HGG Advances

ARTICLE

Polygenic risk scores of endo-phenotypes identify the effect of genetic background in congenital heart disease

Sarah J. Spendlove,[1,2] Leroy Bondhus,[2,3] Gentian Lluri,[4] Jae Hoon Sul,[1,5] and Valerie A. Arboleda[1,2,3,6,*]

Abstract

Congenital heart disease (CHD) is a rare structural defect that occurs in ~1% of live births. Studies on CHD genetic architecture have identified pathogenic single-gene mutations in less than 30% of cases. Single-gene mutations often show incomplete penetrance and variable expressivity. Therefore, we hypothesize that genetic background may play a role in modulating disease expression. Polygenic risk scores (PRSs) aggregate effects of common genetic variants to investigate whether, cumulatively, these variants are associated with disease penetrance or severity. However, the major limitations in this field have been in generating sufficient sample sizes for these studies. Here we used CHD-phenotype matched genome-wide association study (GWAS) summary statistics from the UK Biobank (UKBB) as our base study and whole-genome sequencing data from the CHD cohort ($n_1 = 711$ trios, $n_2 = 362$ European trios) of the Gabriella Miller Kids First dataset as our target study to develop PRSs for CHD. PRSs estimated using a GWAS for heart valve problems and heart murmur explain 2.5% of the variance in case-control status of CHD (all SNVs, $p = 7.90 \times 10^{-3}$; fetal cardiac SNVs, $p = 8.00 \times 10^{-3}$) and 1.8% of the variance in severity of CHD (fetal cardiac SNVs, $p = 6.20 \times 10^{-3}$; all SNVs, $p = 0.015$). These results show that common variants captured in CHD phenotype-matched GWASs have a modest but significant contribution to phenotypic expression of CHD. Further exploration of the cumulative effect of common variants is necessary for understanding the complex genetic etiology of CHD and other rare diseases.

Introduction

Congenital heart disease (CHD) occurs in nearly 1% of live births[1] and is a major challenge in adult and pediatric global health. Despite medical, interventional, and surgical advancements, CHD still accounts for over 200,000 deaths globally per year,[2] including significant morbidity for those undergoing surgical correction of the congenital heart defect. Despite its prevalence in the pediatric population, our understanding of the underlying genetic factors that contribute to CHD remains incomplete.

CHD exists on a wide spectrum of genetic architectures. Single-gene (monogenic) disorders require that mutations in one or both alleles of a gene are necessary and sufficient to cause disease. Even as exome and genome sequencing studies of individuals with CHD have become increasingly common, the proportion of CHD cases explained by single-gene mutations has remained small. In isolated CHD cases, the genetic diagnosis rate is less than 20%,[3–5] although the diagnostic rate increases to greater than 50% in the presence of additional congenital anomalies.[6,7] Recent work has demonstrated that *de novo* and recessive forms of CHD are distinct from each other, occurring in specific gene pathways,[8] and that non-coding *de novo* variants contribute to CHD by disrupting transcriptional regulation[9] during cardiac development. Non-coding variants, *de novo* compound heterozygous mutations, and large copy number changes[5,10,11] remain underexplored sources of clinical genetic variation contributing to genetically undiagnosed CHD.

In families with a clinically established genetic diagnosis for CHD, there are reports of incomplete penetrance and variable expressivity.[12,13] In unrelated individuals harboring the same pathogenic genetic variant, the clinical presentation of the congenital heart defect can be variable.[14,15] The recurrence risk within families with an affected relative ranges between 3-fold and 80-fold increased relative risk.[16] These human genetic data are supported by observations in mouse models, where the common genetic background (strain background) affects the developmental and clinical phenotype of knockout alleles.[17–19] These reports suggest that common variants from the genetic background can contribute to the expressivity of monogenic forms of CHD[9] or may harbor a mutational load sufficient to cause disease.

The study of the contribution of common genetic variation is commonly performed using genome-wide association studies (GWASs) to identify loci and single-nucleotide variants (SNVs) associated with disease phenotypes. Polygenic risk scores (PRSs) are then typically calculated for

[1]Interdepartmental Bioinformatics Program, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA; [2]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA; [3]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA; [4]Ahmanson/UCLA Adult Congenital Heart Disease Center, Division of Cardiology, Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA; [5]Department of Psychiatry, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA; [6]Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
*Correspondence: varboleda@mednet.ucla.edu
https://doi.org/10.1016/j.xhgg.2022.100112.

Table 1. Overview of base study GWAS and phenotype abbreviations

UKBB GWAS name	Abbreviation	No. of cases	No. of controls	GWAS source	Phenotype source	Link
Diagnoses, main ICD10: R00 abnormalities of heartbeat	heartbeat	2,542	358,652	Neale lab	ICD-10	https://broad-ukb-sumstats-us-east-1.s3.amazonaws.com/round2/additive-tsvs/R00.gwas.imputed_v3.both_sexes.tsv.bgz
Non-cancer illness code, self-reported: heart arrhythmia	heart arrhythmia	2,013	359,128	Neale lab	PHESANT	https://broad-ukb-sumstats-us-east-1.s3.amazonaws.com/round2/additive-tsvs/20002_1077.gwas.imputed_v3.both_sexes.tsv.bgz
Non-cancer illness code, self-reported: heart valve problem/ heart murmur	heart valve	2,453	358,688	Neale lab	PHESANT	https://broad-ukb-sumstats-us-east-1.s3.amazonaws.com/round2/additive-tsvs/20002_1078.gwas.imputed_v3.both_sexes.tsv.bgz
395: heart valve disorders	valve disorders	4,239	402,421	PheWeb	Phecode	ftp://share.sph.umich.edu/UKBB_SAIGE_HRC/PheCode_395_SAIGE_MACge20.txt.vcf.gz
396: abnormal heart sounds	heart sounds	1,049	402,421	PheWeb	Phecode	ftp://share.sph.umich.edu/UKBB_SAIGE_HRC/PheCode_396_SAIGE_MACge20.txt.vcf.gz

Five CHD-related GWASs, computed using the UK Biobank (UKBB) and used as our base datasets for computing PRSs. The final copy of the data was downloaded on December 3, 2021.

an individual to assess genome-wide genetic risk in complex disease through cumulative effects of SNVs.[20] Because of GWASs' reliance on common variants of small effect, for practical reasons, the majority of work done with PRSs has been for complex diseases such as diabetes mellitus and[21] coronary artery disease[22] or quantitative phenotypes such as height.[23] In complex disease cohorts, case numbers can be in the hundreds of thousands and are well powered to identify even low-effect common genetic variants. Importantly, the development of PRSs can bridge results from large-scale GWASs into more interpretable and individualized results in the clinic. Early studies have shown promise for complex traits, such as coronary artery disease (CAD), which show significant SNV heritability of $h_{2,SNV}$ of 13.3% ± 0.4%.[24] Recent work has demonstrated that individuals who harbor a higher load of CAD risk-associated variants have an equivalent CAD risk, as seen in monogenic forms of CAD.[22]

However, the major hurdle to exploring the contribution of common variation for individuals with rare diseases, such as CHD, is the lack of well-powered GWASs that are publicly available and the availability of sizable cohorts of affected individuals to test the PRSs. To our knowledge, there are only a few studies that have explored the contribution of common genetic variants to CHD. GWASs have identified loci associated with CHD subtypes such as atrial septal defects[25] and tetralogy of Fallot.[26] A recent study used PRSs to investigate the risk of another subtype of CHD, atrioventricular septal defects (AVSD), that has variable penetrance in Down syndrome infants.[27] A GWAS on dextro-transposition of the great arteries (D-TGA), a severe form of CHD, identified a risk locus on chromosome 3. This study provides additional support for the idea that common variants contribute to CHD risk.[28]

To address the lack of well-powered GWASs for CHD and further dissect the contribution of common variants to CHD risk, we leveraged CHD-matched GWASs of subphenotypic manifestations of abnormal cardiac development, or endo-phenotypes, to develop PRSs aggregating the effect of identified genetic risk variants on these endo-phenotypes (Table 1). These endo-phenotypes reflect CHD diagnosed at birth as well as more "mild" phenotypes indicative of abnormal heart development that can be ascertained in a large biobank cohort. Understanding endo-phenotypes indicative of less severe forms of CHD is important because previous work has shown that there is shared regulation of monogenic disease genes that influences rare and complex traits that are phenotypically similar.[29] Exome and genome data, as well as epidemiological studies, suggest that an individual's heart development is influenced by multiple types of genetic variation. We show in this study that this unique endo-phenotype approach allows us to harness well-powered GWASs while still giving insights into the genetic architecture of more clinically severe CHD subtypes.

To do this, we develop PRSs for CHD endo-phenotypes and test them in a cohort of 711 individuals who have a clinical diagnosis of CHD[9] but have had no causal pathogenic variant identified. We leveraged summary statistics from publicly available GWASs on the UK Biobank (UKBB) cohort as our base studies, comparing cardiac phenotypes collated through PHEnome Scan ANalysis Tool (PHESANT) or using a phecode approach, where each GWAS represents endo-phenotypes related to cardiac development, to build PRSs using PRSice-2.[30] We then tested these scores using the

Gabriella Miller Kids First CHD whole-genome sequencing (WGS) dataset as our target study. To account for the lack of a genetically distinct validation cohort, we used a permutation process to generate empirical p-values. Our study demonstrates a significant contribution of common genetic variation to rare-disease CHD phenotype expressivity. Many of the common variants that were included to build our significant PRS models are found in or within 10 kb of known cardiac genes. This study shows the critical importance of phenotype choice for GWASs to capture the contribution of genetic background on CHD risk and severity. This is one of the first attempts to explore the role of common variation within the individual genetic background using PRSs in a cohort of individuals with rare CHD.

Material and methods

Target data

The Gabriella Miller Kids First Pediatric Research Program[31] was launched in 2015 and has WGS data for individuals with childhood cancer and/or structural birth defects. WGS data from the CHD cohort (dbGAP: phs001138.v3.p2) of the Gabriella Miller Kids First dataset were used as the target dataset for PRS analyses. This cohort focuses on individuals with genetically undiagnosed forms of CHD.[9,32] gVCF files were downloaded that contained data from 2,133 participants in 711 trios. To remove technical artifacts from the WGS samples, joint calling with a batch of 200 gVCF files using the Genome Analysis Toolkit (GATK) CombineGVCFs was performed, followed by joint genotyping using GATK GenotypeGVCFs v.3.7 across all batches. GATK's Variant Quality Score Recalibration (VQSR) was then applied to SNVs and insertions or deletions (indels) separately, and variants failing VQSR were removed. Genotypes with genotype quality (GQ) scores of 20 or less were set to missing.

Target data quality control metrics for PRSs

Quality control (QC) of WGS was performed on a variant and an individual level. For variant QC, (1) any heterozygous haploid calls from male X chromosome and female Y chromosome calls were set to missing, (2) multi-allelic SNVs and indels were removed, (3) monomorphic variants were removed because these are not useful in PRS analyses, (4) SNVs and indels with a missing rate greater than 5% were removed, and (5) variants with a Hardy-Weinberg equilibrium (HWE) p-value of less than 1×10^{-5} were removed.

Mendelian errors were removed using the following logic. We removed common SNVs and indels (minor allele frequency (MAF) $\geq$ 5% among parents) with Mendelian errors among all trios greater than 5 and rare SNVs and indels (MAF $<$ 5% among parents) with Mendelian errors among all trios greater than 3. We then applied Polymutt[33] to further refine genotypes in VCF files according to trio structure and sequencing quality and to remove Mendelian errors. Polymutt was applied using the default parameters. Following Polymutt, about 99.5% of Mendelian errors for SNVs and indels were corrected. After variant-level QC, we ended up with 63,482,867 SNVs and 4,246,140 indels (Table S1).

