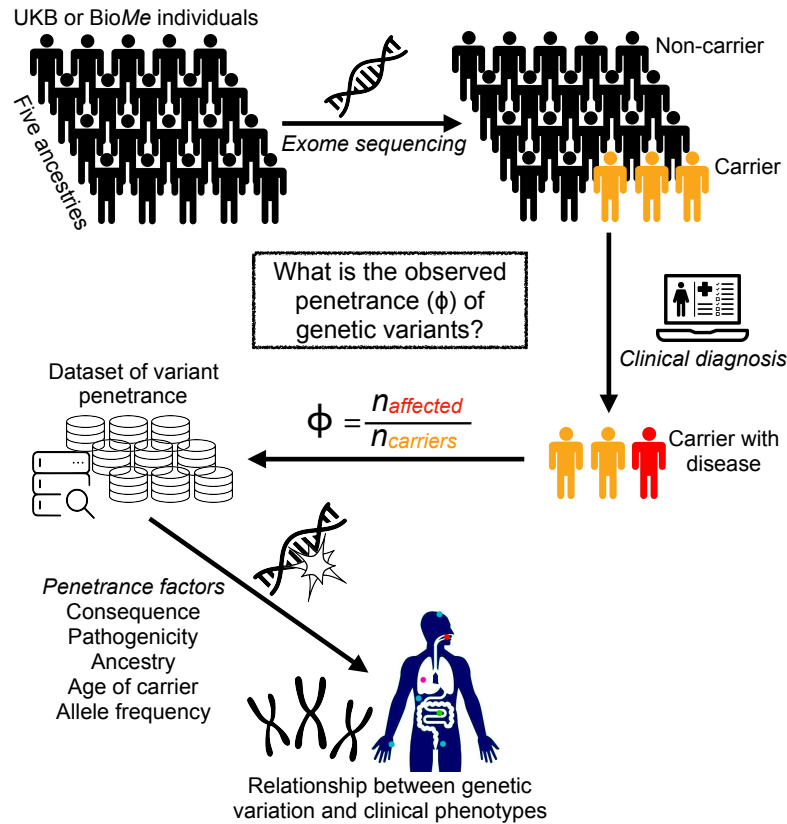


Supplemental Material

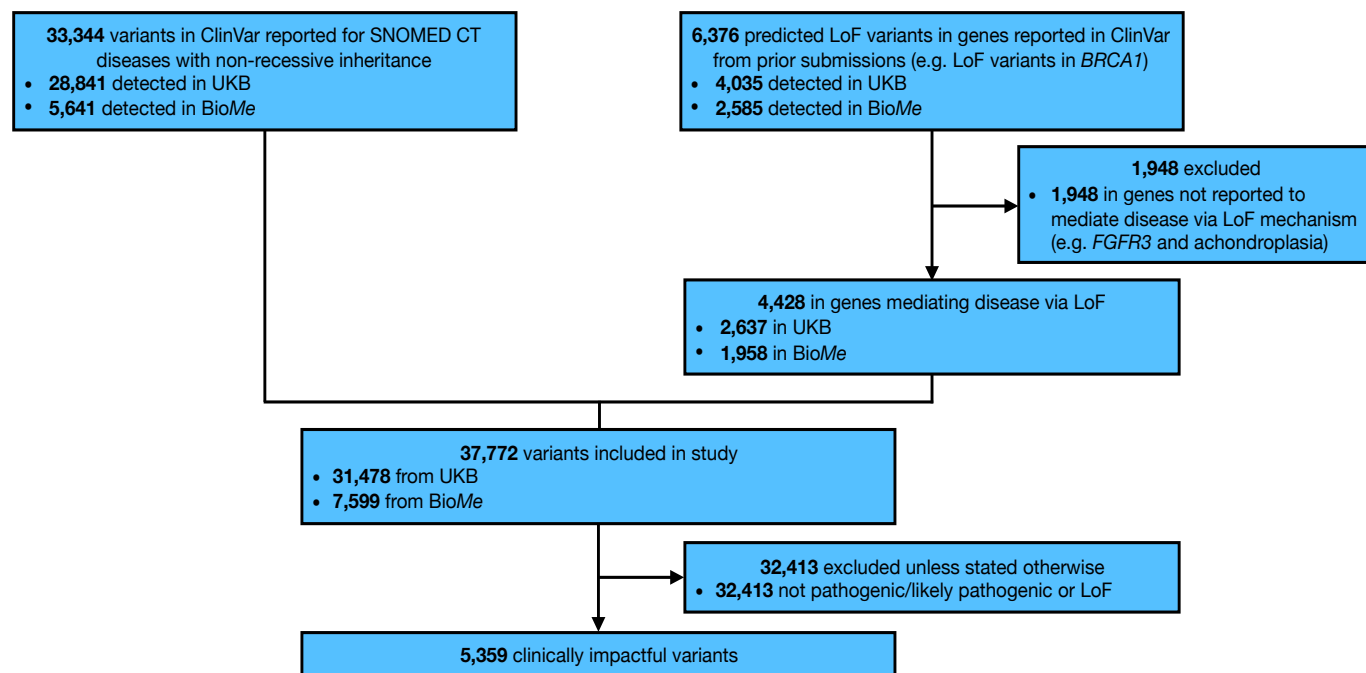
FIGURES

Supplementary Figure 1. Graphical abstract of study design and workflow.



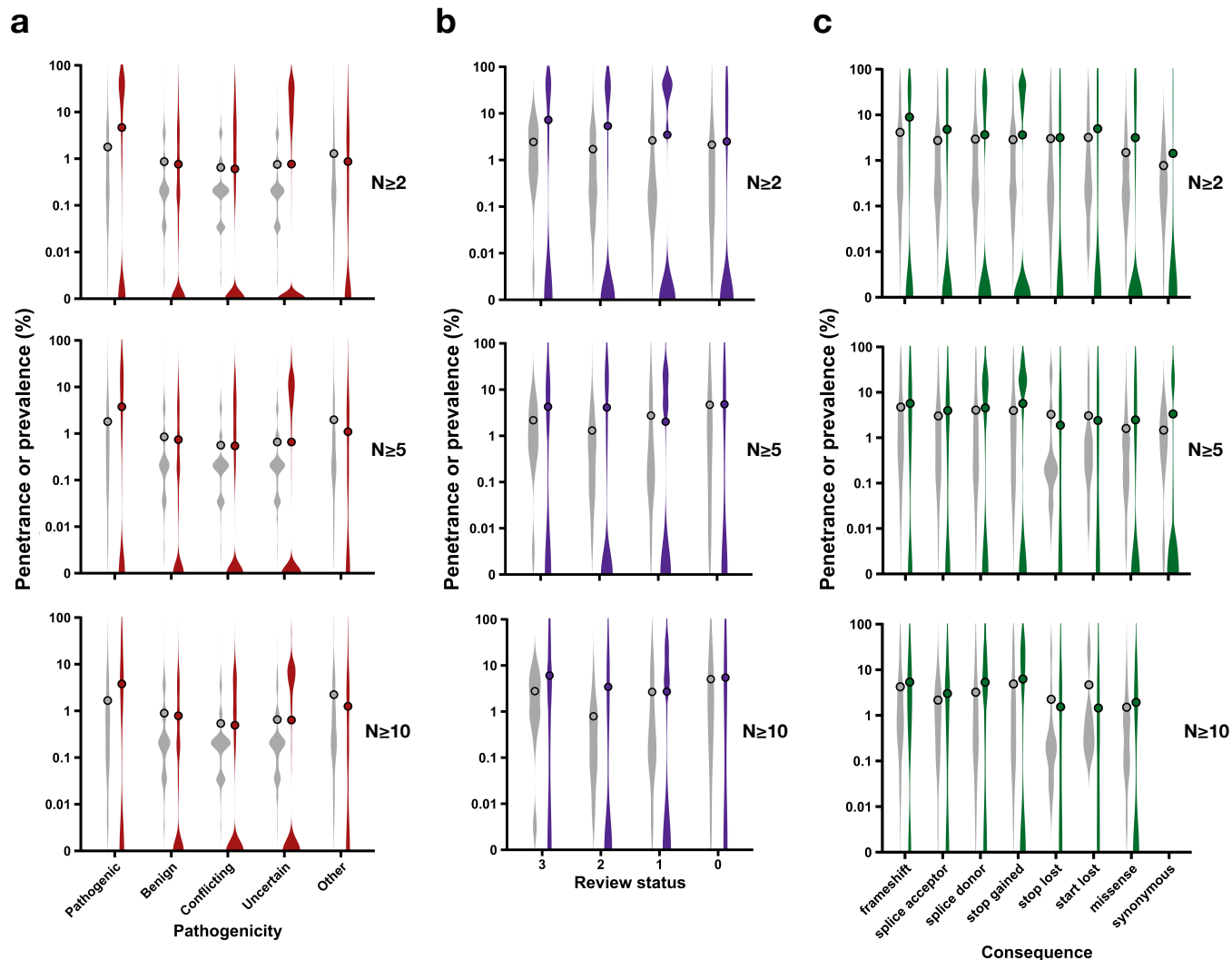
Graphical abstract of study design and workflow. A total of 72,434 individuals from the BioMe Biobank (BioMe) and UK Biobank (UKB) with exome sequence data were used to analyze penetrance of 37,772 clinical variants. Genetic information was linked to clinical phenotypes in the electronic health record (EHR), including diagnosis codes, metabolite levels, biological measurements, and physician notes, and was used to compute penetrance. The results were used to examine how penetrance varies by ClinVar pathogenicity class, molecular consequence, and carrier age and ancestry.

