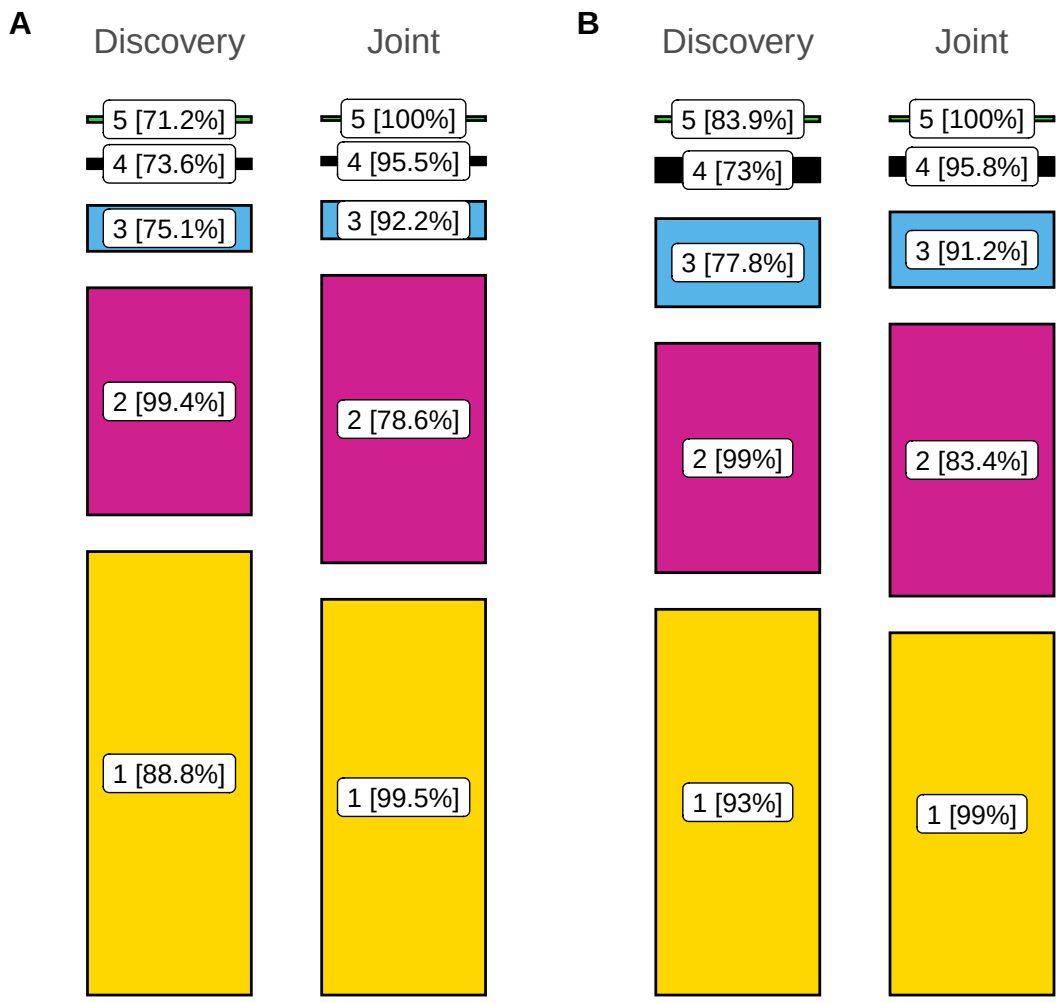


Figure C-4. Distribution of people across clusters of the 5-class models fitted to the discovery (Project MinE) and joint (Project MinE and STRENGTH) datasets

Project MinE dataset is shown in Panel A and STRENGTH is shown in Panel B. 'flows' on the figure indicate the movement of people between classes across the two models. Percentages shown indicate the proportion of people in a given class who remained in the equivalent class for the model fitted to the opposing dataset (e.g., for Panel A, 88.8% of people from Class 1 of the discovery dataset model remained in Class 1 of the joint dataset model). Plotting was performed using ggsankey (v0.0.99999) (Sjoberg, 2022).

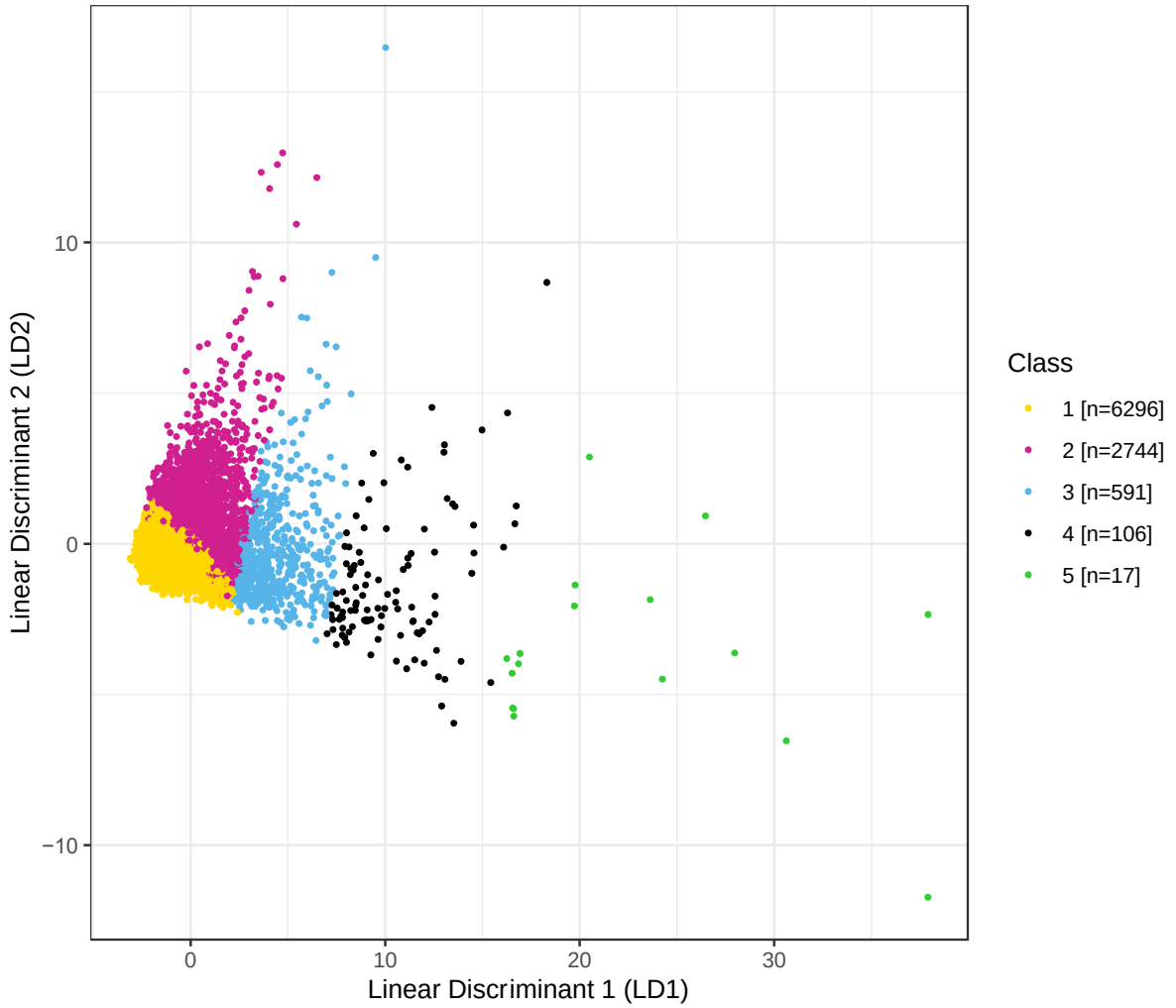


Figure C-5. Distribution of people and clusters across the first two linear discriminant analysis axes when restricting to people with non-censored disease duration

LD1 is highly correlated with diagnostic delay, while LD2 is associated primarily with disease duration (see Table C-6).

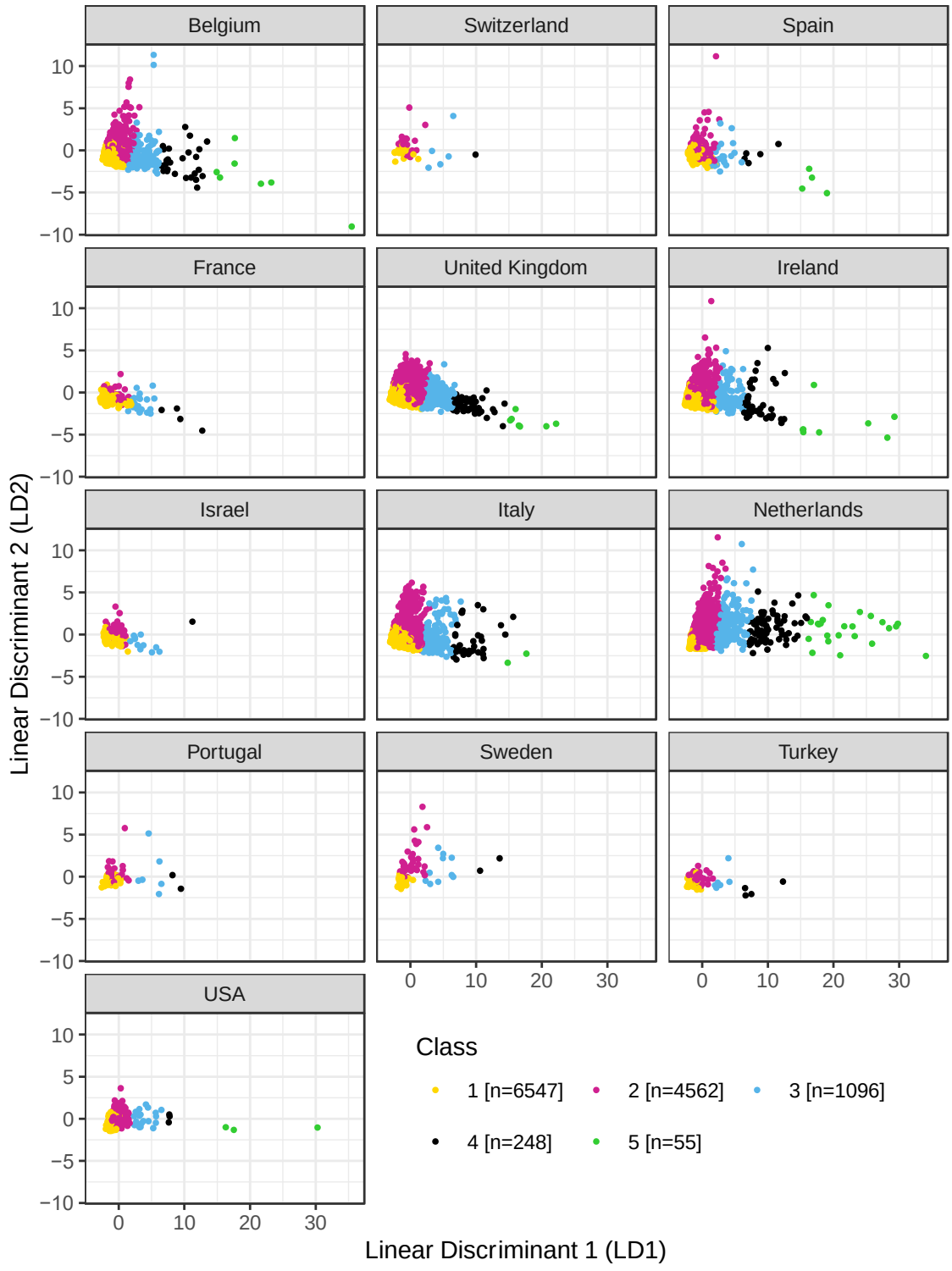


Figure C-6. Distribution of people and clusters across the first two discriminant analysis axes, stratified by country of origin

LD1 is highly correlated with diagnostic delay while LD2 is associated primarily with disease duration (see Table 6-3).