For individual-level QC, individual samples with a genotype missing rate greater than 5% were considered to be of low sequencing quality and removed. In this dataset, all individuals had genotype missing rates of less than 5% and were retained for downstream analysis. Identity-by-descent analysis using Plink 1.9[34] was used to verify that trio structures were consistent with the genetic data. Matching of sample labeled sex and chromosomal sex predictions from Plink 1.9 was done to ensure proper sample labeling.

Target data generation of pseudo-controls

Target data were obtained from the CHD subset of the Gabriella Miller Kids First (GMKF) dataset. Because target data were trio based instead of organized in a case-control structure, pseudo-controls were generated using the non-transmitted parental haplotypes. This is necessary because the parents of the probands cannot be counted as true controls because they share the same common risk alleles the probands have, given their close relatedness. Thus, 711 GMKF trios were translated into 711 cases and 711 pseudo-controls. This was done using the development build of Plink 1.9.[35–37] Figure 1B shows how pseudo-controls were created from non-transmitted alleles. In brief, the non-transmitted parental alleles from both parents are combined into one genotype, whereas transmitted parental alleles (i.e. those that are inherited in the child) are removed, creating one complete control genotype instead of two parent genotypes that are neither true controls nor true cases according to the reasoning explained above. Mendel errors and alleles where the data are missing from one of the parents are set to missing in the children and the resulting pseudo-control. The end result is a set of cases and pseudo-controls that should have unchanged linkage disequilibrium (LD) structure because we are simply taking half of the haplotypes from one parent and half from another to combine them into a control that does not contain the haplotypes shared by the proband. Pseudo-controls have been repeatedly used in the scientific literature in studies where the controls are related to the probands, including in PRS studies.[35–38]

Target data ancestry grouping

Principal-component analysis was performed using the EIGENSTRAT[24,39] software to check the ancestry group of the parents in the trios using the 1000 Genomes dataset[40] as a reference. Prior to principal-component analysis (PCA), related individuals from both datasets were removed (for CHD target data, only parents were included). In addition, indels and non-biallelic variants were removed, variants with a MAF of greater than 15% were filtered out, and LD pruning was performed. The CHD target dataset was further processed to correct Mendel errors, and variants were filtered according to HWE ($p < 1 \times 10^{-5}$) and genotype missing rate of greater than 0.5%. The 1000 Genomes dataset was transferred from GrCH37 to GrCH38 using the LiftOver tool.

The PCA results allowed us to separate trios based on whether both parents had European ancestry or whether one or both parents had non-European ancestry. After PCA, we found 362 trios with two parents of European ancestry and 349 trios with one or both parents of non-European ancestry.

CHD gene sets

We curated a gene set composed of 14,167 expressed genes from RNA sequencing (RNA-seq) data collected from fetal cardiac tissue[41] that we used in our PRS generation (Table S2). When filtering the GMKF genotypes for variants in these gene sets, we included variants in these genes or within 10 kb of these genes. We also curated a smaller set of CHD-related genes (Table S3) that we used to check whether the variants included in our PRS models were in or within 10 kb of known CHD genes.

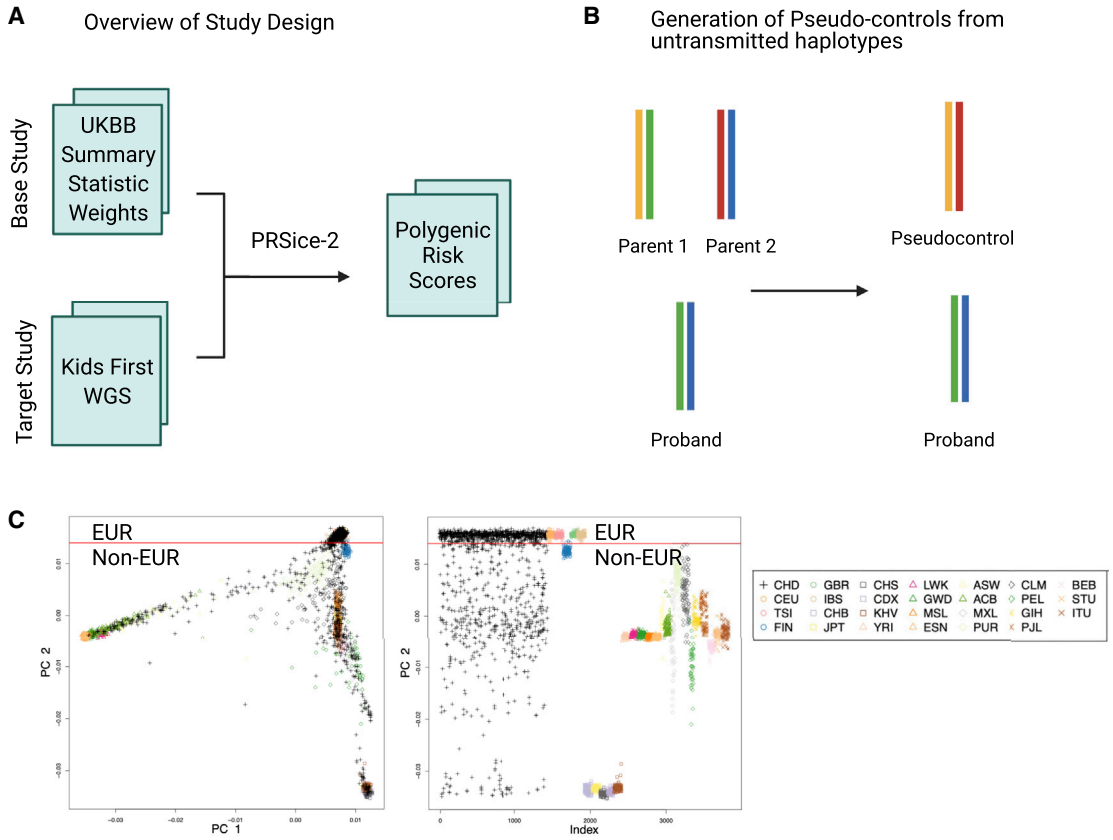

Figure 1. Schematic of the pipeline for results and generation of pseudo-controls

(A) Overview of the study design. Our base study for generation of PRSs is a set of 5 CHD-related GWASs from the UKBB. The summary statistics from these GWASs are used inside PRSice to generate the PRS model. Our target study is a subset of whole-genome sequences from the GMKF dataset. Using the PRS model generated with the base study, PRSice generates a PRS for each individual in the targeted study.

(B) Generation of pseudo-controls from non-transmitted haplotypes. The target study consists of trios of one affected proband and two presumably unaffected parents. Thus, to generate pseudo-controls, we use Plink to combine the non-transmitted parental alleles into a control genotype, working under the assumption that any genetic variants that modulate CHD risk were passed onto the child and are not included in the non-transmitted haplotype.

(C) Genetic ancestry distribution of parental genotypes. Genetic ancestry distribution of parental genotypes (black) compared with 1000 Genomes genotypes (colored), with the red line indicating samples that cluster genetically with European (EU) samples in 1000 Genomes. For our analysis, we separated our target dataset into trios with two parents whose genes cluster in PCA with samples of EU ancestry in the 1000 Genomes dataset and those with at least one parent whose genes do not cluster in PCA with samples of EU ancestry in 1000 Genomes dataset. In the legend, CHD represents our parental haplotypes from our GMKF target dataset, and 1000 Genomes population code descriptions can be found at https://www.internationalgenome.org/category/population/.

GWAS base data

Phenotype-matched summary statistics from GWASs were obtained from two sources running GWASs on the UK Biobank (UKBB) dataset. These GWASs were run using phenotype annotations from distinct methods for identifying patients with endophenotypes for CHD. One challenge with assessment of mild representations of clinical phenotypes such as CHD is that they are often sub-clinical and therefore not uniformly assessed in the electronic health record. The UKBB represents a unique cohort that has life-long electronic health record (EHR) data (ICD codes) for participants but also includes nurse-collected and self-reported data. Obtaining matched and accurate phenotypes is one of the major challenges in rare-disease GWAS and PRS studies, and we explored the utility of GWASs that leverage different data sources to obtain our phenotypes of interest.

In this study, we use two main sources of GWAS base data derived from the UKBB: (1) the Neale lab GWAS analysis[42] and (2) PheWeb GWAS analysis.[42,43] The Neale lab GWAS used phenotypes that were PHESANT curated[44] (a combination of self-reported or nurse-assigned based on an interview) or based on ICD-10 codes. Two of our phenotypes (heart valve and heart arrhythmia) were PHESANT curated, and one (heartbeat) was based on ICD-10 codes. These final summary statistics from the Neale lab were chosen because they are endo-phenotypes that reflect CHD-specific manifestations that are typically diagnosed at birth as well as more "mild" phenotypes that might be commonly ascertained in a large biobank cohort. These phenotypes were binary and identified using the ICD-10 or through an altered version of the PHESANT tool.[44] Two other phenotypes from the Neale lab GWAS we looked at in exploratory analyses, congenital malformations of the heart and great arteries and congenital malformations of the cardiac septa, were ultimately excluded from the study because of a very large imbalance between cases and controls that resulted in insufficient power.

One reason for this imbalance is the fact that more mild CHD phenotypes may not be diagnosed until later in life. In addition, the cohort in the UKBB has a median age of 58. However, surgeries to improve life expectancy in children with severe cases of CHD were not as successful until the late 20th century, and, therefore, we would expect that this dataset might have fewer cases of congenital malformations of the heart and great arteries' compared with more common conditions, such as heart arrhythmia or heart valve problems (heart murmur).

The PheWeb GWAS base data were used to determine whether phecode-based phenotypes provided an improvement of the PRS study. The major differences in these GWASs are as follows: (1) the PheWeb version has denser genotyping and more white British individuals (20 million imputed variants in 400,000 white British individuals),[43] and (2) the phenotypes are all EHR derived using phecodes, which classify ICD codes that represent a spectrum of common phenotypes under the same phecode, which is intended to improve phenotype accuracy and power.[45,46] As a follow-up, two phenotype-matched summary statistics from the UKBB Haplotype Reference Consortium(HRC)-imputed GWASs obtained from the PheWeb website, published as a peer-reviewed correspondence in *Nature Genetics* in June 2020,[43] were downloaded (final download date: December 3, 2021) for two CHD-related phenotypes: heart valve disorders (395) and abnormal heart sounds (396). These phenotypes were created by merging EHR-derived ICD billing codes, as described on the PheWeb website.[47] All base GWASs use the same UKBB population, but the phenotype curation, exact number of individuals, and number of imputed variants varies slightly between the two sources, and so we wanted to see whether different phenotype sources provided stable results in our PRS analysis.

Before performing PRS analysis with PRSice, we filtered out rare variants from the base data (minor allele frequency $< 5\%$) and lifted the data using the LiftOver tool and the associated hg19-to-GRCh38 chain file. When running PRSice, we used the alternate allele as the effective allele (A1) for the Neale lab GWAS and the PheWeb GWAS.