Supplementary Figure 2. Variant selection for analysis of penetrance.



Selection criteria of genetic variants included for analysis of penetrance in the BioMe Biobank (BioMe) and UK Biobank (UKB). Loss-of-function (LoF) variants in a gene were mapped to disease based on prior variant submissions in ClinVar linking the gene to disease (e.g., LoF variants in *BRCA1* were mapped to breast cancer based on prior variant submissions in ClinVar linking *BRCA1* variants to breast cancer). LoF variants in genes that do not mediate disease via LoF mechanism were excluded (e.g., *FGFR3* in achondroplasia). Clinically impactful variants are reported as pathogenic/likely pathogenic in ClinVar or are of LoF consequence in a gene mediating disease via LoF mechanism, and in a gene with non-recessive inheritance.

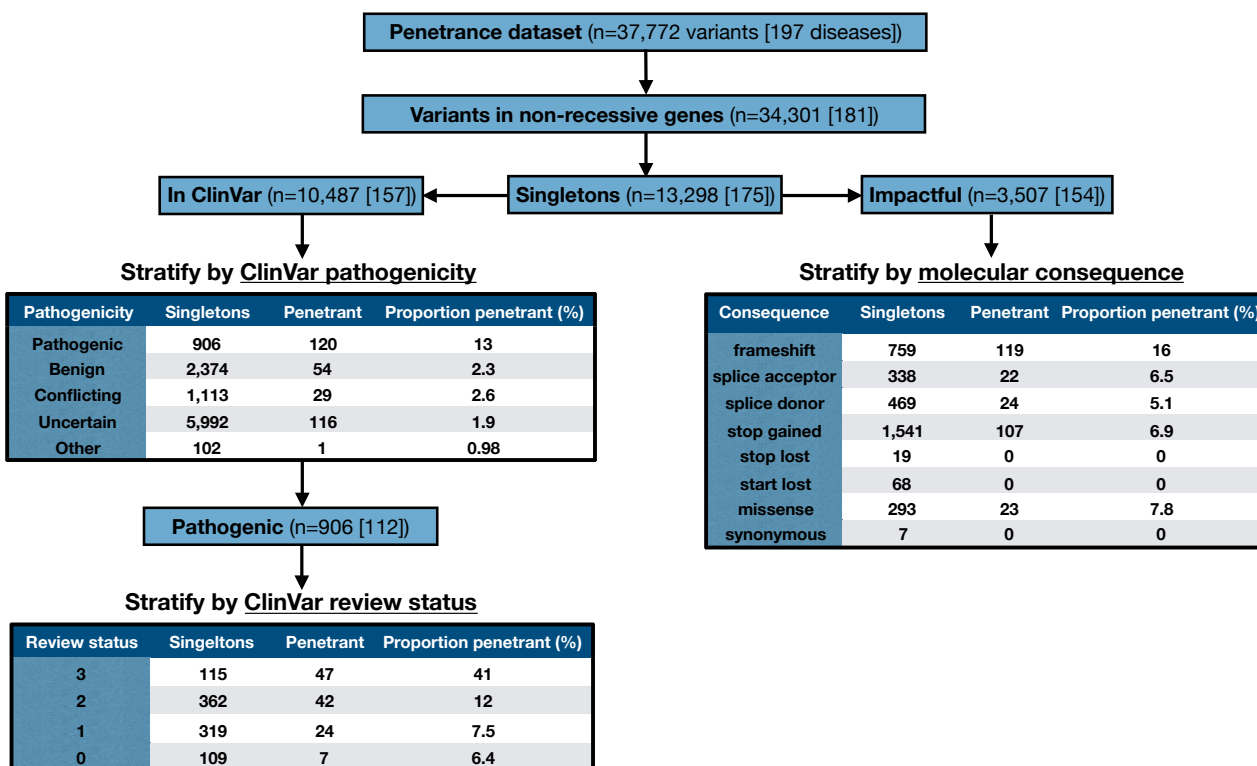
Supplementary Figure 3. Sensitivity analysis of penetrance estimates by sample size.