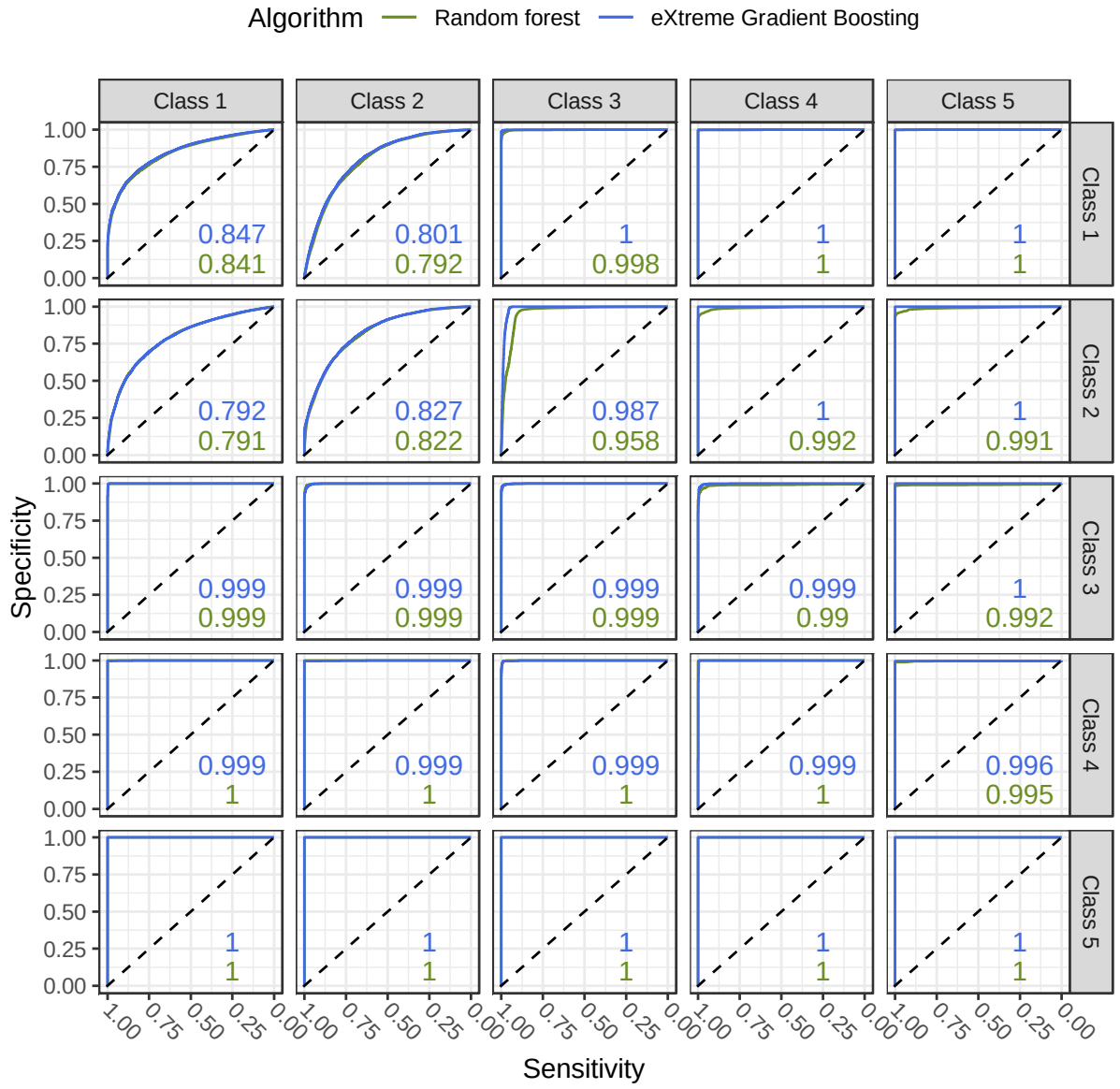


Figure C-7. Receiver operating characteristic curves for random forest and eXtreme Gradient Boosting algorithm predictions of class membership using only clinical data available around the time of diagnosis across all people with complete clinical data

Panels along the plot-diagonal (displayed with background shading) display ROCs for people being in that class versus all other classes. The upper triangle of panels present ROCs where the class represented in the panel row are considered controls and the column for the class is the 'case' group. Case-control coding is reversed for panels in the lower triangle. Colour denotes performance of different machine-learning algorithms, with numbers shown representing the area under the curve for each algorithm in that panel.

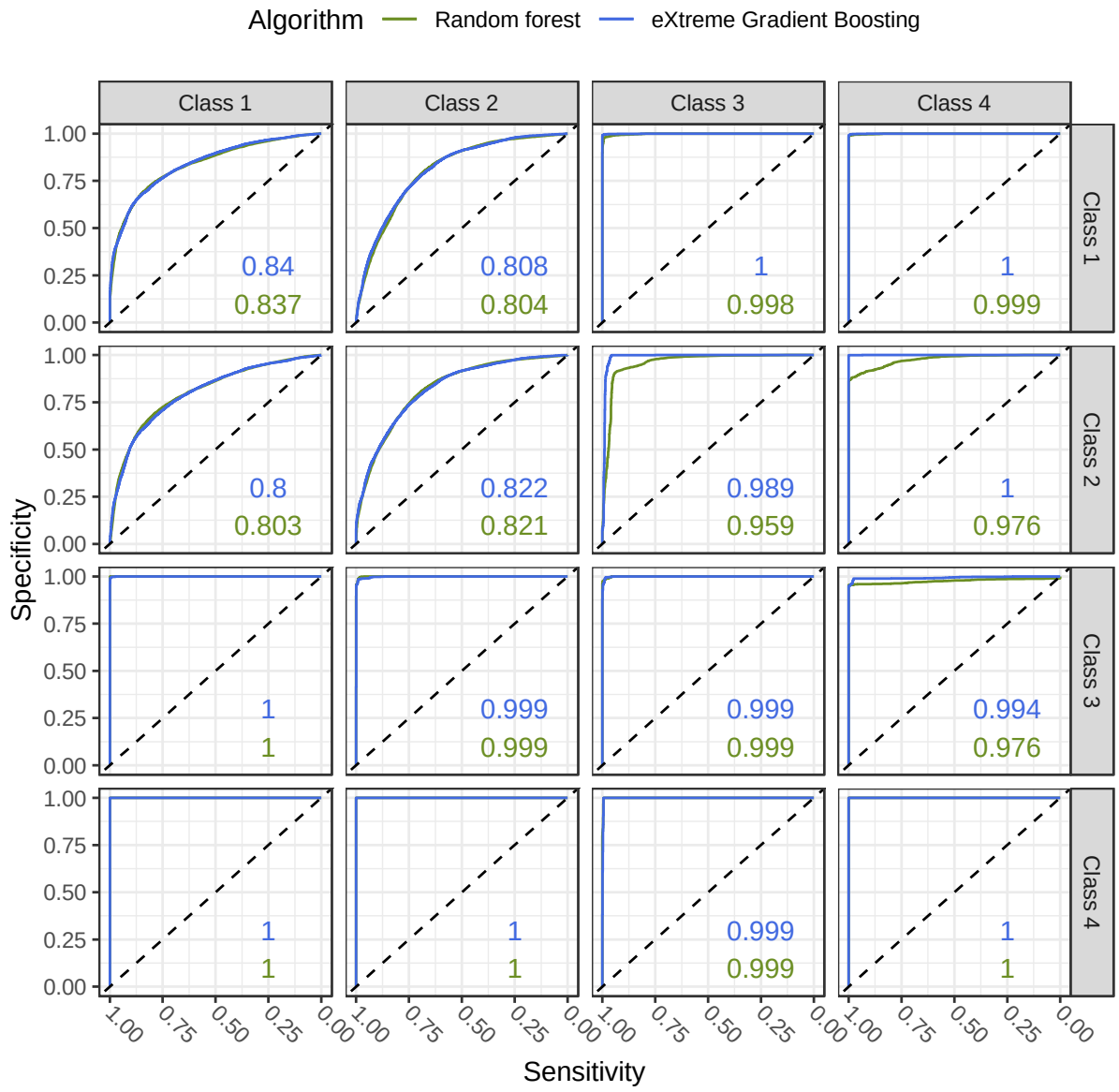


Figure C-8. Receiver operating characteristic curves for random forest and eXtreme Gradient Boosting algorithm predictions of class membership using clinical data available around the time of diagnosis and measures of genetic disease liability

Panels along the plot-diagonal (displayed with background shading) display ROCs for people being in that class versus all other classes. The upper triangle of panels present ROCs where the class represented in the panel row are considered controls and the column for the class is the 'case' group. Case-control coding is reversed for panels in the lower triangle. Colour denotes performance of different machine-learning algorithms, with numbers shown representing the area under the curve for each algorithm in that panel. Class 5 is excluded from this classification analysis owing to small sample size.

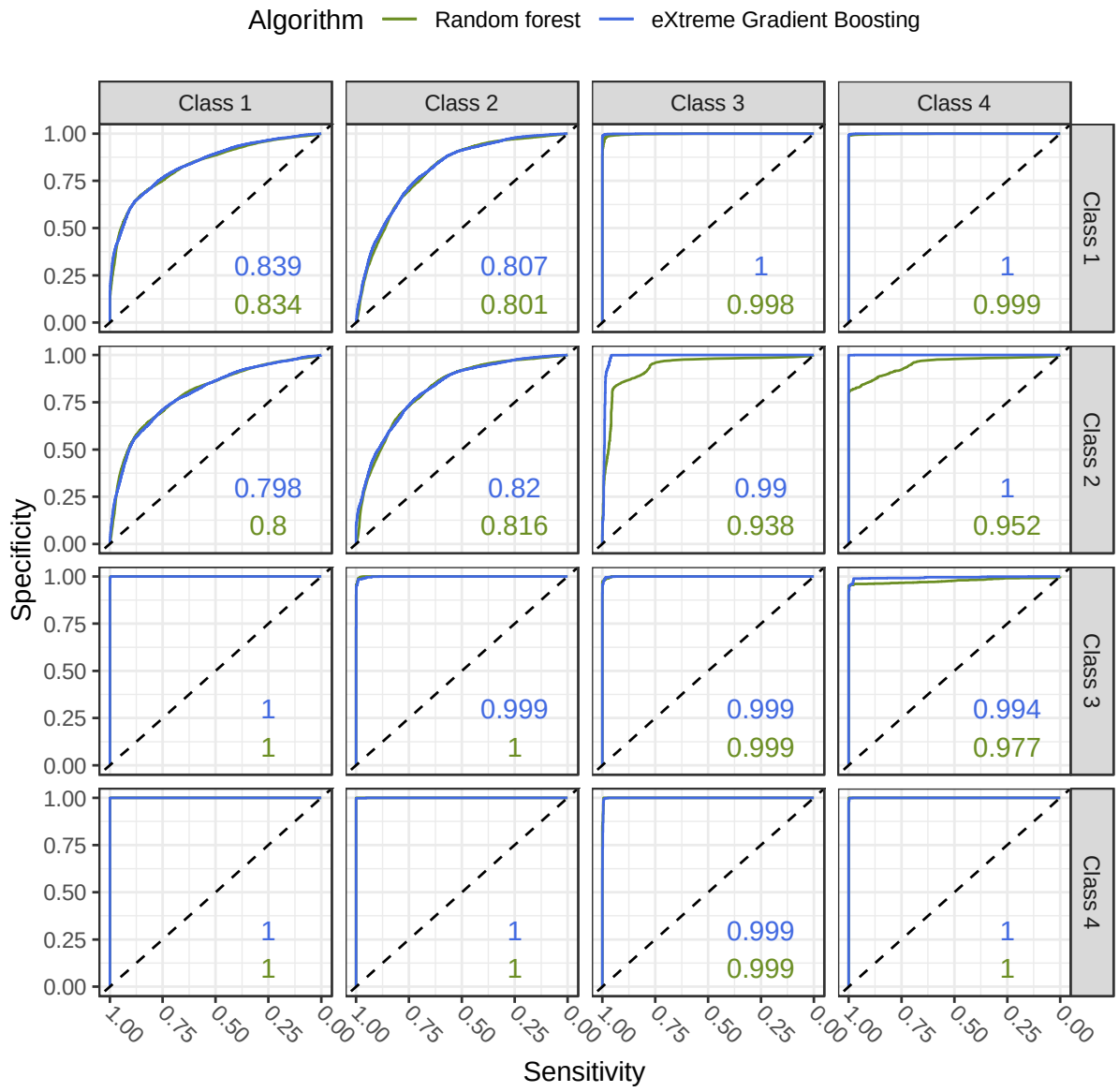


Figure C-9. Receiver operating characteristic curves random forest and eXtreme Gradient Boosting algorithm predictions of class membership using clinical data available around the time of diagnosis

This classification algorithm was trained upon the same features as the algorithm presented in Figure C-7, but was restricted to the same samples that were used in the algorithm presented in Figure C-8. Panels along the plot-diagonal (displayed with background shading) display ROCs for people being in that class versus all other classes. The upper triangle of panels present ROCs where the class represented in the panel row are considered controls and the column for the class is the 'case' group. Case-control coding is reversed for panels in the lower triangle. Colour denotes performance of different machine-learning algorithms, with numbers shown representing the area under the curve for each algorithm in that panel. Class 5 is excluded from this classification analysis owing to small sample size.