PRS calculation with PRSice2

The PRSs for the CHD target data were built using the PRSice2 software[30] and the five UKBB base GWASs. QC on the target dataset and minor allele filtering on the base datasets were completed as described previously. PRSs using all SNVs and SNVs in genes expressed in fetal cardiac tissue were generated. PRSice then uses LD pruning and base GWAS p-value thresholds to select the most informative SNVs to incorporate into the PRS (Figure S1). PRSs from all GWASs were generated for CHD severity group status as well as case-control status, resulting in a total of 10 European-ancestry case-control PRSs using 5 UKBB base GWASs and all variants or variants from genes expressed in fetal cardiac tissue and 6 total European ancestry severity PRSs using the same 5 UKBB base GWASs and the same two sets of variants. We tried to replicate significant case-control results from the Neale lab GWAS using the non-European ancestry subset of our target data but were unsuccessful, likely because of differences in LD because the genetic ancestry of the non-European target data did not match the genetic ancestry of the base data (Figure S4; Table S6). PRSice was run as recorded on our GitHub page (see data and code availability). We specify (1) that the base data provides beta values, (2) whether the PRS was to be calculated using a binary target (as in case-control) or quantitative target (as in severity), (3) to include non-founders, (4) that we wanted to print out the SNVs from the final model, and (5) that empirical p-values were to be calculated using permutation (to compensate for overfitting issues inherent in the PRSice software as well as the multiple testing created in this study by analyzing a handful of different but related phenotypes). The severity PRSs used a phenotype file indicating severity (mild, moderate, severe, pseudo-control) according to expert classification (Table S4), described under clinical phenotype severity classification. Individuals with unknown severity were not included in the phenotype file and so were not used for developing the severity PRS model, although a PRS was still calculated for these individuals.

PRSice p-values of model fit are generated by testing for an association between PRS (continuous) and the phenotype of interest. To account for overfitting, linear regression is used in a permutation procedure to generate empirical p-values. PRSice explains that the linear regression t-statistic is similar to the logistic regression t-statistic. In the figures, the case-control PRS odds ratios according to PRS decile were calculated using logistic regression, and for severity PRSs, the change in phenotype given score in decile was calculated using linear regression.

We would like to make a note regarding the statistical tests we decided to use in this paper. We considered the benefits of recalculating p-values using a conditional logistic regression or a pairwise t-test that match probands with their corresponding pseudo-control. PRSice documentation indicates that the simple logistic regression p-value is similar to the linear regression p-value. We also performed a verification analysis taking the heart valve, all SNVs, phenotype and comparing the p-value of model fit from PRSice with the p-value resulting from a conditional logistic regression (Pheno $\sim$ PRS + strata(FamilyID) using clogit in R). These p-values were similar (linear $= 2.918 \times 10^{-4}$ versus conditional $= 2.953 \times 10^{-4}$). A pairwise t-test in R (pairwise_t_test(PRS $\sim$ Pheno, paired = TRUE)) also gives a p-value of 3.03×10^{-4}. Because all of these p-values are similar, we decided to keep using the default PRSice settings. Several recent papers[34,35] also used pseudo-controls with PRSice or Plink[34] (which uses a similar clumping and thresholding PRS method), showing that our methodology is on par with the standard for recent PRS research using pseudo-controls. Finally, Peyrot et al.[48] have shown that, in GWASs with trio families (i.e., single-proband families) for diseases without known assortative mating, that pseudo-controls are equivalent to unscreened controls.[48]

Clinical phenotype severity classification

Clinical severity phenotype classification on the GMKF target data was based on a 2001 classification strategy.[49] Affected individuals were classified into 4 groups on a scale of 0–3: 0, no CHD; 1, simple; 2, moderate complexity; 3, great complexity (Table S4). All individuals here were classified based on initial presentation. Terms in the phenotype table from the GMKF CHD subset were manually mapped to analogous terms in Warnes et al.[49] Disease severity for individual was set to reflect the most severe CHD phenotype present. If there was an ambiguous phenotype that did not map to a term in the classification strategy,[49] these were manually assigned. Severity assignments were then manually checked and curated by a board-certified clinical cardiologist in adult CHD. Samples lacking a cardiac phenotype assignment or that were of unknown severity were removed from severity PRS analysis ($n = 5$).

Variant and gene analysis

After developing our PRS scores, we looked at the variants included in each of these scores and calculated whether these variants fell within a gene and how many of these genes were within 10 kb of CHD or fetal cardiac genes as defined in our curated gene sets (described

above). We also generated a list of all SNVs that were present in a significant or nominal significant PRS and identified whether they were located near or within 10 kb of known CHD genes.

Results

Our goal in this study was to apply phenotype-matched GWASs to development of a PRS for CHD. To do this, we developed PRSs for a case-control and a quantitative, phenotype-severity model of CHD using GWASs of CHD-related phenotypes from the UKBB[50] as base datasets. We tested these PRSs on a WGS study of individuals with CHD from the GMKF study. The individuals in this study were tested previously for pathogenic causes of rare monogenic diseases, and no causal pathogenic variants were identified; therefore, they represent the genetically undiagnosed subset of individuals with CHD.[32] Figure 1A shows a schematic representing the pipeline we used to obtain our PRS results.

One of the challenges when studying the role of common variants in rare diseases is that the focus is often on the most severe and monogenic forms of disease. Mild forms of the disease are not uniformly ascertained and included in the medical record, particularly when no medical intervention is required. The ascertainment bias toward more severe CHD cases makes it challenging to identify the mild-to-severe spectrum of congenital malformations in heart development. Phenotype quality is critical for high-quality and robust GWASs. Therefore, we sought GWASs that leveraged phenotypes related to CHD and that had at least 1,000 cases within the UKBB to ensure sufficient power.

There are several ways of curating more accurate phenotypes in large scale biobanks. Here we used, as our base data, GWASs that derive phenotypes from the UKBB using three different approaches: self-reported or nurse-reported,[44] raw ICD-10 codes,[51] or phecodes.[47] Each of these rely on different data fields or abstractions of raw phenotype data in the UKBB. Billing codes (ICD-10) are structured but are often inaccurate because the same condition can be classified by multiple different codes. One approach to increase the accuracy of ICD-10 coding for genomic studies is to generate Phecodes,[45,47] which map the more than 60,000 ICD codes representing various disease diagnoses to fewer than 2,000 related phenotypes by leveraging the ICD code hierarchical structure. Phecode-based approaches have been successfully used in EHR-linked biobanks to improve phenotyping efforts.[52,53] The alternative approach is to use self-reported datasets, which have been increasingly used across phenotyping studies and GWASs[54–56] but can be labor intensive to obtain. We used GWASs from two sources: (1) the Neale lab V2 GWASs,[42] which primarily used PHESANT[44] to perform automated aggregation of self-reported and nurse-reported phenotypes, and (2) PheWeb GWASs,[43] which leveraged Phecodes created within the UKBB ICD codes to create more robust phenotypes.

We computed CHD PRSs using UKBB GWAS base datasets for the phenotypes abnormalities of heartbeat (referred to as "heartbeat" hereafter), heart arrhythmia, and heart valve problems/murmur (referred to as "heart valve" hereafter)[42] and from PheWeb for heart valve disorders (referred to as "valve disorders" hereafter) and abnormal heart sounds (referred to as "heart sounds" hereafter) (Table 1). These GWASs include over 350,000 individuals of European ancestry. We chose these phenotypes for three reasons. (1) These GWASs were powered well enough to detect significant genetic associations, whereas GWASs representing more severe manifestations of CHD were grossly underpowered. (2) The broad clinical spectrum of CHD is likely better represented by these more subtle endo-phenotypes. (3) Using a matching genetic ancestry can have a significant influence on PRS results.

Our exploratory analyses showed that traditionally curated CHD phenotypes were grossly underpowered and highly unbalanced with respect to cases versus controls. The GWAS study for "congenital malformations of the heart and great arteries'' from the Neale lab had only 427 cases, and "Congenital malformations of the cardiac septa" had 313 cases in the UKBB study of over 350,000 individuals of European ancestry. Therefore, we opted to use surrogate measures of cardiac electrical and structural defects to serve as proxies in CHD. There were at least 1,000 cases represented in these endo-phenotype GWASs. These included datasets from the Neale lab that used ICD-10 phenotypes: a heartbeat GWAS with 2,542 cases, PHESANT-based phenotype GWASs for heart arrhythmia with 2,013 cases or heart valve with 2,453 cases, and phecode-based GWASs for valve disorder with 4,239 cases and heart sounds with 1,049 cases (Table 1).

When choosing our base studies, we aimed to include a spectrum of CHD phenotypes that is broader than simply the most severe cases, which require medical and surgical intervention and are considered typical CHD "cases." This requires a delicate balance to capture phenotypes that are more likely to have a genetic component (valve disorders) rather than phenotypes that might have a significant environmental contribution (i.e., valve disorder secondary to infection). A significant number of individuals have subtle or mild cardiac developmental phenotypes that are clinically observed as a heart murmur (sound) or abnormalities in the cardiac conduction system that alter the heartbeat. These are often identified through a routine physical exam and represent the mildest aspect of the spectrum of congenital heart malformations and often do not require significant treatment. The "heart valve problem or heart murmur" endo-phenotype is clearly aligned with CHD because heart valve problems are a subset of structural CHD. The "heart arrhythmia" and "abnormalities of heartbeat" endo-phenotypes are important clinical phenotypes that often coexist with mild to severe CHD.[57,58] For example, families with inherited forms of CHD demonstrate variable penetrance of clinical endo-phenotypes where structural and conduction abnormalities

Table 2. Case-control PRSs for congenital heart-related phenotypes in individuals with CHD

GWAS used	SNV list	GWAS source	Best p-value threshold	# Of SNVs included in model	PRS R^2 (variance explained by PRS)	p-value of model fit	Empirical p-value of model fit
Heart valve	**all SNVs**	**Neale**	**4.15005×10^{-3}**	**2,760**	**0.0246**	**2.918×10^{-4}**	**7.899×10^{-3}**
Heart valve	**fetal cardiac**	**Neale**	**3.85005×10^{-3}**	**1,079**	**0.0238**	**3.634×10^{-4}**	**7.999×10^{-3}**
Heart sounds	**all SNVs**	**PheWeb**	**2.5005×10^{-4}**	**178**	**0.01782**	**0.002015**	**0.03630**
Heart arrhythmia	fetal cardiac	Neale	1.05005×10^{-3}	309	0.01154	0.01265	0.1614
Heart sounds	fetal cardiac	PheWeb	0.0671001	11,699	0.01162	0.01242	0.1701
Valve disorders	all SNVs	PheWeb	3.0005×10^{-4}	289	0.007156	0.04923	0.4314
Heartbeat	fetal cardiac	Neale	5.5005×10^{-4}	194	6.481×10^{-3}	0.06127	0.5227
Heart arrhythmia	all SNVs	Neale	1.0005×10^{-4}	95	6.379×10^{-3}	0.06350	0.5375
Valve disorders	fetal cardiac	PheWeb	4.35005×10^{-3}	1,293	0.005302	0.09035	0.6444
Heartbeat	all SNVs	Neale	0.0924501	35,865	4.047×10^{-3}	0.1388	0.7837

This table lists the PRS results generated under a case-control model using PRSice as described in material and methods. PRS with significant empirical p-values indicated with bolded text. The first four columns describe the UKBB phenotype used for the base data, which variants were used, the p-value threshold in the base data used by PRSice in the PRS model, and the number of variants included in the final PRS model. The last three columns indicate the variance in phenotype explained by the PRS model, the p-value of the PRS model's fit, and an empirical p-value of model fit generated by thresholding to account for overfitting that occurs in the PRSice algorithm. Empirical $p \leq 0.05$ is considered significant. Empirical $p > 0.05$ with uncorrected $p \leq 5 \times 10^{-3}$ is considered nominally significant. The R^2 values correspond to the uncorrected p-value and are thus inherently inflated. For more details, see PRSice documentation.

co-occur.[59,60] Some of these phenotypes were curated using ICD-10 codes, whereas others were gathered using an aggregation of self-reports and nurse reports. Use of self-reporting is well documented[54–56,61] and, like the endophenotype approach, allows us to systematically assess a large number of milder phenotypes. Our approach to our base data allowed us to capture a larger spectrum of mildly affected individuals who have a genetic predisposition toward congenital heart disorders.