Sensitivity analysis of penetrance estimates by sample size. For increasing thresholds of sample size ($N \geq 2$ carriers [excludes singletons], $N \geq 5$ carriers, and $N \geq 10$ carriers), penetrance distributions are shown as violin plots in red, purple, and green color on a base-10 logarithmic scale with the mean penetrance overlaid as points, alongside disease prevalence in non-carriers shown as violin plots in grey color with the mean disease prevalence superimposed as points. Pathogenic/likely pathogenic variants are grouped as pathogenic and benign/likely benign variants are grouped as benign. **a**,

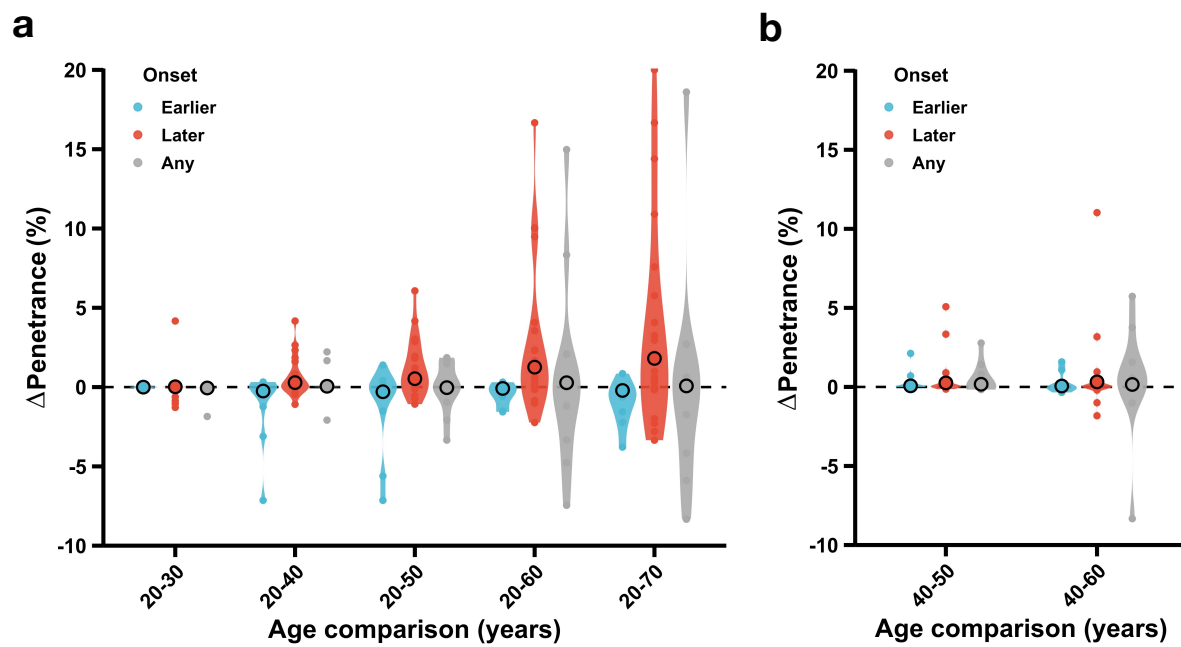
Penetrance is stratified by classification of variant pathogenicity in ClinVar. Pathogenic variants on average have the highest penetrance for all sample size thresholds (4.7% vs. 0.76% next highest mean for uncertain class [$N \geq 2$], 3.7% vs. 1.1 % next highest mean for other class [$N \geq 5$], and 3.8% vs. 1.3% next highest mean for other class [$N \geq 10$]; two-tailed t-test $P=5 \times 10^{-134}$, $P=1 \times 10^{-18}$, and $P=1 \times 10^{-15}$, respectively). Other pathogenicity, variants with a ClinVar pathogenicity classification other than pathogenic, benign, conflicting, or uncertain. **b**, Penetrance is stratified by variant review status in ClinVar as reviewed by experts (review status=3), multiple submitters (review status=2), single submitter (review status=1), or no assertion criteria (review status=0). Variants reviewed by experts (review status=3) have the highest penetrance on average for all sample size thresholds (7.2% vs. 5.4% next highest review status=2 group [$N \geq 2$], 4.3% vs. 4.1% next highest review status=2 group [$N \geq 5$], and 6.0% vs. 3.4% next highest review status=2 group [$N \geq 10$]; $P=2 \times 10^{-1}$, $P=9 \times 10^{-1}$, and $P=3 \times 10^{-1}$, respectively). **c**, Penetrance is stratified by molecular consequence using Variant Effect Predictor (VEP). Frameshift variants on average have the highest penetrance for all sample size thresholds (8.9% vs. 3.2% for missense [$N \geq 2$], 5.7% vs. 2.5% for missense [$N \geq 5$], and 5.4 vs. 1.9% for missense [$N \geq 10$]; $P=6 \times 10^{-10}$, $P=5 \times 10^{-3}$, and $P=1 \times 10^{-2}$, respectively).