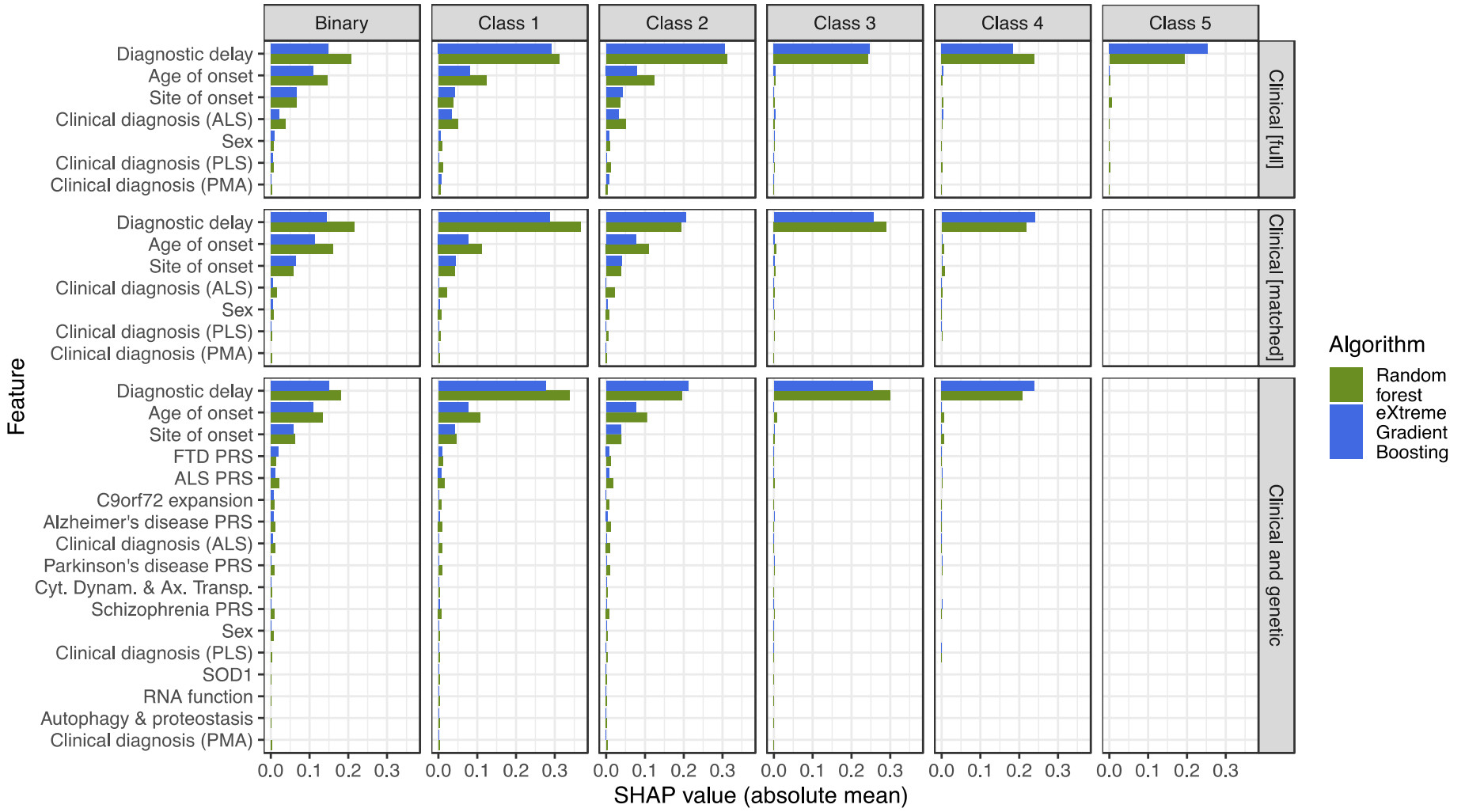


Figure C-10. SHapley Additive exPlanations (SHAP) of feature importance across trained classification algorithms

Panel rows stratify between algorithms using 'Clinical' features only, and with a combination of 'Clinical and genetic' features; the 'Clinical [full]' row is trained upon all samples with complete clinical data, while 'Clinical [matched]' uses the same features but is sample matched to the 'Clinical and genetic' row. Panel columns stratify feature importance for predictions of probability of each class individually; the 'Binary' column describes distinct algorithms trained to predict Classes 1 versus 2 only. Class 5 is only represented within Clinical [all] algorithms. FTD = frontotemporal dementia, ALS = amyotrophic lateral sclerosis, PRS = polygenic risk score. SHAP values are calculated for predicted class probabilities and therefore can range between 0 and 1.

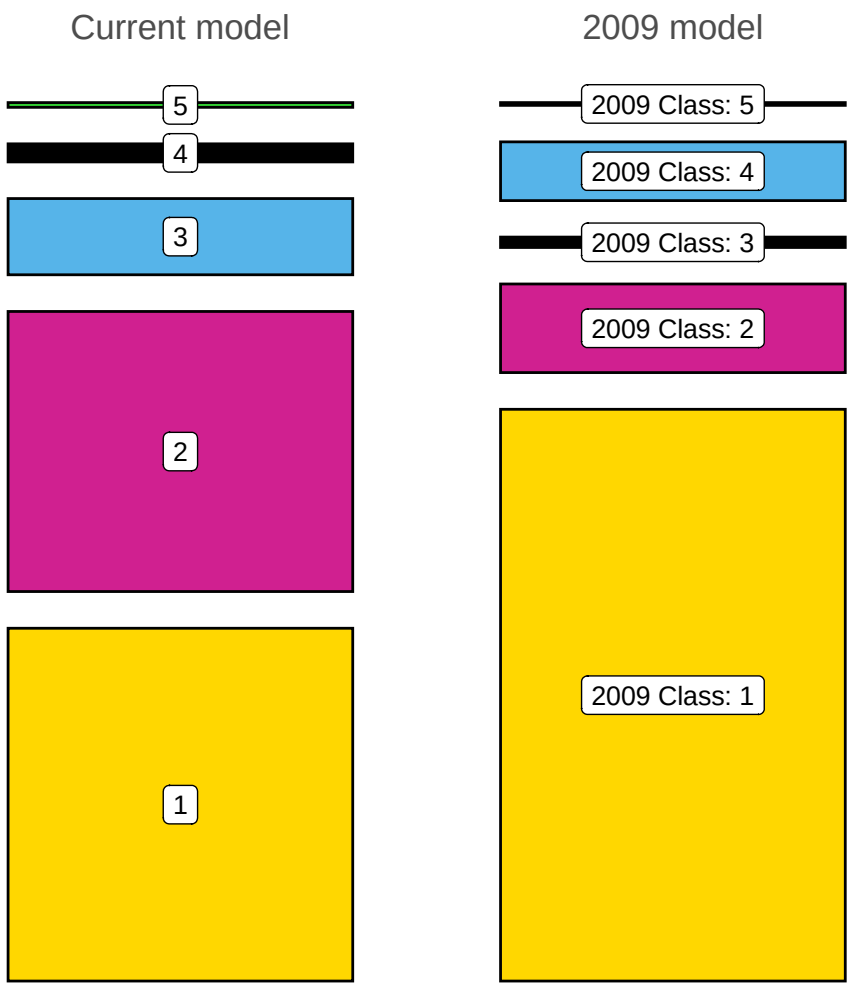


Figure C-11. Comparison of classes from the current latent class model of ALS with the model from a previous study

These data compare people from STRENGTH (N = 5,961). The 'current' model, shown on the left, is the accepted 5-class model trained upon the joint dataset. The '2009 model', shown on the right, is a 5-class model derived in clinical data a United Kingdom cohort of people with amyotrophic lateral sclerosis (Ganesalingam et al., 2009). Colouring of the 2009 model column is with respect to the current model class with the largest sample overlap.

Appendix C.3. Supplemental tables

Table C-1. Mean and standard deviation of diagnostic delay per country of origin across unique samples from Project MinE and STRENGTH

Diagnostic delay values entered into the latent class model and subsequent analyses were centred on the per-country mean and scaled by the per-country standard deviation. The values presented for 'total' diagnostic delay were derived by aggregating across the per-country statistics weighted by sample size; these were not used in any analyses.

Region (Iso2c code)	Number of samples with measured diagnostic delay	Mean diagnostic delay in years	Standard deviation
Belgium (BE)	1583	1.111	1.257
Switzerland (CH)	43	1.357	1.298
Spain (ES)	348	1.142	1.334
France (FR)	166	1.123	0.967
United Kingdom (GB)	3140	1.711	1.461
Ireland (IE)	1902	1.238	1.347
Israel (IL)	101	1.191	1.153
Italy (IT)	1182	0.935	0.779
Netherlands (NL)	3745	1.669	2.609
Portugal (PT)	59	1.695	1.830
Sweden (SE)	90	2.187	2.872
Turkey (TR)	74	1.387	1.584
United States of America (US)	338	2.035	2.782
Total	**12771**	**1.464**	**1.872**

Table C-2. Summary of number of variants identified in ALS-associated genes and assignment of genes to disease pathways according to evidence of role in gene products in pathway

Additional pathways relevant to ALS but without any variants assigned are: DNA repair (D), mitochondrial dysfunction (E), inflammation (F)

Assigned to Pathway	Gene	N with variants in gene	Supporting evidence for role in pathway	Other pathway involvement
A – Autophagy and proteostasis	*UBQLN2*	*11*	*(Al Sultan, Waller, Heath, & Kirby, 2016; Deng et al., 2011; Masrori & Van Damme, 2020)*	*B (Al Sultan et al., 2016)*
	CHMP2B	*23*	*(Al Sultan et al., 2016; Weishaupt et al., 2016)*	*A (Burk & Pasterkamp, 2019)* *C (Burk & Pasterkamp, 2019; Chia et al., 2018; Müller et al., 2018)*
	ERBB4	*62*	*(Al Sultan et al., 2016; Takahashi et al., 2013)*	*E (Müller et al., 2018)*
	SIGMAR1	*17*	*(Al Sultan et al., 2016; Fukunaga, Shinoda, & Tagashira, 2015)*	*E (Fukunaga et al., 2015)*
	CHCHD10	*22*	*(Chia et al., 2018; Nguyen et al., 2018)*	*E (Al Sultan et al., 2016; Chia et al., 2018; Nguyen et al., 2018)*
	DAO	*49*	*(Kondori et al., 2018)*	
	OPTN	*67*	*(Al Sultan et al., 2016; Masrori & Van Damme, 2020; Nguyen et al., 2018; Thakur et al., 2020; Weishaupt et al., 2016)*	*F (Chia et al., 2018)*
	SCFD1	*47*	*(Iacoangeli et al., 2021)*	
	SQSTM1 (p62)	*51*	*(Al Sultan et al., 2016; Masrori & Van Damme, 2020; Weishaupt et al., 2016)*	*F (Chia et al., 2018)*
	TBK1	*96*	*(Al Sultan et al., 2016; Freischmidt et al., 2015; Masrori & Van Damme, 2020; Nguyen et al., 2018)*	*F (Al Sultan et al., 2016; Chia et al., 2018; Müller et al., 2018; Nguyen et al., 2018)*
	UNC13A	*88*	*(A.-L. Brown et al., 2021)*	*C (Hardiman et al., 2017)*
	VAPB	*12*	*(Al Sultan et al., 2016; Nishimura et al., 2004)*	*C (Al Sultan et al., 2016; Morgan & Orrell, 2016; Müller et al.,*

				2018) *D (Nguyen et al., 2018)*
	VCP	*22*	*(Al Sultan et al., 2016; Burk & Pasterkamp, 2019; Johnson et al., 2010; Masrori & Van Damme, 2020; Weishaupt et al., 2016)*	
	VEGFA	*41*	*(Lambrechts et al., 2003; Thakur et al., 2020)*	
B – RNA function	*ANG*	*26*	*(Al Sultan et al., 2016; Greenway et al., 2006)*	*A (Al Sultan et al., 2016; Chia et al., 2018; Müller et al., 2018; Thakur et al., 2020)*
	ATXN2	*177*	*(Elden et al., 2010; Masrori & Van Damme, 2020)*	*C (Müller et al., 2018)*
	FUS	*52*	*(Al Sultan et al., 2016; Masrori & Van Damme, 2020)*	*A (Burk & Pasterkamp, 2019)* *D (Nguyen et al., 2018; Weishaupt et al., 2016)* *E (Al Sultan et al., 2016)*
	hnRNPA1	*10*	*(Kim et al., 2013; Masrori & Van Damme, 2020; Nguyen et al., 2018)*	
	MATR3	*46*	*(Al Sultan et al., 2016; Chia et al., 2018; Masrori & Van Damme, 2020)*	*A (R. H. Brown & Al-Chalabi, 2017)*
	SETX	*213*	*(Al Sultan et al., 2016; Yüce & West Stephen, 2013)*	
	TAF15	*62*	*(Al Sultan et al., 2016)*	
	TARDBP	*23*	*(Al Sultan et al., 2016; Kirby et al., 2010)*	*A (Burk & Pasterkamp, 2019)* *D (Nguyen et al., 2018)*
C – Cytoskeletal dynamics and axonal transport	*ALS2*	*104*	*(Castellanos-Montiel, Chaineau, & Durcan, 2020; Weishaupt et al., 2016)*	*D (Nguyen et al., 2018)*
	ANXA11	*67*	*(Nguyen et al., 2018; Smith et al., 2017)*	
	CFAP410 (C21ORF2)	*78*	*(Chia et al., 2018; Nguyen et al., 2018)*	*A (R. H. Brown & Al-Chalabi, 2017)* *D (Chia et al., 2018; Farg, Konopka, Soo, Ito, & Atkin, 2017;*