Finally, we used GWASs that were limited to individuals of European ancestry to develop the PRSs because of the known challenges and non-transportability of multi-ethnic PRSs.[62] Because different genetic ancestries have different allele frequencies and LD structure, this important problem requires additional research and methodological development to develop a universal PRS. Our target data, obtained from the CHD subset of the GMKF dataset, consists of 711 cases and 711 pseudo-controls created from non-transmitted parental alleles (Figure 1B).[35–37] Because of the multi-ethnic nature of our GMKF target data, we divided our 711 families into 362 families where both parents had genetic ancestry clustering with samples from Europe and 349 families where one or both parents had genetic ancestry clustering with samples from elsewhere in the world (Figure 1C). The cases and pseudo-controls (see material and methods) from the 362 families with European genetic ancestry were used for testing our PRSs developed from our base studies in an independent target dataset.

Case-control PRSs for congenital heart defects

We started by developing PRSs using a case-control model of CHD. We built PRSs for the five European ancestry GWAS phenotypes using all SNVs shared between the base and target data as our input. We then built three additional PRSs using SNVs located in or within 10 kb of genes that were expressed in fetal cardiac tissue. We applied these six PRS models to the target data, generating individual-level polygenic scores for each individual in a subset of the GMFK data composed of trios of European ancestry.

The next step required validation of our PRS results. This can be accomplished by using an independent dataset to replicate results, performing cross-validation with a subset of the initial data, or by using permutation methods to generate an empirical p-value, as discussed in the PRSice-2 documentation. We decided to use PRSice to generate empirical p-values of model fit through permutation methods because we did not have access to a non-overlapping set of individual-level CHD, whole-genome sequences with which to validate our results and decided against reducing our statistical power by further subdividing our data into a training subset and testing subset. In our results (Tables 2 and 3), the p-value of model fit represents the initial, uncorrected p-value, whereas the empirical p-value represents the validated p-value. This empirical p-value corrects for the overfitting that naturally occurs when using parameter optimization to develop a PRS.[30,63] Empirical $p \leq 0.05$ was considered significant. Empirical $p > 0.05$ but with non-corrected $p \leq 5.00 \times 10^{-3}$ (correcting for 10 multiple tests) was considered nominally significant.

When calculating the PRS using all SNVs (Table 2), the heart valve phenotype reaches significance and explains 2.5% of the variance in the CHD phenotype (initial $p = 2.92 \times 10^{-4}$, empirical $p = 7.90 \times 10^{-3}$). Within our target GMKF dataset, the mean heart valve PRS is higher in controls compared with cases (Figure 2A). Next we calculated the odds ratio of this CHD phenotype across 10 PRS deciles and found that, in the highest decile, there was a significant decrease in CHD risk compared with the 50th percentile of CHD scores (odds ratio [OR] = 0.498; 95%

Table 3. Severity PRS for congenital heart related phenotypes in patients with CHD

GWAS used	SNV list	GWAS source	Best p-value threshold	No. of SNVs included in model	PRS R^2 (variance explained by PRS)	p-value of model fit	Empirical p-value of model fit
Heart valve	**fetal cardiac**	**Neale**	**3.85005×10^{-3}**	**1,079**	**0.01836**	**2.608×10^{-4}**	**6.199×10^{-3}**
Heart sounds	**all SNVs**	**PheWeb**	**2.5005×10^{-4}**	**178**	**0.01289**	**2.248×10^{-3}**	**0.03740**
Heart valve	**all SNVs**	**Neale**	**4.15005×10^{-3}**	**2,760**	**0.01594**	**6.726×10^{-4}**	**0.01510**
Heart arrhythmia	**fetal cardiac**	**Neale**	**1.40005×10^{-3}**	**412**	**0.01153**	**3.865×10^{-3}**	**0.05939**
Heart sounds	**fetal cardiac**	**PheWeb**	**0.0767001**	**13,023**	**0.01071**	**5.382×10^{-3}**	**0.08729**
Heartbeat	fetal cardiac	Neale	0.3073	36,773	5.253×10^{-3}	0.05157	0.4655
Valve disorders	all SNVs	PheWeb	3.0005×10^{-4}	289	4.187×10^{-3}	0.08229	0.5859
Valve disorders	fetal cardiac	PheWeb	4.35005×10^{-3}	1,293	4.337×10^{-3}	0.07698	0.5880
Heart arrhythmia	all SNVs	Neale	1.10005×10^{-3}	840	3.300×10^{-3}	0.1230	0.7677
Heartbeat	all SNVs	Neale	0.0923001	35,818	2.950×10^{-3}	0.1449	0.8069

This table lists the PRS results generated using a severity model using PRSice as described in material and methods. PRS with significant empirical p-values indicated with bolded text. The first four columns describe the UKBB phenotype used for the base data, which variants were used, the p-value threshold in the base data used by PRSice in the PRS model, and the number of variants included in the final PRS model. The last three columns indicate the variance in phenotype explained by the PRS model, the p-value of the PRS model's fit, and an empirical p-value of model fit generated by thresholding to account for overfitting that occurs in the PRSice algorithm. Empirical $p \leq 0.05$ is considered significant. Empirical $p > 0.05$ with uncorrected $p \leq 5 \times 10^{-3}$ is considered nominally significant. The R^2 values correspond to the uncorrected p-value and are thus inherently inflated. For more details, see PRSice documentation.

confidence interval [CI], 0.257–0.965) (Figure 2B; Table S5). This suggests a protective effect of common variants included in this PRS. To further verify these results, we looked at the distribution of beta values and minor allele frequencies for all effective SNVs included in our significant PRS score (Figure S2).

Next, because we were exploring a heart-specific phenotype, we sought to determine whether we could improve our PRSs by limiting to SNVs identified near 18,421 fetal cardiac genes identified through published RNA-seq analyses.[41] For the heart valve PRS, restriction of the fetal cardiac gene set showed a similar p-value and PRS R^2 despite including less than half of the SNVs from the all SNV PRS (1,079 SNVs) in the model (initial $p = 3.63 \times 10^{-4}$, empirical $p = 8.00 \times 10^{-3}$) (Table 2; Figure S5). Although the distribution of PRSs showed higher mean PRSs in the controls, and our empirical p-value used for validation stayed significant, the OR for the top PRS decile of the CHD phenotype using this approach did not replicate, suggesting that limiting our SNVs to those expressed in fetal cardiac tissue did not have sufficient power to capture additional CHD risk (Table S5). Our annotation of fetal cardiac transcripts was limited to a single snapshot in time and did not encompass the full developmental transcriptional time that would be important for cardiac development. The other two GWAS phenotypes we evaluated, heart arrhythmia and heartbeat, did not show any significant association in our case-control analysis.

We also sought to test the use of alternative phenotype classifications that are often observed in EHR-linked biobanks, phecodes (see material and methods), which are becoming increasingly common.[52,64] For phecodes, distinct ICD-10 codes are assigned to a common phecode to normalize differences across institutions in phenotype and diagnosis coding. We then calculated PRSs based on the GWAS summary statistics using phecodes for phenotype curation in the UKBB.[43] The phecodes for "abnormal heart sounds" include heart murmurs and are somewhat analogous to the "heart valve problem or heart murmur" problem or "heart murmur" phenotype GWASs in the Neale lab GWASs. Our results show that the abnormal heart sounds PRS using all SNVs in a case-control format gives similarly significant results (Table 2; Figure S6) (case-control initial $p = 2.01 \times 10^{-3}$, case-control empirical $p = 0.036$). As with the Neale lab heart valve PRS, we again see a significant protective effect in the PheWeb abnormal heart sounds PRS However, we do not see a significant effect with the PheWeb heart valve disorder PRS. These data demonstrate that the choice of phenotype and method of derivation of core phenotype is of critical importance and that inclusion of non-genetic etiology (e.g., valvular disorders caused by bacterial endocarditis) can dilute out a signal of common genetic contribution. These data drive home the point that the decision of which summary statistics to use greatly affects the resulting PRS score.

Finally, we tested the portability of these GWAS PRSs in individuals of non-European ancestry. One of the major struggles with PRSs is poor performance in individuals of non-European or admixed ancestry[62,65,66] because of lack of GWAS data in matching populations. Using the heart valve Neale lab UKBB GWAS (Table 1), we computed PRSs for the heart valve phenotype in the non-European dataset composed of 349 trios where one of more parents were of non-European ancestry (Figure S4; Table S6). In our analysis, neither of the PRS models using all SNVs or fetal cardiac SNVs showed significance (Table S4) in the non-European population of GKMF CHD cases. Our results highlight the importance of well-powered and matched

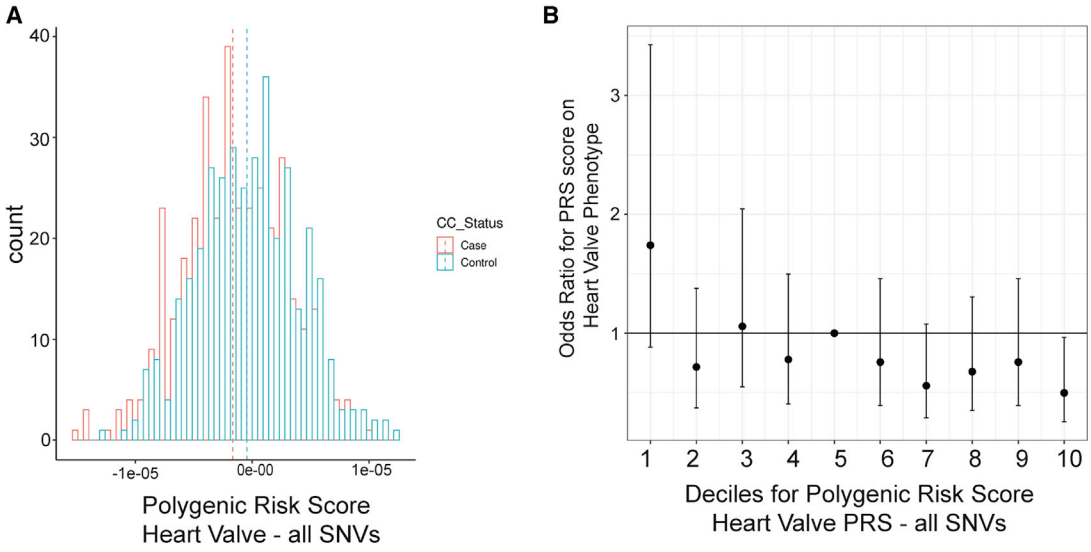

Figure 2. Top decile of the case-control PRS (heart valve, all SNVs) has decreased odds for a CHD classification
(A) Histogram of the PRS distribution of cases and controls, showing that the mean PRS of controls (-2.26×10^{-7}) is greater than that of cases (-8.33×10^{-7}).
(B) Odds ratios (ORs) and confidence intervals (CIs) of PRS deciles, calculated via PRSice, showing that the top decile of PRSs is significantly lower than the 50% decile (OR = 0.498, 95% CI = 0.257–0.965).
See also Table S4.

genetic ancestry to obtain meaningful results, particularly for rare genetic disorders where the study population is often smaller and underpowered.