Supplementary Figure 4. Penetrance of 13,298 singletons stratified by ClinVar pathogenicity, ClinVar review status, and molecular consequence.



Penetrance of 13,298 singletons stratified by ClinVar pathogenicity, ClinVar review status, and molecular consequence. Review status, level of evidence for a variant's assertion of pathogenicity in ClinVar ranging from 0 (no assertion criteria) to 3 (reviewed by experts). Clinically impactful variants are reported as pathogenic/likely pathogenic in ClinVar or are of loss-of-function consequence in a gene mediating disease via LoF mechanism, and in a gene with non-recessive inheritance.

Supplementary Figure 5. Association between age of disease onset and age-dependent change in penetrance for 157 diseases in the BioMe Biobank and UK Biobank.



Association between age of disease onset and age-dependent change in penetrance for 157 diseases in the BioMe Biobank and UK Biobank. Diseases correspond to the 5,359 clinically impactful variants and are grouped according to their age of onset: Earlier, Later, or Any. Change in penetrance is displayed as a violin plot for each age of onset group when comparing two carrier age thresholds with the mean change in penetrance superimposed as a point. Δ Penetrance (%), change in variant penetrance represented as a percent (+ values indicate greater penetrance estimate with the older age threshold and - values indicate greater penetrance estimate with the younger age threshold); age comparison, two carrier age thresholds for which penetrance is compared (e.g. 20-70 compares penetrance with carriers ≥ 20 years old and penetrance with carriers ≥ 70 years old); onset, disease groups according to age of onset. **a**, The BioMe Biobank has

carriers of age 20-90 years old and Δ Penetrance is assessed over five carrier age thresholds: Later diseases have greater Δ Penetrance on average than Earlier diseases for 20-40 (two-tailed t-test $P=0.006$), 20-50 ($P=0.002$), 20-60 ($P=0.02$), and 20-70 ($P=0.02$) age comparisons. **b**, UK Biobank has carriers of age 40-69 years old and Δ Penetrance is evaluated over two carrier age thresholds: there is no difference in Δ Penetrance between Later and Earlier diseases for either 40-50 ($P=0.2$) or 40-60 ($P=0.2$) age comparisons.

TABLES

Supplementary Table 1. Summary of cases/controls, ICD-10 diagnosis codes, and genes for 197 diseases analyzed for penetrance of genetic variants.

Supplementary Table 2. Summary of 37,772 clinical variants assessed for penetrance.

Supplementary Table 3. Validation of ICD-10-based phenotypes with clinical algorithms in the assessment of penetrance for nine diseases in the BioMe Biobank.

Supplementary Table 4. Tabulated list of 208 penetrance measurements computed with ICD-10-based phenotypes and clinical algorithms for ClinVar pathogenic variants in the BioMe Biobank.

Supplementary Table 5. Validation of ICD-10 phenotypes with manual curation of physician notes in the problem list (PL) for six diseases in the BioMe Biobank.

Supplementary Table 6. Summary of risk difference for 5,359 clinically impactful variants across 197 diseases.

Supplementary Table 7. Age of onset for 197 diseases included in study.

Supplementary Table 8. Examples of 20 clinically impactful variants associated with elevated risk of seven diseases.

Supplementary Table 9. Penetrance dataset for 12,844 variants analyzed in the BioMe Biobank.

Supplementary Table 10. Penetrance dataset for 60,407 variants analyzed in the UK Biobank.