				Weishaupt et al., 2016), *E (Chia et al., 2018)*
	FIG4	*80*	*(Al Sultan et al., 2016; Burk & Pasterkamp, 2019)*	
	NEFH	*114*	*(Castellanos-Montiel et al., 2020; Marriott et al., 2022; Weishaupt et al., 2016)*	
	NEK1	*160*	*(Chia et al., 2018; Nguyen et al., 2018)*	*A (R. H. Brown & Al-Chalabi, 2017)* *D* (Chia et al., 2018; Farg et al., 2017; Weishaupt et al., 2016) *E (Chia et al., 2018; Nguyen et al., 2018)*
	ATXN1	*64*	*(Lattante et al., 2018; Tazelaar et al., 2020)*	
	DCTN1	*100*	*(Castellanos-Montiel et al., 2020; Weishaupt et al., 2016)*	
	MOBP	*6*	*(H. Chen, Kankel, Su, Han, & Ofengeim, 2018; Cirulli et al., 2015; The UniProt, 2021)*	
	PFN1	*6*	*(Castellanos-Montiel et al., 2020; Weishaupt et al., 2016)*	*A (Castellanos-Montiel et al., 2020)* *B (Castellanos-Montiel et al., 2020)*
	SPG11	*236*	*(Al Sultan et al., 2016; Burk & Pasterkamp, 2019; Orlacchio et al., 2010)*	*A (Burk & Pasterkamp, 2019)* *D (Weishaupt et al., 2016)*
	TUBA4A	*7*	*(Castellanos-Montiel et al., 2020; Weishaupt et al., 2016)*	
Not assigned to pathway	*SOD1*	*53*	*(Bunton-Stasyshyn et al., 2015)*	-
	C9orf72	*366*	*(Balendra & Isaacs, 2018)*	-

Table C-3. Comparison of latent class model solutions for the Project MinE and for the Joint datasets

The joint dataset is a combination of unique individuals from Project MinE and STRENGTH. AIC = Akaike information criterion, (a)BIC = (adjusted) Bayesian information criterion

Dataset		1	2	3	4	5	6	7	8	9
Project MinE (discovery sample)	Loglikelihood	-50036.833	-48440.246	-47557.668	-46908.637	-46502.638	-	-	-	-
	AIC	100089.667	96912.492	95163.337	93881.274	93085.276	-	-	-	-
	BIC	100143.932	97021.022	95326.131	94098.333	93356.6	-	-	-	-
	aBIC	100118.51	96970.178	95249.865	93996.645	93229.49	-	-	-	-
	Entropy	1	0.937	0.912	0.787	0.791	-	-	-	-
	N per class	6523	243: 6280	415: 109: 5999	318: 81: 2096: 4028	3952: 2023: 409: 87: 52	-	-	-	-
	Lowest average class probability	NA	0.877	0.802	0.843	0.818	-	-	-	-
Joint	Loglikelihood	-100154.4	-96218.178	-94005.047	-92337.947	-91200.445	-90418.286	-90000.826	-89507.894	-89228.181
	AIC	200324.807	192468.356	188058.095	184739.894	182480.889	180932.572	180113.652	179143.788	178600.362
	BIC	200385.38	192589.502	188239.814	184982.187	182783.755	181296.011	180537.664	179628.373	179145.52
	aBIC	200359.957	192538.655	188163.544	184880.493	182656.638	181143.472	180359.701	179424.987	178916.71
	Entropy	1	0.954	0.933	0.82	0.836	0.841	0.79	0.792	0.752
	N per class	14352	524: 13828	136: 1100: 13116	5400: 108: 717: 8127	7401: 5470: 1138: 259: 84	1260: 329: 47: 88: 5346: 7282	1395: 326: 6601: 1263: 23: 92: 4652	404: 1408: 55: 16: 1471: 139: 6172: 4687	1254: 4269: 3032: 1387: 55: 3798: 138: 404: 15
	Lowest average class probability	NA	0.926	0.894	0.884	0.879	0.828	0.748	0.738	0.711

Table C-4. Five-class latent class model solutions when restricting to samples with recorded diagnostic delay and disease duration

Numbers presented in bold refer to average probability of belonging to the assigned class. †Total percentage of people in the equivalent classes across the missingness vs full models were: 99.2% for the discovery model; 99.6% for the joint dataset model.

Dataset [General statistics]	Assigned class	Percentage of class in equivalent class of full-dataset model †	N in class (% of dataset) based on		Average posterior probability of belonging to class				
			posterior probabilities	most likely class membership	1	2	3	4	5
Discovery (Project MinE) [N = 5377; entropy = 0.850]	1	99.0	3156.30 (0.587)	3319 (0.617)	**0.908**	0.09	0.002	0	0
	2	99.7	1699.04 (0.316)	1551 (0.288)	0.089	**0.883**	0.028	0	0
	3	99.3	422.66 (0.079)	408 (0.076)	0.008	0.075	**0.909**	0.007	0
	4	97.6	82.99 (0.015)	83 (0.015)	0	0	0.037	**0.963**	0
	5	1	16.01 (0.003)	16 (0.003)	0	0	0	0	**1**
Joint [N = 12771; entropy = 0.881]	1	99.4	6348.46 (0.497)	6624 (0.519)	**0.923**	0.076	0.001	0	0
	2	99.9	4949.39 (0.388)	4698 (0.368)	0.049	**0.926**	0.025	0	0
	3	99.9	1153.11 (0.090)	1132 (0.089)	0.003	0.084	**0.902**	0.011	0
	4	99.6	259.75 (0.020)	257 (0.020)	0	0	0.038	**0.961**	0.001
	5	1	60.30 (0.005)	60 (0.005)	0	0	0	0	**1**

Table C-5. Validation of the Project MinE dataset 5-class model solution using independent data from STRENGTH within a K-nearest neighbours (KNN) classification algorithm

Class assignments were determined by MPlus, according to the parameters of the 5-class model fitted to the Project MinE dataset. KNN was trained to predict class membership using the clinical features of LCA. The training dataset for KNN was complete cases from the Project MinE sample (N=5320), and prediction was for complete cases (N=7188) from STRENGTH. Area under the receiver operating characteristic curve (AUC) for the KNN model in prediction of each assigned class vs any other class was consistently high. Numbers presented in bold denote people predictions which align with the assigned class. The algorithm was applied in R using a fixed seed, and 5 neighbours were considered (see Figure C-3).

Class assigned by latent class model	Number of people predicted to belong to class by KNN					AUC for KNN prediction of assigned class vs other
	1	2	3	4	5	
1	**3646**	110	5	0	0	0.938
2	300	**1907**	19	0	0	0.906
3	17	111	**770**	11	0	0.919
4	0	0	32	**209**	0	0.932
5	0	0	0	11	**40**	0.892

Table C-6. Results of linear discriminant analysis after restricting to people with non-censored disease duration

Proportion of trace describes the proportion of the separation between classes accounted for by each linear discriminant (LD) axis. Pooled within-group correlations greater than 0.5 are presented in bold and are considered variables associated with a given LD. Reference groups for categorical variables are: 'not-bulbar' for site of onset, 'male' for sex, 'ALS' for clinical diagnosis. Figure C-5 visualises the distribution of people and classes across the first two LD axes.

Statistic	Variable	LD1	LD2	LD3	LD4
Eigenvalue	-	95.48	31.93	3.80	2.16
Proportion of trace	-	0.898	0.100	1.42×10^{-3}	4.59×10^{-4}
Pooled within-group correlation	Diagnostic delay	**0.923**	-0.375	-0.082	-0.001
	Age of onset	-0.067	-0.342	-0.353	**-0.580**
	Disease duration	**0.507**	**0.800**	0.089	-0.084
	Site of onset (bulbar)	-0.073	-0.264	0.475	0.177
	Sex (female)	-0.012	-0.081	-0.208	-0.214
	Clinical diagnosis (PLS)	0.099	0.013	**0.699**	**-0.646**
	Clinical diagnosis (PMA)	0.030	0.123	-0.268	-0.386

Table C-7. Results of multinomial regression analysis of all people with no missingness across predictors, including people with censored disease duration

Class 1 is used as the outcome variable reference category because this is the largest subgroup. Continuous variables were standardised to have a mean of 0 and standard deviation of 1; diagnostic delay is standardised per-country of origin. Categorical variable reference categories are: 'not-bulbar' for site of onset; "ALS" for clinical diagnosis. Sex at birth (male or female) was entered into the regression model but removed in stepwise feature selection. ALS = amyotrophic lateral sclerosis, PLS = primary lateral sclerosis, PMA = progressive muscular atrophy.

Class	Predictor	Standardised beta (95% Confidence Interval)	Standard error	Z-score	p-value
2	Diagnostic delay	1.18 (0.98, 1.4)	0.104	11	5.05×10^{-30}
	Age of onset	-0.743 (-0.83, -0.66)	0.0442	-17	2.83×10^{-63}
	Disease duration	9.72 (9.3, 10)	0.225	43	0
	Site of onset (bulbar)	-1.04 (-1.2, -0.88)	0.0849	-12	1.14×10^{-34}
	Clinical diagnosis (PLS)	2.5 (1.8, 3.2)	0.344	7.3	4.04×10^{-13}
	Clinical diagnosis (PMA)	1.44 (1.1, 1.8)	0.179	8.1	7.52×10^{-16}
3	Diagnostic delay	41.4 (34, 49)	3.96	10	1.50×10^{-25}
	Age of onset	-0.0527 (-0.5, 0.39)	0.226	-0.23	0.815
	Disease duration	11.8 (11, 13)	0.373	32	1.05×10^{-221}
	Site of onset (bulbar)	-2.99 (-4, -2)	0.531	-5.6	1.72×10^{-8}
	Clinical diagnosis (PLS)	9.23 (7.2, 11)	1.04	8.9	6.91×10^{-19}
	Clinical diagnosis (PMA)	2.86 (1.4, 4.3)	0.757	3.8	0.000162
4	Diagnostic delay	87.6 (61, 110)	13.4	6.5	6.25×10^{-11}
	Age of onset	-1.64 (-3.1, -0.23)	0.72	-2.3	0.0227
	Disease duration	12.9 (12, 14)	0.629	21	1.62×10^{-93}
	Site of onset (bulbar)	-5.92 (-13, 0.69)	3.37	-1.8	0.0791
	Clinical diagnosis (PLS)	12.3 (8.3, 16)	2.02	6.1	1.13×10^{-09}
	Clinical diagnosis (PMA)	4.6 (-0.064, 9.3)	2.38	1.9	0.0532
5	Diagnostic delay	111 (94, 130)	8.33	13	3.38×10^{-40}
	Age of onset	-2.87 (-6.5, 0.74)	1.84	-1.6	0.119
	Disease duration	12.7 (10, 15)	1.31	9.7	2.95×10^{-22}
	Site of onset (bulbar)	-11.4 (-51, 28)	20.3	-0.56	0.573
	Clinical diagnosis (PLS)	9.24 (-2.7, 21)	6.1	1.5	0.13
	Clinical diagnosis (PMA)	0.733 (-16, 17)	8.39	0.087	0.93

Table C-8. Results of multinomial regression analysis of all people with no missingness across predictors, restricted to people with non-censored disease duration

Class 1 is used as the outcome variable reference category because this is the largest subgroup. Continuous variables were standardised to have a mean of 0 and standard deviation of 1; diagnostic delay is standardised per-country of origin. Categorical variable reference categories are: 'not-bulbar' for site of onset; "ALS" for clinical diagnosis, 'male' for sex. No features were dropped from the model within stepwise feature selection. No person with PMA remained in this class for this data subsample. ALS = amyotrophic lateral sclerosis, PLS = primary lateral sclerosis, PMA = progressive muscular atrophy.*