Severity PRSs for congenital heart defects

We next developed PRSs using a quantitative severity model of CHD to explore whether increased PRSs were associated with severity of the CHD phenotype (Table 3). GKMF CHD individuals were classified as having mild, moderate, or severe complexity in accordance with 2001 guidance[49] by a specialist in adult CHD. Phenotype severity was coded as follows: 0 for unaffected pseudo-controls, 1 for mild, 2 for moderate, and 3 for severe cases of CHD. We then ran the PRSs on these four groups across the same five GWAS studies and used empirical p-values to validate our PRS results as before. We performed a total of 10 PRSs (using five phenotypes and two SNV sets). Tests with empirical $p \leq 0.05$ were considered significant, and tests with empirical $p > 0.05$ but initial p of model fit $\leq 8.33 \times 10^{-3}$ were considered nominally significant.

The severity PRS results were similar to the results found in our case-control PRSs for CHD. Across the four groups of increasingly complex CHD phenotypes (0–3), the PRS built from the heart valve GWAS with fetal cardiac SNVs was significant (initial $p = 2.57 \times 10^{-4}$, empirical $p = 5.90 \times 10^{-3}$) (Table 3; Figure 3C), and the mean PRS decreased with increasing severity, suggesting that the PRS is partially driven by variants that are protective of increasingly severe CHD phenotypes (Figures 3A and 3B) (initial $p = 2.61 \times 10^{-4}$, empirical $p = 6.20 \times 10^{-3}$) (Table 3). The heart valve phenotype using all SNVs(initial $p = 6.73 \times 10^{-4}$, empirical $p = 0.015$) also produced significant results (Table 3; Figure 3D), consistent with our results of the case-control PRS for the heart valve phenotype using all SNVs (Figure 2B).

When reviewing the quantitative severity model of CHD PRSs, we found that heart arrhythmia GWAS using fetal cardiac SNVs showed a nominally significant difference between cases and controls (initial $p = 3.86 \times 10^{-3}$, empirical $p = 0.0594$). The direction of the effect was opposite to that observed in the heart valve PRS (Figure S5), indicating that the variants in this score increase risk of CHD severity instead of providing a protective effect like in the other phenotypes. We performed a similar severity analysis based on the PheWeb datasets and found that the abnormal heart sounds PRS showed significant cumulative protection from CHD phenotypes (severity initial $p = 2.25 \times 10^{-3}$, severity empirical $p = 0.0374$). The severity of abnormal heart sounds PRS using fetal cardiac SNVs was nominally significant (initial $p = 5.38 \times 10^{-3}$, empirical $p = 0.0873$) (Table 3; Figure S6).Our analysis is one of the first to quantify the contribution of common genetic variants to a rare disease, congenital heart defects, using PRSs. These results demonstrate the potential utility to quantify the effect of genetic background on risk of CHD and modification of disease severity.

Variant and gene analysis of PRS

To further investigate these results, we looked at the variants included in each of the significant PRS (Table S7). We also looked at how many of these variants were within 10 kb of known CHD genes or fetal cardiac genes, as described in the Material and methods (Table S8), and which variants contain any expression quantitative trait loci (eQTLs) with $p < 0.05$ at that SNP (for any alternative allele) in the GTEX v.8 “heart atrial appendage” or “heart

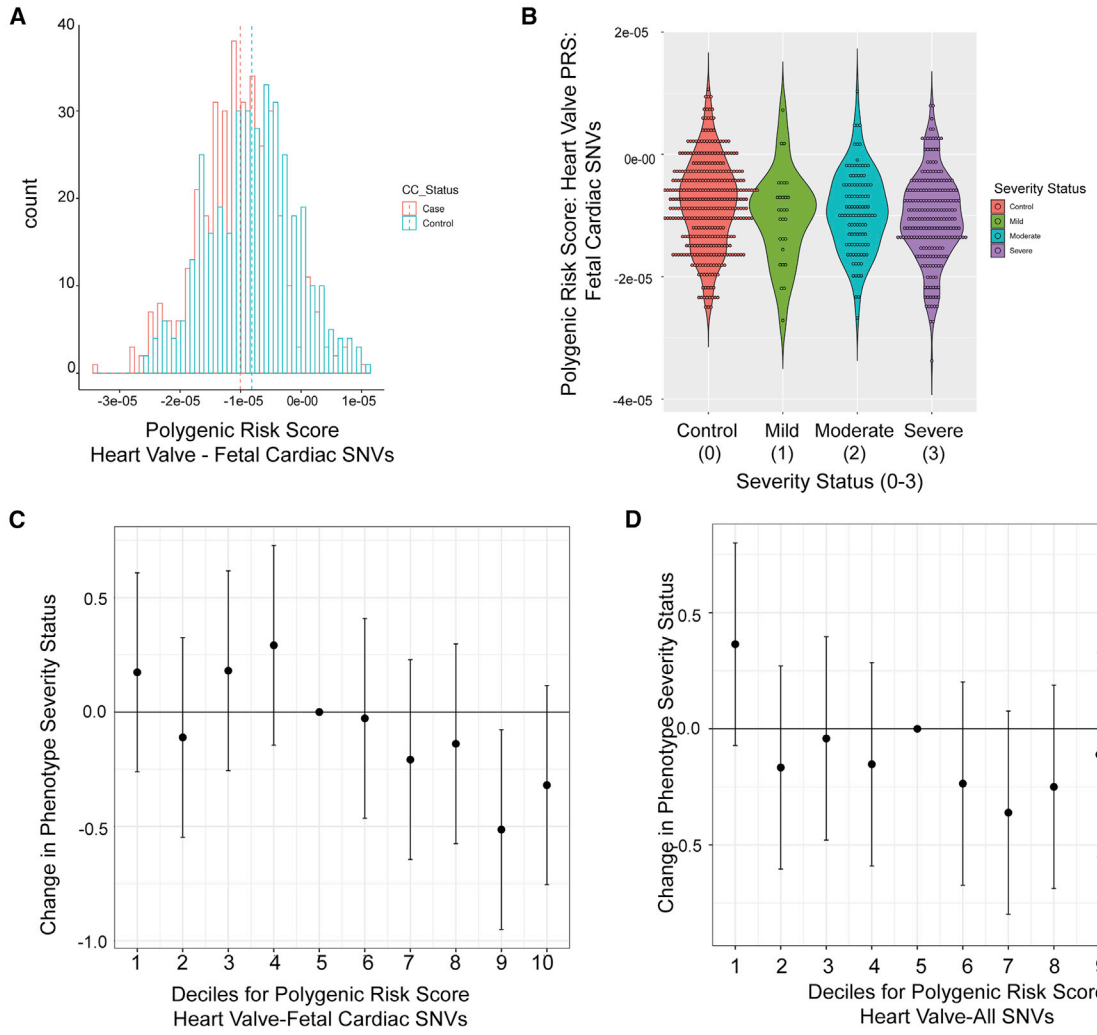

Figure 3. Increasing severity of CHD is associated with decreasing PRS (heart valve, fetal cardiac SNVs, and heart valve, all SNVs)
(A) Histogram of the PRS distribution of cases and controls, showing that the mean PRS of controls (-4.07×10^{-6}) is greater than that of cases (-5.03×10^{-6}).
(B) Violin plot showing distribution of PRSs of control individuals and individuals with mild, moderate, and severe disease. The correlation between the PRS and the severity status is −0.117 using Spearman's correlation and ordinals where control = 0, mild = 1, moderate = 2, and severe = 3.
(C) Change in severity phenotype (heart valve, fetal cardiac SNVs) given PRS increases over deciles (represented by linear regression coefficient and associated confidence intervals [CIs]).
(D) Change in severity phenotype (heart valve, all SNVs) PRS increases over deciles (represented by linear regression coefficient and associated CIs).

left ventricle" datasets (Table S7). We can see that even when switching from a case-control model to a severity model of PRS, we generally keep the same variants in the PRS. We also see that, for the heart valve PRS using fetal cardiac SNVs, 21 of 1,079 (1.9%) SNVs are in or within 10 kb of known CHD genes (Table S3). For the heart valve PRS using all SNVs, 30 of 2,760 (1.1%) of the SNVs are in or within 10 kb of known CHD genes. These 30 SNVs overlap most of the same 21 CHD SNVs from the fetal cardiac PRS. Our analysis suggests that there may be additional long-range effects of SNVs or still undiscovered genes associated with modulation of cardiac development phenotypes.

One interesting variant observed in the heart valve PRS is at chr9:136519087 and falls within the *NOTCH1* gene (MIM: 190198). Rare mutations in *NOTCH1* have been known to cause bicuspid aortic valve, a common type of isolated CHD.[13,67,68] The GWAS summary statistics show that this variant has a p-value of 1.90×10^{-3} and a beta value of -8.89×10^{-4}, indicating that it has a protective effect. This SNV has a relatively high minor allele frequency of 0.133. This suggests that relatively common alleles may dysregulate established monogenic disease genes, a finding that has been observed previously in GWASs.[29] We also identified two SNVs in *PRDM16* (MIM: 605557), a gene associated with congenital cardiomyopathy.[69] These data demonstrate that most of the SNVs in these genes are not necessarily associated with monogenic forms of disease, consistent with other GWASs exploring

the spectrum between rare and common disorders. For a full list of variants and genes included in the significant and nominally significant PRSs, see Table S7.

Finally, in Table S9, we list all individual PRS scores for all six of the significant PRSs listed in Tables 2 and 3 (heart valve case-control fetal SNVs, all heart sounds case-control SNVs, all heart valve case-control SNVs, heart valve severity fetal SNVs, all heart sounds severity SNVs, and all heart valve severity SNVs). This table will allow future researchers to better utilize our results. This will also enable future analyses involving only specific types of CHD when combined with Table S4.

Overall, our study has identified several PRSs that are significantly associated and account for up to 2.5% of the genetic variance observed in cases of CHD. The importance of phenotype curation within these large EHRs is highlighted by the challenge of obtaining sufficient, high-quality phenotypes for the base GWAS study. Our study demonstrates the feasibility of leveraging large-scale GWASs from biobanks such as the UKBB and highlights the importance of development phenotype curation methods to improve reproducibility across cohorts. We used different methods for phenotype curation, ICD code based and self-reported, to demonstrate how small changes in phenotype curation can affect PRS results. Overall, our study suggests that self-reported/nurse-reported datasets, such as those used in the Neale lab GWAS, are cleaner and have more significant findings compared with phecodes.

Discussion

Our work investigates the common variant genetic basis of CHD by computing PRSs for CHD risk using GWASs from UKBB endo-phenotypes. Phenotype-matched approaches work to link GWAS significant loci with phenotype-matched monogenic disease genes to further fine-mapping efforts.[29] We use the same logic for our investigation of CHD, where phenotypes that capture mild representations of a broad class of malformations are leveraged to gain insight into the less accessible and more severe CHD. To our knowledge, these are some of the first PRSs developed to explain some of the variance in CHD phenotype severity. We use phenotype-matched GWASs for common diseases, such as "heart valve problem/heart murmur," which is an audible sound made by non-laminar turbulent blood flow in the heart's chambers often caused by cardiac valve anomalies. This audible murmur can be innocuous and have little cause for clinical concern or can represent a clinically relevant CHD requiring treatment. Using mild manifestations of congenital anomalies, we can quantify the effect of common genetic background in rare diseases such as CHD.