Class	Predictor	Standardised beta (95% Confidence Interval)	Standard error	Z-score	p-value
2	Diagnostic delay	15.6 (12, 19)	1.59	9.8	8.06×10^{-23}
	Age of onset	-6.8 (-8.2, -5.5)	0.689	-9.9	5.43×10^{-23}
	Disease duration	213 (170, 250)	20.2	11	3.33×10^{-26}
	Site of onset (bulbar)	-10.1 (-12, -8)	1.07	-9.5	3.12×10^{-21}
	Sex (female)	-2.09 (-3, -1.2)	0.464	-4.5	6.48×10^{-6}
	Clinical diagnosis (PLS)	41.5 (33, 50)	4.23	9.8	1.06×10^{-22}
	Clinical diagnosis (PMA)	15.8 (12, 19)	1.69	9.3	1.05×10^{-20}
3	Diagnostic delay	158 (120, 190)	18.6	8.5	2.14×10^{-17}
	Age of onset	-5.74 (-7.4, -4.1)	0.85	-6.8	1.45×10^{-11}
	Disease duration	219 (180, 260)	20.6	11	2.06×10^{-26}
	Site of onset (bulbar)	-16.9 (-21, -13)	2.07	-8.2	3.64×10^{-16}
	Sex (female)	-0.0116 (-2, 1.9)	0.99	-0.012	0.991
	Clinical diagnosis (PLS)	66.6 (50, 84)	8.69	7.7	1.83×10^{-14}
	Clinical diagnosis (PMA)	19.5 (14, 25)	2.59	7.5	6.15×10^{-14}
4	Diagnostic delay	193 (150, 240)	22	8.8	1.82×10^{-18}
	Age of onset	-6.95 (-9.6, -4.3)	1.36	-5.1	3.43×10^{-7}
	Disease duration	222 (180, 260)	20.8	11	9.74×10^{-27}
	Site of onset (bulbar)	-19.8 (-27, -13)	3.6	-5.5	4.03×10^{-8}
	Sex (female)	0.768 (-3.1, 4.6)	1.95	0.39	0.694
	Clinical diagnosis (PLS)	63.1 (41, 85)	11.3	5.6	2.10×10^{-08}
	Clinical diagnosis (PMA)	26.3 (-21, 74)	24.1	1.1	0.276
5	Diagnostic delay	200 (160, 240)	22.3	9	2.76×10^{-19}
	Age of onset	-7.23 (-15, 0.93)	4.16	-1.7	0.0826
	Disease duration	222 (180, 260)	20.8	11	1.92×10^{-26}
	Site of onset (bulbar)	-22.1 (-36, -8.4)	7	-3.2	1.56×10^{-03}
	Sex (female)	0.696 (-11, 12)	5.79	0.12	0.904
	Clinical diagnosis (PLS)	52.6 (24, 81)	14.4	3.6	2.67×10^{-04}
	Clinical diagnosis (PMA)*	-23.02 (-)	-	-	-

Table C-9. Summary of cox proportional-hazards model predicting disease duration from onset until death or censoring using Class and all other clinical features from LCA

Continuous predictors (age of onset and diagnostic delay) were standardised to have a mean of 0 and standard deviation of 1; Diagnostic delay was standardised by country of origin (see Table C-1; Figure C-2). Hazard ratios greater than 1 indicate association between the variable and shorter disease duration.

Variable	Factor Level	Number sampled	Hazard Ratio [95% Confidence Interval]	Inverse hazard ratio	Z-score	P-value
Class	1	6547	-	-	-	-
	2	4562	0.010 [0.009, 0.012]	99.3	-68	$<2x10^{-16}$
	3	1096	0.011 [0.010, 0.014]	87.8	-48.8	$<2x10^{-16}$
	4	248	0.016 [0.011, 0.021]	64.6	-26.1	$<2x10^{-16}$
	5	55	0.045 [0.024, 0.086]	22	-9.46	$<2x10^{-16}$
Site of onset	Other	8736	-	-	-	-
	Bulbar	3772	0.940 [0.899, 0.983]	1.06	-2.7	0.00699
Sex	Male	7412	-	-	-	-
	Female	5096	0.925 [0.888, 0.964]	1.08	-3.68	0.000237
Clinical diagnosis	ALS	11406	-	-	-	-
	PLS	473	0.327 [0.274, 0.391]	3.05	-12.4	$<2x10^{-16}$
	PMA	629	0.824 [0.742, 0.915]	1.21	-3.63	0.000281
Age of onset in years	-	12508	1.21 [1.18, 1.24]	0.828	16	$<2x10^{-16}$
Diagnostic delay	-	12508	0.644 [0.610, 0.681]	1.55	-15.8	$<2x10^{-16}$

Table C-10. Results of differential expression analysis between BrainBank samples in Class 1 and 2

log2FoldChange results are for Class 2 (n = 18) relative to class 1 (n = 70). Adjusted P-values (padj) were adjusted via independent hypothesis weight

Available in Excel workbook: Thesis_TSpargo_ExcelTables.xlsx

Table C-11. gProfiler gene enrichment results for the top 500 differentially expressed genes between BrainBank samples in Class 1 and 2

Sample size in class: 1 = 70 and 2 = 18. See Table C-10 for differential expression analysis results.

Available in Excel workbook: Thesis_TSpargo_ExcelTables.xlsx

Table C-12. Optimum hyperparameter tuning settings and overall AUC for all machine-learning algorithms trained

Predictions were made based on 'Clinical' features only or a combination of 'Clinical and genetic' features; the 'Clinical [full]' columns are trained based on samples with complete clinical data, while 'Clinical [matched]' uses the same features but is sample matched to the 'Clinical and genetic' column In binary classifications, data were further subsampled to only people assigned to classes 1 and 2. AUC values shown in bold indicate the best performing model for a given dataset configuration. Note that the Clinical [full] model describes a multiclass model predicting 5 classes, while the other multiclass models exclude class 5 owing to small sample size. Therefore, these AUC values cannot directly be compared for the multiclass models between Clinical [Full] and with Clinical and genetic or Clinical [matched] datasets. Refer instead to the comparisons shown across Figure C-7, Figure C-8, and Figure C-9. †calculated within the R caret package multiClassSummary function for multiclass and twoClassSummary for binary objectives.

		Algorithm objective					
		Multiclass			**Binary (Classes 1 and 2)**		
Dataset		Clinical [Full]	Clinical and genetic	Clinical [matched]	Clinical [Full]	Clinical and genetic	Clinical [matched]
eXtreme Gradient Boosting tuning parameters	nrounds	172	65	62	769	234	354
	max_depth	4	6	6	3	3	2
	eta	0.18	0.1	0.16	0.02	0.02	0.02
	gamma	0	0.1	0.1	0	0.2	0.1
	colsample_bytree	1	0.7	0.7	0.7	0.8	0.65
	min_child_weight	1	4	4	6	9	1
	subsample	1	0.85	0.8	0.95	0.95	0.9
Random forest tuning parameters	ntree	1001	1001	1001	1001	1001	1001
	nodesize	30	37	32	38	39	40
	mtry	4	9	4	3	5	3
Overall AUC† for prediction across all included classes	eXtreme Gradient Boosting	**0.935**	**0.916**	**0.915**	**0.803**	**0.813**	**0.812**
	Random forest	0.932	0.915	0.913	0.794	0.807	0.805

Appendix D. Chapter 7 supplementary materials

Appendix D.1. Supplemental figures

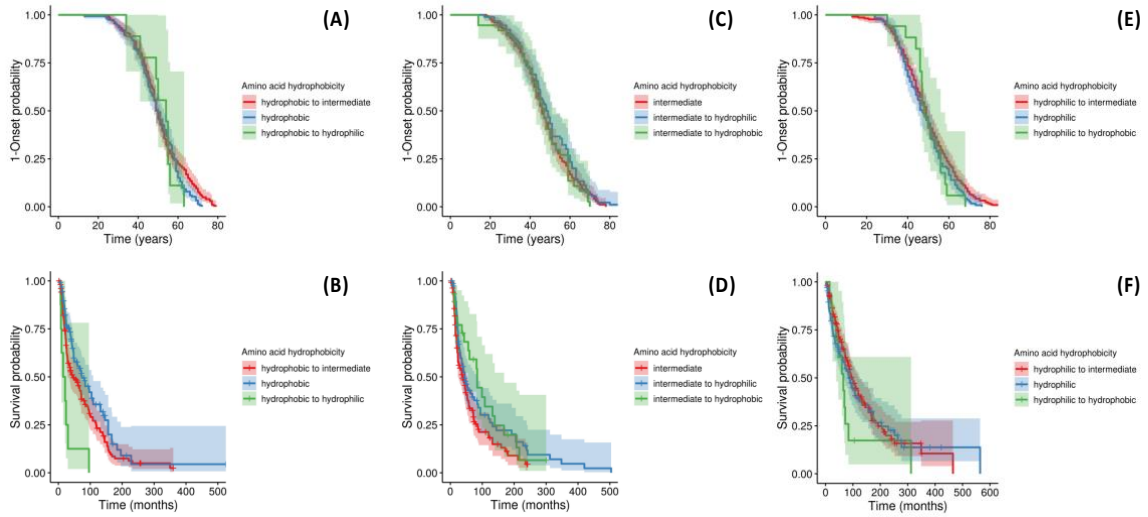


Figure D-1. Kaplan-Meier survival curves for age of onset and disease duration analyses of trends associated with wild type and variant amino acid hydrophobicity

The panels stratify according to the wild type SOD1 hydrophobicity group for the residue containing the variant: ***panels A-B*** *show analysis in hydrophobic residues,* ***panels C-D*** *for intermediate residues, and* ***panels E-F*** *for hydrophilic residues.*

Appendix D.2. Supplemental tables

Table D-1. Sample composition by variant including hydrophobicity group assignments

SOD1 protein variant	Hydrophobicity group	Number of records		
		total	with age of onset	with disease duration
K4E	hydrophilic	11	10	11
A5S	intermediate	4	3	3
A5T	intermediate	18	18	14
A5V	intermediate to hydrophobic (excluded from analysis)	312	298	260
C7F	intermediate to hydrophobic	10	10	7
C7G	intermediate	5	5	5
C7S	intermediate	12	12	10
C7W	intermediate	4	4	4
C7Y	intermediate	1	1	1
V8E	hydrophobic to hydrophilic	1	1	1

L9Q	hydrophobic to hydrophilic	7	7	6
L9V	hydrophobic	2	2	0
G11A	intermediate	1	1	1
G11R	intermediate to hydrophilic	2	2	2
G11V	intermediate to hydrophobic	2	2	2
D12Y	hydrophilic to intermediate	3	3	3
G13D	intermediate to hydrophilic	1	1	0
G13R	intermediate to hydrophilic	3	3	2
V15G	hydrophobic to intermediate	5	5	3
V15M	hydrophobic	12	12	12
G17A	intermediate	7	6	5
G17C	intermediate	2	2	2
G17S	intermediate	2	2	1
N20S	hydrophilic to intermediate	11	11	10
F21C	hydrophobic to intermediate	10	10	8
F21V	hydrophobic	1	1	1
E22G	hydrophilic to intermediate	21	21	17
Q23L	hydrophilic to hydrophobic	12	12	11
Q23R	hydrophilic	1	1	0
V32A	hydrophobic to intermediate	3	3	2
W33G	intermediate	1	1	1
G38R	intermediate to hydrophilic	10	10	6
G38V	intermediate to hydrophobic	3	3	3
L39P	hydrophobic to intermediate	1	1	1
L39R	hydrophobic to hydrophilic	1	1	1
L39V	hydrophobic	5	5	4
E41D	hydrophilic	1	1	0
E41G	hydrophilic to intermediate	6	6	5
G42D	intermediate to hydrophilic	25	21	17
G42S	intermediate	21	21	14
H44R	hydrophilic	17	16	15
F46C	hydrophobic to intermediate	2	2	1
F46S	hydrophobic to intermediate	2	2	1
H47R	hydrophilic	19	19	18
V48A	hydrophobic to intermediate	2	2	1
V48F	hydrophobic	1	1	1
H49Q	hydrophilic	2	2	2
H49R	hydrophilic	6	5	6
E50K	hydrophilic	8	8	7
G62R	intermediate to hydrophilic	1	1	1
F65L	hydrophobic	3	3	2
N66S	hydrophilic to intermediate	11	11	4

P67A	intermediate	1	1	1
P67R	intermediate to hydrophilic	1	1	0
P67S	intermediate	3	3	2
L68P	hydrophobic to intermediate	2	2	1
S69P	intermediate	1	1	1
H72Y	hydrophilic to intermediate	2	2	0
G73C	intermediate	1	1	1
G73D	intermediate to hydrophilic	1	1	1
G73S	intermediate	2	2	2
P75S	intermediate	1	1	1
D77V	hydrophilic to hydrophobic	2	2	1
D77Y	hydrophilic to intermediate	5	5	4
H81R	hydrophilic	1	1	1
D84V	hydrophilic to hydrophobic	1	1	1
L85F	hydrophobic	12	12	10
L85V	hydrophobic	4	2	4
G86C	intermediate	1	0	0
G86R	intermediate to hydrophilic	18	17	17
G86S	intermediate	4	3	3
N87K	hydrophilic	1	1	1
N87S	hydrophilic to intermediate	16	14	13
V88A	hydrophobic to intermediate	1	1	1
V88M	hydrophobic	3	3	3
A90V	intermediate to hydrophobic	15	15	11
D91A	hydrophilic to intermediate	83	79	61
D91N	hydrophilic	1	1	1
D91V	hydrophilic to hydrophobic	1	1	1
G94A	intermediate	27	27	26
G94C	intermediate	14	14	9
G94D	intermediate to hydrophilic	15	15	14
G94R	intermediate to hydrophilic	1	1	1
G94S	intermediate	2	2	0
G94V	intermediate to hydrophobic	4	4	2
V95G	hydrophobic to intermediate	1	1	1
A96G	intermediate	1	1	1
A96T	intermediate	3	3	3
V98M	hydrophobic	1	1	1
I100V	hydrophobic	1	1	0
E101G	hydrophilic to intermediate	47	42	32
E101K	hydrophilic	23	23	20
D102G	hydrophilic to intermediate	3	3	3
D102H	hydrophilic	2	2	2

D102N	hydrophilic	5	5	5
D102Y	hydrophilic to intermediate	1	1	1
I105F	hydrophobic	4	4	3
S106L	intermediate to hydrophobic	2	1	2
L107F	hydrophobic	5	5	4
L107P	hydrophobic to intermediate	1	1	1
L107V	hydrophobic	14	14	13
G109R	intermediate to hydrophilic	1	1	1
G109V	intermediate to hydrophobic	2	2	2
D110Y	hydrophilic to intermediate	4	4	4
C112Y	intermediate	4	4	3
I113M	hydrophobic	2	2	2
I113T	hydrophobic to intermediate	3	3	3
I114F	hydrophobic	2	2	2
I114T	hydrophobic to intermediate	120	108	86
G115A	intermediate	1	1	1
R116C	hydrophilic to intermediate	1	1	1
R116G	hydrophilic to intermediate	1	1	1
L118V	hydrophobic	9	9	8
V119L	hydrophobic	4	4	4
V119M	hydrophobic	1	1	1
V120F	hydrophobic	1	1	1
V120L	hydrophobic	3	3	1
H121Q	hydrophilic	3	3	3
E122G	hydrophilic to intermediate	5	5	4
D125G	hydrophilic to intermediate	3	3	3
D126A	hydrophilic to intermediate	1	1	1
D126H	hydrophilic	3	3	1
D126N	hydrophilic	1	1	1
L127S	hydrophobic to intermediate	9	9	9
L127X	NA	2	2	1
G128R	intermediate to hydrophilic	1	1	1
E133G	hydrophilic to intermediate	1	0	1
E133X	NA	1	1	1
E134A	hydrophilic to intermediate	2	2	2
E134K	hydrophilic	2	2	2
E134V	hydrophilic to hydrophobic	1	1	1
S135G	intermediate	1	1	1
S135N	intermediate to hydrophilic	3	3	3
S135T	intermediate	1	1	1
K137X	NA	1	1	1
T138A	intermediate	7	7	4

T138R	intermediate to hydrophilic	2	2	2
G139E	intermediate to hydrophilic	1	1	1
N140D	hydrophilic	4	4	4
N140H	hydrophilic	4	4	3
N140K	hydrophilic	3	3	3
A141G	intermediate	3	3	0
G142A	intermediate	6	6	4
G142E	intermediate to hydrophilic	3	3	2
G142X	NA	2	2	2
L145F	hydrophobic	69	62	48
L145S	hydrophobic to intermediate	7	7	4
A146G	intermediate	3	3	1
A146T	intermediate	6	6	5
C147R	intermediate to hydrophilic	3	3	2
C147X	NA	1	1	1
G148A	intermediate	1	1	0
G148C	intermediate	1	1	1
G148D	intermediate to hydrophilic	4	3	3
G148S	intermediate	2	2	1
V149A	hydrophobic to intermediate	2	2	2
V149G	hydrophobic to intermediate	43	36	25
I150T	hydrophobic to intermediate	5	5	4
I150V	hydrophobic	2	2	2
I152T	hydrophobic to intermediate	2	2	0
A153P	intermediate	1	1	1
A153T	intermediate	1	1	1

Table D-2. Cox Proportional-Hazards survival analysis for age of onset across all hydrophobicity groups

Comparisons between hydrophobicity groups with p<0.05 compared to the reference group are in bold. CI = confidence interval

Variable	Factor level	N	Hazard Ratio [95% CI]	P-value
Hydrophobicity group	Intermediate	173	-	-
	Hydrophilic to intermediate	215	0.829 [0.666, 1.03]	0.0922
	Hydrophilic	115	1.03 [0.814, 1.31]	0.803
	Hydrophilic to hydrophobic	17	1.04 [0.737, 1.46]	0.831
	Hydrophobic	152	0.957 [0.782, 1.17]	0.675
	Hydrophobic to hydrophilic	9	0.958 [0.656, 1.4]	0.823
	Hydrophobic to intermediate	201	0.81 [0.657, 0.999]	**0.0493**
	Intermediate to hydrophilic	90	0.841 [0.638, 1.11]	0.221
	Intermediate to hydrophobic	37	1.1 [0.778, 1.55]	0.592
Sex	Male	524	-	-
	Female	485	0.947 [0.837, 1.07]	0.39

Table D-3. Cox Proportional-Hazards survival analysis for age of onset across all hydrophobicity groups

Comparisons between hydrophobicity groups with p<0.05 compared to the reference group are in bold. CI = confidence interval

Variable	Factor level	N	Hazard Ratio [95% CI]	P-value
Hydrophobicity group	Intermediate	134	-	-
	Hydrophilic to intermediate	166	0.403 [0.3, 0.542]	**1.65×10^{-9}**
	Hydrophilic	104	0.471 [0.335, 0.663]	**1.56×10^{-5}**
	Hydrophilic to hydrophobic	15	0.65 [0.363, 1.16]	0.147
	Hydrophobic	124	0.548 [0.399, 0.752]	**0.000197**
	Hydrophobic to hydrophilic	8	2.97 [1.65, 5.36]	**0.000282**
	Hydrophobic to intermediate	153	0.733 [0.549, 0.98]	**0.0362**
	Intermediate to hydrophilic	76	0.697 [0.504, 0.963]	**0.0287**
	Intermediate to hydrophobic	28	0.567 [0.346, 0.927]	**0.0236**
Sex	Male	414	-	-
	Female	394	0.7 [0.592, 0.828]	3.13×10^{-5}
Age of onset	-	808	1.03 [1.02, 1.04]	1.94×10^{-14}

Table D-4 SOD1 variants reported in gnomAD v2.1.1 that have a recorded protein consequence and their hydrophobicity group assignments

Chr=chromosome, Ref= Reference allele, Alt = Alternative allele

Chr	position	rsIDs	Ref	Alt	Protein Consequence	Transcript Consequence	VEP Annotation	Hydrophobicity group
21	33032087	rs1297567794	C	T	p.Ala2Val	c.5C>T	missense_variant	intermediate to hydrophobic
21	33032095	rs121912444	GC	G	p.Val6CysfsTer4	c.15delC	frameshift_variant	NA
21	33032095	rs121912444	G	A	p.Ala5Thr	c.13G>A	missense_variant	intermediate
21	33032096	rs121912442	C	T	p.Ala5Val	c.14C>T	missense_variant	intermediate to hydrophobic
21	33032097	rs199766524	C	T	p.Ala5Ala	c.15C>T	splice_region_variant	Synonymous
21	33032101	rs1312702973	T	G	p.Cys7Gly	c.19T>G	missense_variant	intermediate
21	33032102	rs121912448	G	A	p.Cys7Tyr	c.20G>A	missense_variant	intermediate
21	33032104	rs1380854315	G	A	p.Val8Met	c.22G>A	missense_variant	hydrophobic
21	33032109	rs772764888	G	A	p.Leu9Leu	c.27G>A	synonymous_variant	Synonymous
21	33032116	rs762628133	G	T	p.Asp12Tyr	c.34G>T	missense_variant	hydrophilic to intermediate
21	33032121	rs377178013	C	A	p.Gly13Gly	c.39C>A	synonymous_variant	Synonymous
21	33032126	rs1202989817	T	C	p.Val15Ala	c.44T>C	missense_variant	hydrophobic to intermediate
21	33032127	rs1251222457	G	T	p.Val15Val	c.45G>T	synonymous_variant	Synonymous
21	33032132	rs1200906022	G	C	p.Gly17Ala	c.50G>C	missense_variant	intermediate
21	33032134	rs1460554436	A	G	p.Ile18Val	c.52A>G	missense_variant	hydrophobic
21	33032136	rs1447729350	C	T	p.Ile18Ile	c.54C>T	synonymous_variant	Synonymous
21	33032136	rs1447729350	C	A	p.Ile18Ile	c.54C>A	synonymous_variant	Synonymous
21	33032139	rs1182088847	C	G	p.Ile19Met	c.57C>G	missense_variant	hydrophobic
21	33032139	rs1182088847	C	T	p.Ile19Ile	c.57C>T	synonymous_variant	Synonymous
21	33032141	rs768029813	A	G	p.Asn20Ser	c.59A>G	missense_variant	hydrophilic to intermediate
21	33032148	rs756458346	G	A	p.Glu22Glu	c.66G>A	synonymous_variant	Synonymous
21	33032151	rs1424217272	G	T	p.Gln23His	c.69G>T	missense_variant	hydrophilic

21	33032151	rs1424217272	G	C	p.Gln23His	c.69G>C	missense_variant	hydrophilic
21	33032154	rs1467183070	G	A	p.Lys24Lys	c.72G>A	splice_region_variant	Synonymous
21	33036098	rs1314754106	TAAAGG	T	p.Glu25Ter	c.73_77delGAAAG	splice_acceptor_variant	hydrophilic to STOP
21	33036107	rs747214897	G	C	p.Ser26Thr	c.77G>C	missense_variant	intermediate
21	33036107	rs747214897	G	A	p.Ser26Asn	c.77G>A	missense_variant	intermediate to hydrophilic
21	33036111	rs1489004175	T	C	p.Asn27Asn	c.81T>C	synonymous_variant	Synonymous
21	33036114	rs1217375069	A	G	p.Gly28Gly	c.84A>G	synonymous_variant	Synonymous
21	33036117	rs145198224	A	G	p.Pro29Pro	c.87A>G	synonymous_variant	Synonymous
21	33036120	rs748040402	G	A	p.Val30Val	c.90G>A	synonymous_variant	Synonymous
21	33036126	rs769715106	G	A	p.Val32Val	c.96G>A	synonymous_variant	Synonymous
21	33036134	rs777560607	G	T	p.Ser35Ile	c.104G>T	missense_variant	intermediate to hydrophobic
21	33036144	rs1405534640	A	G	p.Gly38Gly	c.114A>G	synonymous_variant	Synonymous
21	33036145	rs121912432	C	G	p.Leu39Val	c.115C>G	missense_variant	hydrophobic
21	33036154	rs121912433	G	A	p.Gly42Ser	c.124G>A	missense_variant	intermediate
21	33036161	rs121912435	A	G	p.His44Arg	c.131A>G	missense_variant	hydrophilic
21	33036168	rs1421563256	C	T	p.Phe46Phe	c.138C>T	synonymous_variant	Synonymous
21	33036169	rs748897491	C	G	p.His47Asp	c.139C>G	missense_variant	hydrophilic
21	33036178	rs752237082	G	A	p.Glu50Lys	c.148G>A	missense_variant	hydrophilic
21	33036178	rs752237082	GA	G	p.Glu50GlyfsTer39	c.149delA	frameshift_variant	NA
21	33036180	rs201045805	G	A	p.Glu50Glu	c.150G>A	synonymous_variant	Synonymous
21	33036182	rs759149157	T	G	p.Phe51Cys	c.152T>G	missense_variant	hydrophobic to intermediate
21	33036189	rs1276917683	T	C	p.Asp53Asp	c.159T>C	synonymous_variant	Synonymous
21	33038766	rs1446483921	T	C	p.Cys58Cys	c.174T>C	synonymous_variant	Synonymous
21	33038772	rs373888553	T	C	p.Ser60Ser	c.180T>C	synonymous_variant	Synonymous
21	33038774	rs1378635853	C	A	p.Ala61Glu	c.182C>A	missense_variant	intermediate to hydrophilic
21	33038787	rs147620646	T	C	p.Phe65Phe	c.195T>C	synonymous_variant	Synonymous
21	33038788	rs1283021712	AATCCTCT	A	p.Leu68GlufsTer19	c.201_207delTCTATCC	frameshift_variant	NA