Interest in PRS to assess clinical risk for complex disorders has gained traction over the past several years, with PRSs having been developed for cancer risk,[70–72] CAD,[22] and diabetes.[22,73] One area that has had limited uptake of PRSs is a field that has a long history of genetic and genomic testing for rare congenital syndromes: clinical genetics. The classic population in medical genetics has been children with rare, phenotypically classifiable, genetic syndromes caused by single-gene mutations. Given the rarity of these types of congenital phenotypes, GWASs have been few and limited in scope until the more recent releases of EHR-linked genomic biobanks such as the UKBB[50] and large-cohort sequencing efforts such as the Gabriella Miller Foundation[31] and the UK Deciphering Developmental Delay studies.[74]

Taking advantage of the rapid generation and sharing of these large-scale genomic datasets, we used, as our base studies, GWASs performed in the UKBB with structural heart-related phenotypes and used GMKF WGS data from a variety of CHD phenotypes to test the PRS model. PRSs estimated using the heart valve problem or heart murmur GWAS explain 2.5% of variance in case-control status of CHD and 1.8% of variance in severity of CHD. This shows that cumulative, common genetic variants affect the variable penetrance and expressivity observed in large family studies of CHD. From the standpoint of rare disease genetics, we can consider PRS to be a "modifier" in cases with unidentified primary genetic or environmental etiology. In our analysis, the effective (alternate) alleles in the heart valve and heart sounds PRSs showed an overall protective effect from CHD risk and severity. Our results show that small differences in severity can be partly explained using common genetic variants.

Improvements of methods for phenotyping mild CHD phenotypes across the UKBB would allow generation of more powerful PRSs for CHD. We identified several well-powered GWASs for CHD-related phenotypes outside of the UKBB, but none released summary statistics data. As more EHR-linked biobanks come online, it will be feasible to assemble CHD cohorts with sufficient power to include larger cross-validation sets and improve the PRS developed here. Our data demonstrates that the choice of base GWAS phenotype is critically important for identifying accurate and robust contributions of common genetic variants to CHD risk. Although we found significant associations of PRS with CHD for the self-reported phenotype "heart valve problem/heart murmur," we were surprised to see that, in the phecode PRS, only the heart sounds phenotype remained significant, whereas the valve disorders phenotype was not significant in our analysis. We believe that valve disorders, which had nearly twice the number of cases as the other PRS GWASs, likely included non-genetic etiologies of heart valve dysfunction that were secondary to infection or other cardiac diseases. Ultimately, improved phenotyping in the base and target validation studies will be required to improve the precision of the PRS and add clinical utility of PRSs for individuals with rare genetic diseases.

Currently, one major limitation of nearly all PRS models is that they are less accurate in individuals with non-European ancestry.[75–77] Genetic datasets over the past 15 years have been overwhelmingly Euro-centric, and cumulative data have shown that only 19% of all genomes sequenced has been from individuals with non-European ancestry[76] despite

the fact that non-Europeans make up 84% of the global population. The variants identified in GWASs used to build PRSs, including our PRSs, are focused on the UKBB European population because of the power and size of the base-study population. Given the rarity of the phenotypes we are testing, having tools and methods that can take into account admixture and diverse ancestral populations would improve the power and accuracy of our genetic models for taking into account the effects of genetic background on congenital phenotypes. The necessity of diversifying large-scale genetic studies such as GWASs is critical, given that European-centric GWASs and PRSs are not importable into non-European datasets (Figure S4).[62,76] Our focus in this work was to develop some of the first PRSs for a rare congenital disorder. However, more work remains to refine the base-data GWAS and to improve multi-ethnic PRS approaches to account for admixture in human populations.

Although our focus is on the genetic contributions, we acknowledge that complex patterning events within human development are also influenced by maternal and environmental factors. PRSs do not account for the interaction between genetic factors and environment or the role of rare genetic variants. Previous work studying the genetic architecture of CHD has focused on rare and *de novo* variants within the protein coding genome. However, a recent publication has explored the role of rare *de novo* variants in the same GKMF dataset[9] and identified several *de novo* non-coding variants controlling expression of cardiac development genes or binding of specific gene-regulatory factors. Our study complements the work in rare variants by focusing exclusively on the genetic background and common variant contribution to CHD. Ultimately, integrating the relative contributions of rare and common non-coding variants, alongside copy number variants and other forms of genomic variation, will yield stronger predictors of CHD risk and severity.

Clinical utility of PRSs remains limited because of the low relative risk predicted and heterogeneity of scores among individuals of the same phenotype. However, continuing studies are exploring clinical scenarios where these scores might provide added benefit to existing algorithms that rely on serum and clinical diagnostic markers[78,79] or where they can be used to understand the phenotypic variability for individuals with a high-risk monogenic predisposition.[80–82] Despite the many challenges in the field of rare disease genetics and PRSs in general, we believe that PRSs can contribute significantly to our understanding of the genetic architecture of rare, understudied disorders like CHD, and we hope that, as PRS methods continue to improve, additional insight into and knowledge of the causes of CHD can be used to improve long-term outcomes in affected individuals.

In this work, we successfully used PRS to find a significant association of the cumulative effects of common genetic variants with CHD. This is one of the first publications to explore the role of PRS in this rare congenital disease, where the GWAS is not limited to a narrow subset of individuals with CHD. We found common genetic variants to not only be associated with the case-control status of CHD but also with severity of the disease. We also identified GWAS phenotypes that exist on the spectrum of CHD and therefore can be used to build a PRS for CHD when there are no high-quality, matched GWAS data. As more genetic data become available and linked to clinical phenotypes, our ability to quantify risk associated with multiple variants through polygenic risk[83] will only improve. In addition, by further elucidating the genetic basis of CHD, this research contributes to the groundwork for new discoveries in treating CHD and its related health problems.

Statement of ethics

Procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national), and proper informed consent was obtained.

Data and code availability

The key code for this project is at https://github.com/Spendlove/congenital_heart_disease_PRS. All datasets and summary data will be made available. Raw and summary datasets used in this study can be found at the following locations: Neale lab UKBB GWAS (http://www.nealelab.is/uk-biobank), GMKF CHD cohort (dbGAP: phs001138.v3.p2, phs001138.v4.p2, https://kidsfirstdrc.org), and PheWeb GWAS (https://pheweb.org/UKB-SAIGE/). GTEX data used for the analyses described in this manuscript were obtained from the eQTL catalog (https://www.ebi.ac.uk/eqtl/Data_access/) on February 7, 2022.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2022.100112.

Acknowledgments

This work was funded by DP5OD024579 (to V.A.A.) from NIH Common Fund and R03HL150604 from NHLBI (to V.A.A. and J.H.S.). S.S. and L.B. were supported by NHGRI T32 T32HG002536 (2020–2022). L.B. was supported by T32M012424 (2019–2020). We thank Gabriella Miller's Kids First Foundation for generating these datasets and all affected individuals and families and the physicians who have worked hard to care for them. Schematic overview figures were created with BioRender. We also thank Darren Lin and Angela Wei for help with curating genetic data. The Genotype-Tissue Expression (GTEX) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. https://www.ebi.ac.uk/eqtl/Data_access/

Author contributions

S.S., V.A.A., and J.H.S. wrote the paper. S.S., L.B., and J.H.S. performed data analyses with input from G.L. and V.A.A. All authors contributed to data interpretation.

Declaration of interests

Dr. Sul is an employee of Merck as of April 2021.

Received: December 13, 2021
Accepted: April 19, 2022

References

1. van der Linde, D., Konings, E.E.M., Slager, M.A., Witsenburg, M., Helbing, W.A., Takkenberg, J.J.M., and Roos-Hesselink, J.W. (2011). Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. J. Am. Coll. Cardiol. *58*, 2241–2247. https://doi.org/10.1016/j.jacc.2011.08.025.
2. Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J., Adair, T., Aggarwal, R., Ahn, S.Y., et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet *380*, 2095–2128. https://doi.org/10.1016/S0140-6736(12)61728-0.
3. Jin, S.C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., DePalma, S.R., Zeng, X., Qi, H., Chang, W., Sierant, M.C., et al. (2017). Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nat. Genet. *49*, 1593–1601. https://doi.org/10.1038/ng.3970.
4. Homsy, J., Zaidi, S., Shen, Y., Ware, J.S., Samocha, K.E., Karczewski, K.J., DePalma, S.R., McKean, D., Wakimoto, H., Gorham, J., et al. (2015). De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science *350*, 1262–1266. https://doi.org/10.1126/science.aac9396.
5. Miller, D.B., and Piccolo, S.R. (2021). A survey of compound heterozygous variants in pediatric cancers and structural birth defects. Front. Genet. *12*, 640242. https://doi.org/10.3389/fgene.2021.640242.
6. Bolkier, Y., Barel, O., Marek-Yagel, D., Atias-Varon, D., Kagan, M., Vardi, A., Mishali, D., Katz, U., Salem, Y., Tirosh-Wagner, T., et al. (2021). Whole-exome sequencing reveals a monogenic cause in 56% of individuals with laterality disorders and associated congenital heart defects. J. Med. Genet. https://doi.org/10.1136/jmedgenet-2021-107775.
7. Diab, N.S., Barish, S., Dong, W., Zhao, S., Allington, G., Yu, X., Kahle, K.T., Brueckner, M., and Jin, S.C. (2021). Molecular genetics and complex inheritance of congenital heart disease. Genes *12*, 1020.
8. Watkins, W.S., Hernandez, E.J., Wesolowski, S., Bisgrove, B.W., Sunderland, R.T., Lin, E., Lemmon, G., Demarest, B.L., Miller, T.A., Bernstein, D., et al. (2019). De novo and recessive forms of congenital heart disease have distinct genetic and phenotypic landscapes. Nat. Commun. *10*, 4722. https://doi.org/10.1038/s41467-019-12582-y.
9. Richter, F., Morton, S.U., Kim, S.W., Kitaygorodsky, A., Wasson, L.K., Chen, K.M., Zhou, J., Qi, H., Patel, N., DePalma, S.R., et al. (2020). Genomic analyses implicate noncoding de novo variants in congenital heart disease. Nat. Genet. *52*, 769–777. https://doi.org/10.1038/s41588-020-0652-z.
10. Gross, A.M., Ajay, S.S., Rajan, V., Brown, C., Bluske, K., Burns, N.J., Chawla, A., Coffey, A.J., Malhotra, A., Scocchia, A., et al. (2019). Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. Genet. Med. *21*, 1121–1130. https://doi.org/10.1038/s41436-018-0295-y.
11. Hiatt, S.M., Lawlor, J.M.J., Handley, L.H., Ramaker, R.C., Rogers, B.B., Partridge, C., Partridge, E.C., Boston, L.B., Williams, M., Plott, C.B., et al. (2021). Long-read genome sequencing for the molecular diagnosis of neurodevelopmental disorders. HGG Adv. *132*, S274. https://doi.org/10.1016/s1096-7192(21)00504-7.
12. Atemin, S., Todorov, T., Maver, A., Chamova, T., Georgieva, B., Tincheva, S., Pacheva, I., Ivanov, I., Taneva, A., Zlatareva, D., et al. (2021). MYH7-related disorders in two Bulgarian families: novel variants in the same region associated with different clinical manifestation and disease penetrance. Neuromuscul. Disord. *31*, 633–641. https://doi.org/10.1016/j.nmd.2021.04.004.
13. Roifman, M., Chung, B.H.Y., Reid, D.M., Teitelbaum, R., Martin, N., Nield, L.E., Thompson, M., Shannon, P., and Chitayat, D. (2021). Heterozygous *NOTCH1* deletion associated with variable congenital heart defects. Clin. Genet. *99*, 836–841. https://doi.org/10.1111/cge.13948.
14. Fahed, A.C., Gelb, B.D., Seidman, J.G., and Seidman, C.E. (2013). Genetics of congenital heart disease: the glass half empty. Circ. Res. *112*, 707–720. https://doi.org/10.1161/circresaha.112.300853.
15. Prendiville, T., Jay, P.Y., and Pu, W.T. (2014). Insights into the genetic structure of congenital heart disease from human and murine studies on monogenic disorders. Cold Spring Harb. Perspect. Med. *4*, a013946. https://doi.org/10.1101/cshperspect.a013946.
16. Oyen, N., Poulsen, G., Boyd, H.A., Wohlfahrt, J., Jensen, P.K., and Melbye, M. (2009). Recurrence of congenital heart defects in families. Circulation *120*, 295–301. https://doi.org/10.1161/circulationaha.109.857987.
17. Arboleda, V.A., Fleming, A., Barseghyan, H., Délot, E., Sinsheimer, J.S., and Vilain, E. (2014). Regulation of sex determination in mice by a non-coding genomic region. Genetics *197*, 885–897. https://doi.org/10.1534/genetics.113.160259.
18. del Pilar Jiménez-A, M., Viriyakosol, S., Walls, L., Datta, S.K., Kirkland, T., Heinsbroek, S.E.M., Brown, G., and Fierer, J. (2008). Susceptibility to Coccidioides species in C57BL/6 mice is associated with expression of a truncated splice variant of Dectin-1 (Clec7a). Genes Immun. *9*, 338–348. https://doi.org/10.1038/gene.2008.23.
19. Moss, D.J.H., Pardinas, A.F., Langbehn, D., Lo, K., Leavitt, B.R., Roos, R., Durr, A., Mead, S., Holmans, P., Jones, L., et al. (2017). Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. Lancet Neurol. *16*, 701–711. https://doi.org/10.1016/S1474-4422(17)30161-8.
20. Lewis, C.M., and Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. Genome Med. *12*, 44. https://doi.org/10.1186/s13073-020-00742-5.
21. Thakarakkattil Narayanan Nair, A., Donnelly, L.A., Dawed, A.Y., Gan, S., Anjana, R.M., Viswanathan, M., Palmer, C.N.A., and Pearson, E.R. (2020). The impact of phenotype, ethnicity and genotype on progression of type 2 diabetes mellitus. Endocrinol. Diabetes Metab. *3*, e00108. https://doi.org/10.1002/edm2.108.
22. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224. https://doi.org/10.1038/s41588-018-0183-z.