21	33038791	rs1356474292	C	T	p.Pro67Ser	c.199C>T	missense_variant	intermediate
21	33038793	rs770457607	T	C	p.Pro67Pro	c.201T>C	synonymous_variant	Synonymous
21	33038798	rs778327622	C	A	p.Ser69Tyr	c.206C>A	missense_variant	intermediate
21	33038798	rs778327622	C	T	p.Ser69Phe	c.206C>T	missense_variant	intermediate to hydrophobic
21	33038800	rs1457291290	A	G	p.Arg70Gly	c.208A>G	missense_variant	hydrophilic to intermediate
21	33038808	rs368042695	C	T	p.His72His	c.216C>T	synonymous_variant	Synonymous
21	33038809	rs121912455	G	A	p.Gly73Ser	c.217G>A	missense_variant	intermediate
21	33038814	rs1459217491	G	A	p.Gly74Gly	c.222G>A	synonymous_variant	Synonymous
21	33038823	rs550738116	T	C	p.Asp77Asp	c.231T>C	synonymous_variant	Synonymous
21	33039573	rs121912458	A	G	p.His81Arg	c.242A>G	missense_variant	hydrophilic
21	33039584	rs121912452	T	C	p.Leu85Leu	c.253T>C	synonymous_variant	Synonymous
21	33039586	rs1315541036	G	C	p.Leu85Phe	c.255G>C	missense_variant	hydrophobic
21	33039586	rs1315541036	G	A	p.Leu85Leu	c.255G>A	synonymous_variant	Synonymous
21	33039595	rs749831011	G	T	p.Val88Val	c.264G>T	synonymous_variant	Synonymous
21	33039600	rs1280042397	C	T	p.Ala90Val	c.269C>T	missense_variant	intermediate to hydrophobic
21	33039602	rs1343616996	G	A	p.Asp91Asn	c.271G>A	missense_variant	hydrophilic
21	33039603	rs80265967	A	C	p.Asp91Ala	c.272A>C	missense_variant	hydrophilic to intermediate
21	33039604	rs1256439749	C	G	p.Asp91Glu	c.273C>G	missense_variant	hydrophilic
21	33039605	rs1345907062	A	G	p.Lys92Glu	c.274A>G	missense_variant	hydrophilic
21	33039609	rs774994509	A	G	p.Asp93Gly	c.278A>G	missense_variant	hydrophilic to intermediate
21	33039610	rs759731506	T	C	p.Asp93Asp	c.279T>C	synonymous_variant	Synonymous
21	33039612	rs121912438	G	C	p.Gly94Ala	c.281G>C	missense_variant	intermediate
21	33039612	rs121912438	G	A	p.Gly94Asp	c.281G>A	missense_variant	intermediate to hydrophilic
21	33039619	rs557930089	C	T	p.Ala96Ala	c.288C>T	synonymous_variant	Synonymous
21	33039619	rs557930089	CG	C	p.Asp97MetfsTer8	c.289delG	frameshift_variant	NA
21	33039620	rs121912459	G	A	p.Asp97Asn	c.289G>A	missense_variant	hydrophilic
21	33039622	rs111229903	T	A	p.Asp97Glu	c.291T>A	missense_variant	hydrophilic

21	33039622	rs111229903	T	C	p.Asp97Asp	c.291T>C	synonymous_variant	Synonymous
21	33039628	rs1432013430	T	C	p.Ser99Ser	c.297T>C	synonymous_variant	Synonymous
21	33039629	rs760740095	A	G	p.Ile100Val	c.298A>G	missense_variant	hydrophobic
21	33039640	rs1220006790	T	C	p.Ser103Ser	c.309T>C	synonymous_variant	Synonymous
21	33039648	rs1378590183	C	T	p.Ser106Leu	c.317C>T	missense_variant	intermediate to hydrophobic
21	33039649	rs763963373	A	T	p.Ser106Ser	c.318A>T	synonymous_variant	Synonymous
21	33039657	rs1359299834	G	A	p.Gly109Glu	c.326G>A	missense_variant	intermediate to hydrophilic
21	33039659	rs567432143	G	T	p.Asp110Tyr	c.328G>T	missense_variant	hydrophilic to intermediate
21	33039670	rs1299542356	C	T	p.Ile113Ile	c.339C>T	synonymous_variant	Synonymous
21	33039672	rs121912441	T	C	p.Ile114Thr	c.341T>C	missense_variant	hydrophobic to intermediate
21	33039673	rs750335577	T	G	p.Ile114Met	c.342T>G	missense_variant	hydrophobic
21	33039677	rs1301635320	C	T	p.Arg116Cys	c.346C>T	missense_variant	hydrophilic to intermediate
21	33039677	rs1301635320	C	G	p.Arg116Gly	c.346C>G	missense_variant	hydrophilic to intermediate
21	33039682	rs757951479	A	C	p.Thr117Thr	c.351A>C	synonymous_variant	Synonymous
21	33039686	rs1235629842	G	C	p.Val119Leu	c.355G>C	missense_variant	hydrophobic
21	33040784	rs1457889952	G	C	p.Val120Leu	c.358G>C	missense_variant	hydrophobic
21	33040786	rs1180155915	C	A	p.Val120Val	c.360C>A	splice_region_variant	Synonymous
21	33040788	rs1410925719	A	G	p.His121Arg	c.362A>G	missense_variant	hydrophilic
21	33040789	rs758204711	T	C	p.His121His	c.363T>C	synonymous_variant	Synonymous
21	33040804	rs765931024	C	T	p.Asp126Asp	c.378C>T	synonymous_variant	Synonymous
21	33040814	rs1464048449	G	A	p.Gly130Ser	c.388G>A	missense_variant	intermediate
21	33040816		T	C	p.Gly130Gly	c.390T>C	synonymous_variant	Synonymous
21	33040818	rs1169621300	G	A	p.Gly131Glu	c.392G>A	missense_variant	intermediate to hydrophilic
21	33040819	rs566007659	A	G	p.Gly131Gly	c.393A>G	synonymous_variant	Synonymous
21	33040821	rs1447747586	A	G	p.Asn132Ser	c.395A>G	missense_variant	hydrophilic to intermediate
21	33040830	rs121912451	G	A	p.Ser135Asn	c.404G>A	missense_variant	intermediate to hydrophilic
21	33040833	rs781031581	C	T	p.Thr136Ile	c.407C>T	missense_variant	intermediate to hydrophobic

21	33040837	rs1395498224	G	A	p.Lys137Lys	c.411G>A	synonymous_variant	Synonymous
21	33040840	rs752378382	A	G	p.Thr138Thr	c.414A>G	synonymous_variant	Synonymous
21	33040846	rs1804449	C	T	p.Asn140Asn	c.420C>T	synonymous_variant	Synonymous
21	33040847	rs1217353001	G	A	p.Ala141Thr	c.421G>A	missense_variant	intermediate
21	33040849	rs143100660	T	A	p.Ala141Ala	c.423T>A	synonymous_variant	Synonymous
21	33040854	rs1200970313	G	A	p.Ser143Asn	c.428G>A	missense_variant	intermediate to hydrophilic
21	33040856	rs746397967	C	T	p.Arg144Cys	c.430C>T	missense_variant	hydrophilic to intermediate
21	33040856	rs746397967	C	G	p.Arg144Gly	c.430C>G	missense_variant	hydrophilic to intermediate
21	33040861	rs1482760341	G	C	p.Leu145Phe	c.435G>C	missense_variant	hydrophobic
21	33040870	rs372540831	T	C	p.Gly148Gly	c.444T>C	synonymous_variant	Synonymous
21	33040871	rs567511139	G	A	p.Val149Ile	c.445G>A	missense_variant	hydrophobic
21	33040872	rs1476760624	T	G	p.Val149Gly	c.446T>G	missense_variant	hydrophobic to intermediate
21	33040874	rs1169917994	A	G	p.Ile150Val	c.448A>G	missense_variant	hydrophobic
21	33040875	rs1424014997	T	C	p.Ile150Thr	c.449T>C	missense_variant	hydrophobic to intermediate
21	33040882	rs1173749241	C	T	p.Ile152Ile	c.456C>T	synonymous_variant	Synonymous
21	33040883	rs747094021	G	A	p.Ala153Thr	c.457G>A	missense_variant	intermediate
21	33040885		C	T	p.Ala153Ala	c.459C>T	synonymous_variant	Synonymous
21	33040888	rs1465781849	A	G	p.Gln154Gln	c.462A>G	synonymous_variant	Synonymous
21	33040890	rs768697100	A	C	p.Ter155SerextTer6	c.464A>C	stop_lost	NA

Appendix E. Chapter 8 supplementary materials

Appendix E.1. Supplemental figures

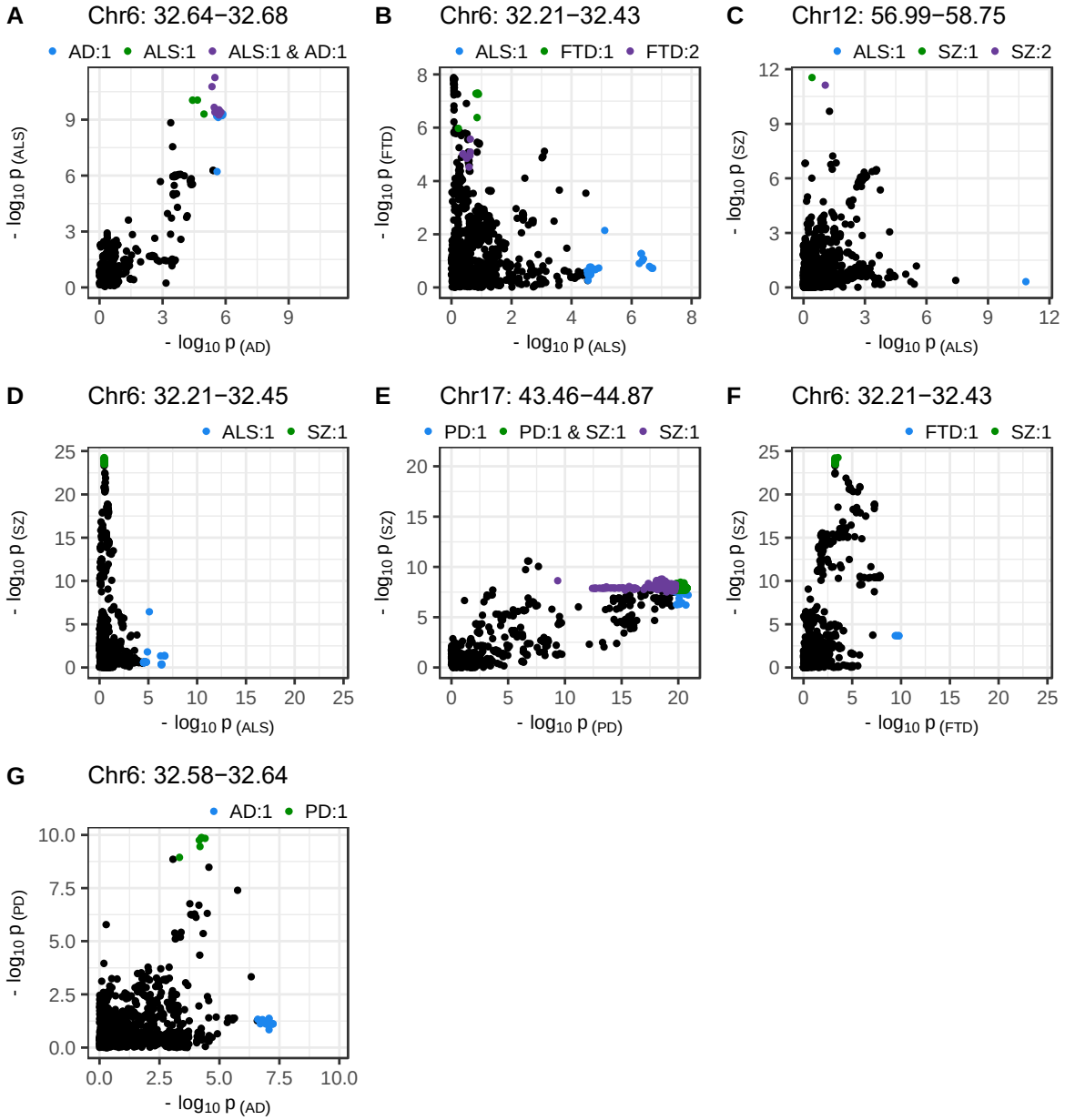


Figure E-1. SNP-wise p-value distribution between trait pairs in comparisons where colocalisation analysis suggested a causal variant in both traits

Colocalisation analysis supported the shared variant hypothesis for the comparison in panel A, and presence of distinct variants for each trait in all other panels (see Table 8-3). Colouring indicates fine-mapping credible sets assigned to SNPs across the traits compared; the legend above each panel is in the format 'Trait: credible set number'. The genomic position range shown above each panel is in Mb. AD = Alzheimer's disease, ALS = amyotrophic lateral sclerosis, FTD = frontotemporal dementia, PD = Parkinson's disease, SZ = schizophrenia.