23. You, C., Zhou, Z., Wen, J., Li, Y., Pang, C.H., Du, H., Wang, Z., Zhou, X.-H., King, D.A., Liu, C.-T., and Huang, J. (2021). Polygenic scores and parental predictors: an adult height study based on the United Kingdom biobank and the Framingham heart study. Front. Genet. *12*, 669441. https://doi.org/10.3389/fgene.2021.669441.
24. Nikpay, M., Goel, A., Won, H.H., Hall, L.M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C.P., Hopewell, J.C., et al. (2015). A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat. Genet. *47*, 1121–1130. https://doi.org/10.1038/ng.3396.
25. Cordell, H.J., Bentham, J., Topf, A., Zelenika, D., Heath, S., Mamasoula, C., Cosgrove, C., Blue, G., Granados-Riveron, J., Setchfield, K., et al. (2013). Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16. Nat. Genet. *45*, 822–824. https://doi.org/10.1038/ng.2637.
26. Xu, J., Lin, Y., Si, L., Jin, G., Dai, J., Wang, C., Chen, J., Da, M., Hu, Y., Yi, C., et al. (2014). Genetic variants at 10p11 confer risk of Tetralogy of Fallot in Chinese of nanjing. PLoS One *9*, e89636. https://doi.org/10.1371/journal.pone.0089636.
27. Trevino, C.E., Holleman, A.M., Corbitt, H., Maslen, C.L., Rosser, T.C., Cutler, D.J., Johnston, H.R., Rambo-Martin, B.L., Oberoi, J., Dooley, K.J., et al. (2020). Identifying genetic factors that contribute to the increased risk of congenital heart defects in infants with Down syndrome. Sci. Rep. *10*, 18051. https://doi.org/10.1038/s41598-020-74650-4.
28. Škorić-Milosavljević, D., Tadros, R., Bosada, F.M., Tessadori, F., van Weerd, J.H., Woudstra, O.I., Tjong, F.V.Y., Lahrouchi, N., Bajolle, F., Cordell, H.J., et al. (2022). Common genetic variants contribute to risk of transposition of the great arteries. Circ. Res. *130*, 166–180. https://doi.org/10.1161/CIRCRESAHA.120.317107.
29. Freund, M.K., Burch, K.S., Shi, H., Mancuso, N., Kichaev, G., Garske, K.M., Pan, D.Z., Miao, Z., Mohlke, K.L., Laakso, M., et al. (2018). Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits. Am. J. Hum. Genet. *103*, 535–552. https://doi.org/10.1016/j.ajhg.2018.08.017.
30. Choi, S.W., and O'Reilly, P.F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience *8*, giz082. https://doi.org/10.1093/gigascience/giz082.
31. Gabriella Miller Kids First (2015). Retrieved from https://kidsfirstdrc.org.
32. 2019. 2018 X01 Projects.
33. Li, B., Chen, W., Zhan, X., Busonero, F., Sanna, S., Sidore, C., Cucca, F., Kang, H.M., and Abecasis, G.R. (2012). A likelihood-based framework for variant calling and de novo mutation detection in families. PLoS Genet. *8*, e1002944. https://doi.org/10.1371/journal.pgen.1002944.
34. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.
35. Jansen, A.G., Dieleman, G.C., Jansen, P.R., Verhulst, F.C., Posthuma, D., and Polderman, T.J.C. (2020). Psychiatric polygenic risk scores as predictor for attention deficit/hyperactivity disorder and autism spectrum disorder in a clinical child and adolescent sample. Behav. Genet. *50*, 203–212. https://doi.org/10.1007/s10519-019-09965-8.
36. Sokolowski, M., Wasserman, J., and Wasserman, D. (2016). Polygenic associations of neurodevelopmental genes in suicide attempt. Mol. Psychiatry *21*, 1381–1390. https://doi.org/10.1038/mp.2015.187.
37. Bates, T.C., Maher, B.S., Medland, S.E., McAloney, K., Wright, M.J., Hansell, N.K., Kendler, K.S., Martin, N.G., and Gillespie, N.A. (2018). The nature of nurture: using a virtual-parent design to test parenting effects on children's educational attainment in genotyped families. Twin Res. Hum. Genet. *21*, 73–83. https://doi.org/10.1017/thg.2018.11.
38. Cordell, H.J., Barratt, B.J., and Clayton, D.G. (2004). Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. Genet. Epidemiol. *26*, 167–185. https://doi.org/10.1002/gepi.10307.
39. Wang, L., Zhang, W., and Li, Q. EIGENSTRAT for Correcting for Population Stratification.
40. Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Toneva, I., Vaughan, B., Leinonen, R., Shumway, M., Flicek, P., Shumway, M., et al. (2012). The 1000 Genomes Project: data management and community access. Nat. Methods *9*, 459–462. https://doi.org/10.1038/nmeth.1974.
41. Pervolaraki, E., Dachtler, J., Anderson, R.A., and Holden, A.V. (2018). The developmental transcriptome of the human heart. Sci. Rep. *8*, 15362. https://doi.org/10.1038/s41598-018-33837-6.
42. Neale Lab (2018). UK Biobank GWAS round 2. Retrieved from. http://www.nealelab.is/uk-biobank.
43. Gagliano Taliun, S.A., VandeHaar, P., Boughton, A.P., Welch, R.P., Taliun, D., Schmidt, E.M., Zhou, W., Nielsen, J.B., Willer, C.J., Lee, S., et al. (2020). Exploring and visualizing large-scale genetic associations by using PheWeb. Nat. Genet. *52*, 550–552. https://doi.org/10.1038/s41588-020-0622-5.
44. Millard, L.A.C., Davies, N.M., Gaunt, T.R., Davey Smith, G., and Tilling, K. (2018). Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. Int. J. Epidemiol. *47*, 29–35. https://doi.org/10.1093/ije/dyx204.
45. Wei, W.Q., Bastarache, L.A., Carroll, R.J., Marlo, J.E., Osterman, T.J., Gamazon, E.R., Cox, N.J., Roden, D.M., and Denny, J.C. (2017). Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLoS One *12*, e0175508. https://doi.org/10.1371/journal.pone.0175508.
46. Bastarache, L., Hughey, J.J., Hebbring, S., Marlo, J., Zhao, W., Ho, W.T., Van Driest, S.L., McGregor, T.L., Mosley, J.D., Wells, Q.S., et al. (2018). Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science *359*, 1233–1239. https://doi.org/10.1126/science.aal4043.
47. Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll, R., Bastarache, L., Denny, J.C., et al. (2019). Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. JMIR Med. Inform. *7*, e14325. https://doi.org/10.2196/14325.
48. Peyrot, W.J., Boomsma, D.I., Penninx, B.W.J.H., and Wray, N.R. (2016). Disease and polygenic architecture: avoid trio design and appropriately account for unscreened control subjects for common disease. Am. J. Hum. Genet. *98*, 382–391. https://doi.org/10.1016/j.ajhg.2015.12.017.