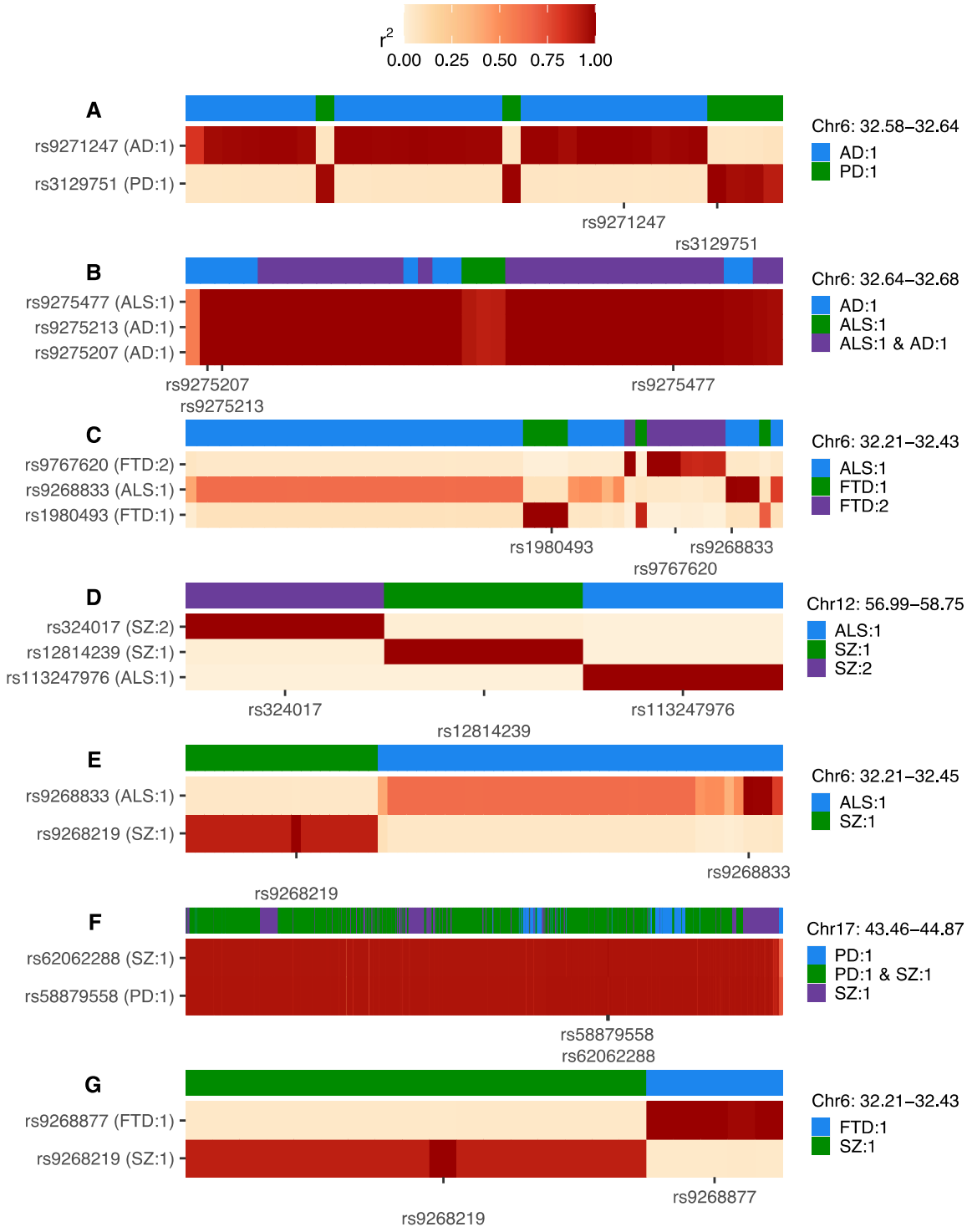


Figure E-2. Heatmaps of linkage disequilibrium (LD) in the 1000 Genomes European reference population across variants assigned to any credible set during univariate fine-mapping of trait pairs

LD is shown relative to the SNPs with the highest posterior inclusion probability (PIP) for each credible set, displaying all top PIP SNPs when ties occur. The y-axis splits by top PIP SNPs and the x-axis displays SNPs ordered by genomic position, marking only the positions of the top PIP SNPs. Credible set assignments for each variant are shown in the colour bar at the top of each panel and for the top PIP SNPs in the y-axis label; these are annotated in the format: 'trait: credible set number'.

The genomic range indicated at the top right of each panel refers to the positions spanned across all SNPs analysed and is in Mb. AD = Alzheimer's disease, ALS = amyotrophic lateral sclerosis, FTD = frontotemporal dementia, PD = Parkinson's disease, SZ = schizophrenia.

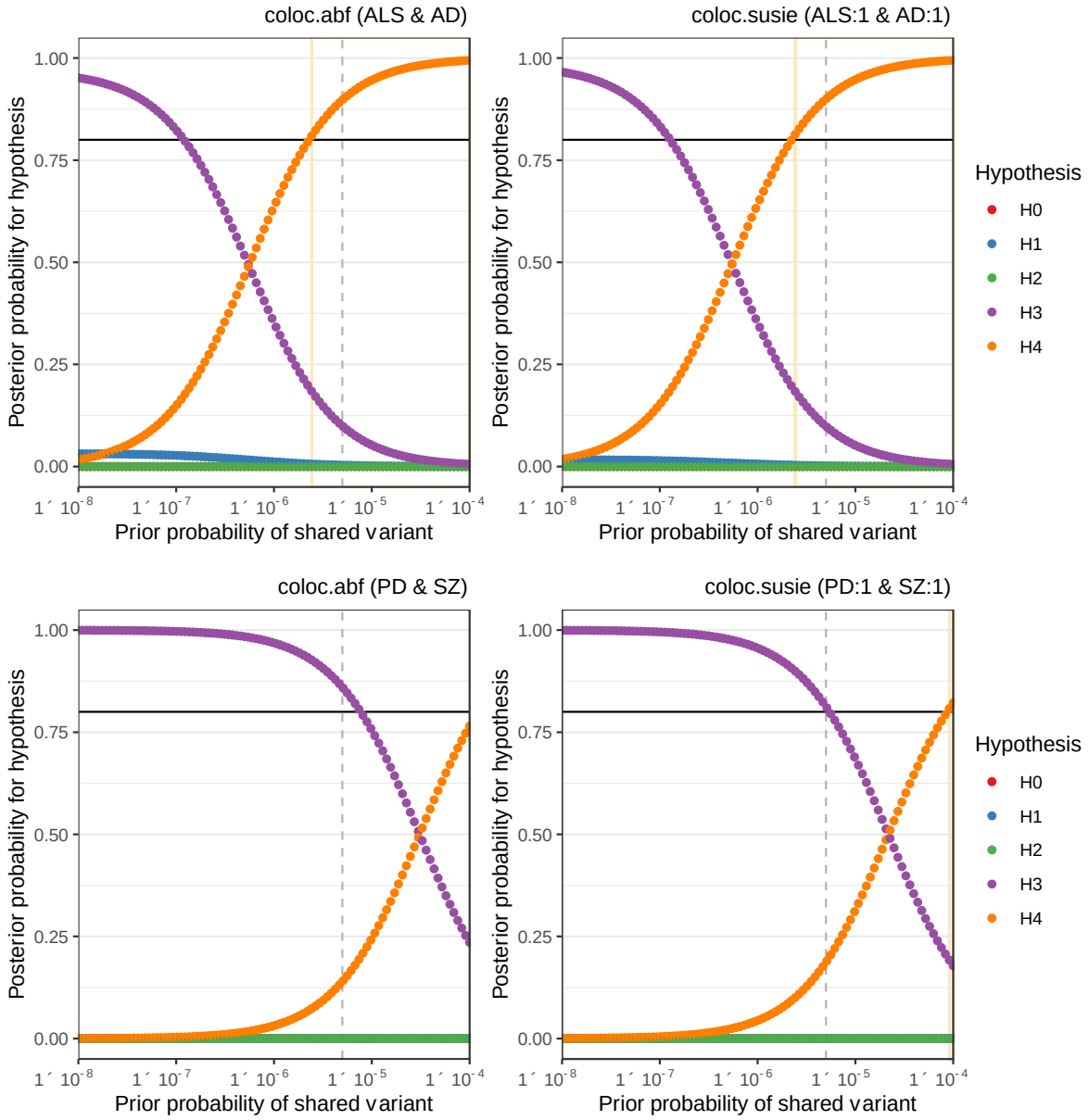


Figure E-3. Sensitivity of colocalisation analysis to the prior probability of a shared variant between traits

The upper panels display analysis at Chr6:32629240-32682213 between amyotrophic lateral sclerosis (ALS) and Alzheimer's disease (AD). The lower panels are for Chr17:43460501-44865832 between Parkinson's disease (PD) and schizophrenia (SZ). Panels labelled 'coloc.abf' display analysis across all SNPs in the region and 'coloc.susie' indicates analysis across the SNPs within the pair of fine-mapping credible sets identified across trait pairs. Plot points indicate posterior probability of each hypothesis (H0 = no causal variant for either trait, H1 = variant causal for the trait one, H2 = variant causal for trait two, H3 = distinct causal variants for each trait, H4 = a shared causal variant between traits), according to the prior probability of H4. The vertical hatched line indicates the prior H4 probability defined for the reported analysis; the black horizontal line indicates the defined threshold for acceptance of H4: posterior H4 probability >0.8. Cream shading of the plot area indicates prior H4 probabilities which yield a posterior probability of H4 above the threshold.

Appendix E.2. Supplemental tables

Table E-1. Results of all bivariate local genetic correlation analyses

Available in Excel workbook: Thesis_TSpargo_ExcelTables.xlsx

Table E-2. Results of colocalisation analyses performed across all SNPs sampled in region

*Number of SNPs refers to the number of SNPs in common between the two traits analysed and present within the 1000 genomes reference panel. Comparisons where univariate fine-mapping identified at least one credible set in each trait were also performed on the basis of these credible sets (see Table 8-3).ø H0 = no causal variant for either trait, H1 = variant causal for trait 1, H2 = variant causal for trait 2, H3 = distinct causal variants for each trait, H4 = a shared causal variant between traits. Annotations column: * denotes comparisons with genetic correlation analysis p-values below the strict Bonferroni correction threshold; all others passed FDR correction. Δ Denotes locus with boundaries extended by ±10kb compared to the region partition defined in genetic correlation analysis. Chr=chromosome, start/stop = GRCh37 base pair position range analysed*

Trait		Genomic position			N of fine-mapping credible sets for trait		Number of SNPs	Posterior probability for hypothesis ø					Annotation
1	2	chr	start	stop	1	2		H0	H1	H2	H3	H4	
AD	PD	1	18427821	19238924	0	0	2676	**0.79**	0.10	0.10	0.01	<0.01	
AD	PD	13	71010209	71976862	0	0	2938	**0.72**	0.12	0.13	0.02	0.01	
AD	PD	6	16566883	17391994	0	0	2059	**0.79**	0.08	0.11	0.01	<0.01	
AD	PD	6	32454578	32539567	2	0	109	0.05	**0.90**	<0.01	0.01	0.04	*
AD	PD	6	32576785	32639239	1	1	958	<0.01	<0.01	<0.01	**0.95**	0.05	Δ
AD	SZ	3	187939200	189451805	0	1	2691	**0.52**	0.04	0.42	0.03	<0.01	
AD	SZ	7	109095559	110479479	0	1	2637	<0.01	<0.01	**0.66**	0.28	0.06	
ALS	AD	17	9619357	10572617	0	0	2287	**0.86**	0.08	0.06	0.01	<0.01	
ALS	AD	6	2746532	3964072	0	0	3255	**0.75**	0.10	0.12	0.02	<0.01	
ALS	AD	6	32629240	32682213	1	1	475	<0.01	<0.01	<0.01	0.10	**0.90**	*

ALS	FTD	6	32208902	32454577	1	2	1709	<0.01	<0.01	0.04	**0.96**	<0.01	
ALS	PD	12	6948955	8123317	0	0	2382	**0.70**	0.18	0.10	0.02	<0.01	
ALS	PD	15	35118928	36583327	0	0	2969	**0.82**	0.09	0.08	0.01	<0.01	
ALS	PD	16	78973581	79662634	0	0	2236	**0.79**	0.06	0.13	0.01	<0.01	
ALS	PD	16	82310205	83077869	0	0	3215	**0.52**	0.06	0.37	0.04	<0.01	
ALS	PD	16	86058598	86748867	0	0	2909	**0.77**	0.08	0.13	0.01	<0.01	
ALS	PD	18	5575813	6756496	0	0	2752	**0.81**	0.07	0.11	0.01	<0.01	
ALS	PD	9	132999453	134141936	0	0	2437	**0.73**	0.06	0.19	0.02	<0.01	
ALS	PD	9	28067455	28766060	1	0	1982	0.16	**0.73**	0.02	0.08	0.01	
ALS	SZ	12	56987106	58748139	1	2	2260	<0.01	<0.01	<0.01	**1.00**	<0.01	
ALS	SZ	6	32208902	32454577	1	1	1711	<0.01	<0.01	0.04	**0.96**	<0.01	
ALS	SZ	9	8262304	9170120	0	0	2528	**0.81**	0.11	0.07	0.01	<0.01	
PD	SZ	17	43460501	44865832	1	1	2453	<0.01	<0.01	<0.01	**0.86**	0.14	
PD	SZ	22	48269178	48977538	0	0	1881	**0.67**	0.10	0.19	0.03	<0.01	
PD	SZ	3	71223282	72334704	0	1	2270	<0.01	<0.01	**0.91**	0.08	0.01	
PD	SZ	7	8718352	9225106	0	0	1591	**0.57**	0.04	0.37	0.02	<0.01	
SZ	FTD	6	32208902	32454577	1	1	1657	<0.01	<0.01	<0.01	**1.00**	<0.01	

Table E-3. Overview of credible sets identified across fine-mapping analyses in summary statistics harmonised across trait pairs

Available in Excel workbook: Thesis_TSpargo_ExcelTables.xlsx