49. Warnes, C.A., Liberthson, R., Danielson, G.K., Dore, A., Harris, L., Hoffman, J.I., Somerville, J., Williams, R.G., and Webb, G.D. (2001). Task force 1: the changing profile of congenital heart disease in adult life. J. Am. Coll. Cardiol. *37*, 1170–1175. https://doi.org/10.1016/s0735-1097(01)01272-4.
50. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779. https://doi.org/10.1371/journal.pmed.1001779.
51. Baysal, B.E. (2013). Natural selection increases mutational robustness in complex diseases: Mendelian evidence from early versus late onset common diseases. PeerJ. Preprint at.
52. Johnson, R.D., Ding, Y., and Venkateswaran, V. (2021). Leveraging genomic diversity for discovery in an EHR-linked biobank: the UCLA ATLAS Community Health Initiative. medRxiv. . Preprint at. https://doi.org/10.1101/2021.09.22.21263987.
53. Bastarache, L., Hughey, J.J., Goldstein, J.A., Bastraache, J.A., Das, S., Zaki, N.C., Zeng, C., Tang, L.A., Roden, D.M., and Denny, J.C. (2019). Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. J. Am. Med. Inform. Assoc. *26*, 1437–1447. https://doi.org/10.1093/jamia/ocz179.
54. Wang, R.T., Silverstein Fadlon, C.A., Ulm, J.W., Jankovic, I., Eskin, A., Lu, A., Rangel Miller, V., Cantor, R.M., Li, N., Elashoff, R., et al. (2014). Online self-report data for duchenne muscular dystrophy confirms natural history and can be used to assess for therapeutic benefits. PLoS Curr. *6*. https://doi.org/10.1371/currents.md.e1e8f2be7c949f9ffe81ec6fca1cce6a.
55. Kennedy, J., Goudie, D., Blair, E., Chandler, K., Joss, S., McKay, V., Green, A., Armstrong, R., Lees, M., Kamien, B., et al. (2019). KAT6A Syndrome: genotype-phenotype correlation in 76 patients with pathogenic KAT6A variants. Genet. Med. *21*, 850–860. https://doi.org/10.1038/s41436-018-0259-2.
56. Hu, Y., Shmygelska, A., Tran, D., Eriksson, N., Tung, J.Y., and Hinds, D.A. (2016). GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. Nat. Commun. *7*, 10448. https://doi.org/10.1038/ncomms10448.
57. Drakopoulou, M., Nashat, H., Kempny, A., Alonso-Gonzalez, R., Swan, L., Wort, S.J., Price, L.C., McCabe, C., Wong, T., Gatzoulis, M.A., et al. (2018). Arrhythmias in adult patients with congenital heart disease and pulmonary arterial hypertension. Heart *104*, 1963–1969. https://doi.org/10.1136/heartjnl-2017-312881.
58. Walsh, E.P., and Cecchin, F. (2007). Arrhythmias in adult patients with congenital heart disease. Circulation *115*, 534–545. https://doi.org/10.1161/circulationaha.105.592410.
59. Ellesøe, S.G., Johansen, M.M., Bjerre, J.V., Hjortdal, V.E., Brunak, S., and Larsen, L.A. (2016). Familial atrial septal defect and sudden cardiac death: identification of a novel NKX2-5 mutation and a review of the literature. Congenit. Heart Dis. *11*, 283–290. https://doi.org/10.1111/chd.12317.
60. Glancy, D.L., Wilklow, F.E., Devarapalli, S.K., Subramaniam, P.N., Moustoukas, N.M., and Kukuy, E. (2007). Sequential heart murmurs in a 43-year-old man with congenital heart disease. Proc (Bayl. Univ. Med. Cent.) *20*, 406–407. https://doi.org/10.1080/08998280.2007.11928335.
61. Winslow, A.R., Hyde, C.L., Wilk, J.B., Eriksson, N., Cannon, P., Miller, M.R., and Hirst, W.D. (2018). Self-report data as a tool for subtype identification in genetically-defined Parkinson's Disease. Sci. Rep. *8*, 12992. https://doi.org/10.1038/s41598-018-30843-6.
62. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. Am. J. Hum. Genet. *100*, 635–649. https://doi.org/10.1016/j.ajhg.2017.03.004.
63. Choi, S.W., Mak, T.S.-H., and O'Reilly, P.F. (2020). Tutorial: a guide to performing polygenic risk score analyses. Nat. Protoc. *15*, 2759–2772. https://doi.org/10.1038/s41596-020-0353-1.
64. Belbin, G.M., Cullina, S., Wenric, S., Soper, E.R., Glicksberg, B.S., Torre, D., Moscati, A., Wojcik, G.L., Shemirani, R., Beckmann, N.D., et al. (2021). Toward a fine-scale population health monitoring system. Cell *184*, 2068–2083.e11. https://doi.org/10.1016/j.cell.2021.03.034.
65. Cavazos, T.B., and Witte, J.S. (2021). Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. HGG Adv. *2*, 100017. https://doi.org/10.1016/j.xhgg.2020.100017.
66. Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., and Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. Nat. Commun. *10*, 3328. https://doi.org/10.1038/s41467-019-11112-0.
67. McKellar, S.H., Tester, D.J., Yagubyan, M., Majumdar, R., Ackerman, M.J., and Sundt, T.M., 3rd. (2007). Novel NOTCH1 mutations in patients with bicuspid aortic valve disease and thoracic aortic aneurysms. J. Thorac. Cardiovasc. Surg. *134*, 290–296. https://doi.org/10.1016/j.jtcvs.2007.02.041.
68. Teekakirikul, P., Zhu, W., Gabriel, G.C., Young, C.B., Williams, K., Martin, L.J., Hill, J.C., Richards, T., Billaud, M., Phillippi, J.A., et al. (2021). Common deletion variants causing protocadherin-α deficiency contribute to the complex genetics of BAV and left-sided congenital heart disease. Hum. Genet. Genomics Adv. *2*, 100037. https://doi.org/10.1016/j.xhgg.2021.100037.
69. Arndt, A.-K., Schafer, S., Drenckhahn, J.-D., Sabeh, M.K., Plovie, E.R., Caliebe, A., Klopocki, E., Musso, G., Werdich, A.A., Kalwa, H., et al. (2013). Fine mapping of the 1p36 deletion syndrome identifies mutation of PRDM16 as a cause of cardiomyopathy. Am. J. Hum. Genet. *93*, 67–77. https://doi.org/10.1016/j.ajhg.2013.05.015.
70. Kachuri, L., Graff, R.E., Smith-Byrne, K., Meyers, T.J., Rashkin, S.R., Ziv, E., Witte, J.S., and Johansson, M. (2020). Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. Nat. Commun. *11*, 6084. https://doi.org/10.1038/s41467-020-19600-4.
71. Kramer, I., Hooning, M.J., Mavaddat, N., Hauptmann, M., Keeman, R., Steyerberg, E.W., Giardiello, D., Antoniou, A.C., Pharoah, P.D.P., Canisius, S., et al. (2020). Breast cancer polygenic risk score and contralateral breast cancer risk. Am. J. Hum. Genet. *107*, 837–848. https://doi.org/10.1016/j.ajhg.2020.09.001.
72. Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J.P., Chen, T.-H., Wang, Q., Bolla, M.K., et al. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. Am. J. Hum. Genet. *104*, 21–34. https://doi.org/10.1016/j.ajhg.2018.11.002.
73. Liu, W., Zhuang, Z., Wang, W., Huang, T., and Liu, Z. (2021). An improved genome-wide polygenic score model for

predicting the risk of type 2 diabetes. Front. Genet. *12*, 632385. https://doi.org/10.3389/fgene.2021.632385.

74. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al. (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet *385*, 1305–1314. https://doi.org/10.1016/s0140-6736(14)61705-0.
75. Karunamuni, R.A., Huynh-Le, M.-P., Fan, C.C., Thompson, W., Lui, A., Martinez, M.E., Rose, B.S., Mahal, B., Eeles, R.A., Kote-Jarai, Z., et al.; UKGPCS Collaborators; and The PRACTICAL Consortium (2021). Performance of African-ancestry-specific polygenic hazard score varies according to local ancestry in 8q24. Prostate Cancer Prostatic Dis., 1–9. https://doi.org/10.1038/s41391-021-00403-7.
76. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591. https://doi.org/10.1038/s41588-019-0379-x.
77. Ewing, A., LCGC (2021). Reimagining Health Equity in Genetic Testing.
78. Lee, J., Kiiskinen, T., Mars, N., Jukarainen, S., Ingelsson, E., Neale, B., Ripatti, S., Natarajan, P., and Ganna, A. (2021). Clinical conditions and their impact on utility of genetic scores for prediction of acute coronary syndrome. Circ. Genom Precis Med. *14*, e003283. https://doi.org/10.1161/circgen.120.003283.
79. Franks, P.W., Melén, E., Friedman, M., Sundström, J., Kockum, I., Klareskog, L., Almqvist, C., Bergen, S.E., Czene, K., Hägg, S., et al. (2021). Technological readiness and implementation of genomic-driven precision medicine for complex diseases. J. Intern. Med. *290*, 602–620. https://doi.org/10.1111/joim.13330.
80. Lakeman, I.M.M., van den Broek, A.J., Vos, J.A.M., Barnes, D.R., Adlard, J., Andrulis, I.L., Arason, A., Arnold, N., Arun, B.K., Balmaña, J., Barrowdale, D., Benitez, J., Borg, A., Caldes, T., Caligo, M.A., Chung, W.K., Claes, K.B.M., GEMO Study Collaborators; EMBRACE Collaborators, Collee, J.M., Couch, F.J., Daly, M.B., Dennis, J., Dhawan, M., Domchek, S.M., Eeles, R., Engel, C., Evans, D.G., Feliubadalo, L., Foretova, L., Friedman, E., Frost, D., Ganz, P.A., Garber, J., Gayther, S.A., Gerdes, A.M., Godwin, A.K., Goldgar, D.E., Hahnen, E., Hake, C.R., Hamann, U., Hogervorst, F.B.L., Hooning, M.J., Hopper, J.L., Hulick, P.J., Imyanitov, E.N., OCGN Investigators; HEBON Investigators; and KconFab Investigators, Isaacs, C., Izatt, L., Jakubowska, A., James, P.A., Janavicius, R., Jensen, U.B., Jiao, Y., John, E.M., Joseph, V., Karlan, B.Y., Kets, C.M., Konstantopoulou, I., Kwong, A., Legrand, C., Leslie, G., Lesueur, F., Loud, J.T., Lubinski, J., Manoukian, S., McGuffog, L., Miller, A., Gomes, D.M., Montagna, M., Mouret-Fourme, E., Nathanson, K.L., Neuhausen, S.L., Nevanlinna, H., Yie, J.N.Y., Olah, E., Olopade, O.I., Park, S.K., Parsons, M.T., Peterlongo, P., Piedmonte, M., Radice, P., Rantala, J., Rennert, G., Risch, H.A., Schmutzler, R.K., Sharma, P., Simard, J., Singer, C.F., Stadler, Z., Stoppa-Lyonnet, D., Sutter, C., Tan, Y.Y., Teixeira, M.R., Teo, S.H., Teule, A., Thomassen, M., Thull, D.L., Tischkowitz, M., Toland, A.E., Tung, N., van Rensburg, E.J., Vega, A., Wappenschmidt, B., Devilee, P., van Asperen, C.J., Bernstein, J.L., Offit, K., Easton, D.F., Rookus, M.A., Chenevix-Trench, G., Antoniou, A.C., Robson, M., and Schmidt, M.K. (2021). The predictive ability of the 313 variant-based polygenic risk score for contralateral breast cancer risk prediction in women of European ancestry with a heterozygous BRCA1 or BRCA2 pathogenic variant. Genet. Med. *23*, 1726–1737. https://doi.org/10.1038/s41436-021-01198-7.
81. Barnes, D.R., Rookus, M.A., McGuffog, L., Leslie, G., Mooij, T.M., Dennis, J., Mavaddat, N., Adlard, J., Ahmed, M., Aittomäki, K., et al. (2020). Polygenic risk scores and breast and epithelial ovarian cancer risks for carriers of BRCA1 and BRCA2 pathogenic variants. Genet. Med. *22*, 1653–1666. https://doi.org/10.1038/s41436-020-0862-x.
82. Trevino, C.E., Rounds, J.C., Charen, K., Shubeck, L., Hipp, H.S., Spencer, J.B., Johnston, H.R., Cutler, D.J., Zwick, M.E., Epstein, M.P., et al. (2021). Identifying susceptibility genes for primary ovarian insufficiency on the high-risk genetic background of a fragile X premutation. Fertil. Steril. *116*, 843–854. https://doi.org/10.1016/j.fertnstert.2021.04.021.
83. Tadros, R., Tan, H.L., ESCAPE-NET Investigators, El Mathari, S., Kors, J.A., Postema, P.G., Lahrouchi, N., Beekman, L., Radivojkov-Blagojevic, M., Meitinger, T., Tanck, M.W., et al. (2019). Predicting cardiac electrical response to sodium-channel blockade and Brugada syndrome using polygenic risk scores. Eur. Heart J. *40*, 3097–3107. https://doi.org/10.1093/eurheartj/ehz435.