Clinical validation, implementation, and reporting of polygenic risk scores for common diseases

Jason Vassy (JVASSY@PARTNERS.ORG)  
Harvard University  https://orcid.org/0000-0001-6113-5564

Limin Hao  
Laboratory for Molecular Medicine

Peter Kraft  
Harvard T.H. Chan School of Public Health  https://orcid.org/0000-0002-4472-8103

Gabriel Berriz  
Laboratory for Molecular Medicine

Elizabeth Hynes  
Laboratory for Molecular Medicine  https://orcid.org/0000-0002-3817-2145

Christopher Koch  
Laboratory for Molecular Medicine

Prathik Kumar  
Laboratory for Molecular Medicine

Shruti Parpattedar  
Laboratory for Molecular Medicine

Marcie Steeves  
Laboratory for Molecular Medicine; Massachusetts General Hospital

Wanfeng Yu  
Laboratory for Molecular Medicine

Ashley Antwi  
VA Boston Healthcare System  https://orcid.org/0000-0002-2591-4410

Charles Brunette  
VA Boston Healthcare System  https://orcid.org/0000-0003-1620-3526

Morgan Danowski  
VA Boston Healthcare System

Manish Gala  
Massachusetts General Hospital

Robert Green  
Brigham and Women's Hospital; Mass General Brigham Personalized Medicine; Ariadne Labs; Medical and Population Genetics, The Broad Institute of MIT and Harvard; Harvard Medical School  https://orcid.org/0000-0001-8472-0424

Natalie Jones
Article

Keywords:

DOI: https://doi.org/10.21203/rs.3.rs-743779/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Implementation of polygenic risk scores (PRS) may improve disease prevention and management but requires the construction and validation of clinical assays, interpretation, and reporting pipelines. We developed a clinical genotype array-based assay for published PRS for 6 common diseases. First, we calculated PRS for 36,423 Mass General Brigham Biobank (MGBB) participants. Finding significant variation in the PRS distributions by race, we implemented adjustment for population structure with ancestry-informative principal components. We replicated published thresholds for odds ratio (OR)>2 in MGBB overall [ranging from 1.75 (1.57, 1.95) for Type 2 diabetes to 2.38 (2.07, 2.73) for breast cancer]. After confirming the high performance and robustness of the pipeline for use as a clinical assay, we analyzed the first 141 prospective samples from the Genomic Medicine at VA Study; frequency of PRS corresponding to published OR>2 ranged from 5/141 (3.6%) for colorectal cancer to 8/48 (16.7%) for breast cancer. Our development of a clinical PRS assay for multiple conditions illustrates the generalizability of this process and necessary technical and reporting decisions for meaningful clinical PRS implementation.

Introduction

For more than a decade, genome-wide association studies (GWAS) have identified thousands of genomic variants significantly associated with a range of common complex human diseases, including cardiometabolic diseases and cancers.\textsuperscript{1,2} Since the risks conferred by common variants are most often individually insignificantly small, investigators have aggregated risk alleles across the genome into genetic risk scores, providing a single measure of genetic association for a given trait due to known common variant effects.\textsuperscript{3–5} While the earliest genetic scores consisted only of variants meeting genome-wide significance, recent computational and methodological advances have leveraged the summary statistics of all available variants from increasingly larger GWAS to calculate \textit{polygenic} risk scores (PRS).\textsuperscript{6–9} For some diseases, a PRS value in the upper tail of the distribution may approximate risks equivalent to those conferred by established clinical risk factors as well as genetic variants associated with monogenic disease.\textsuperscript{7,10} Although PRS are typically derived from weights from cross-sectional GWAS of prevalent disease cases and controls, further work has demonstrated their potential to estimate risk of incident disease.\textsuperscript{11–14}

Suitable clinical implementation of PRS is now an area of active research across many disease areas.\textsuperscript{15–18} However, a key assumption underlying the potential clinical translation of PRS is that clinical laboratories can develop and implement valid clinical assays and interpretation pipelines to report PRS to clinicians and their patients in a meaningful way. The development of a clinical assay from a published PRS is not trivial, and a number of barriers to the process exist. First, uncertainty exists about whether commonly used, cost-effective genotyping arrays and clinical imputation pipelines can calculate a PRS for an individual with the analytic and clinical validity expected of a clinical assay, as opposed to one
adequate for research. Second, laboratories must implement methods to account for the reduced validity of most PRS among patients of non-European and admixed ancestry.\textsuperscript{19,20} This limitation applies both to the calculation of the PRS itself and its clinical interpretation, as published effect sizes are from populations of primarily European ancestry.\textsuperscript{19} Published methods can adjust for population structure, enabling all population groups to be treated similarly in downstream analyses.\textsuperscript{21,22} However, these methods are not immediately applicable to correct a PRS for a prospectively genotyped individual, whose sample is likely part of a small batch with insufficient data for robust adjustment for population structure. Correction thus requires additional decisions about how to adjust for population structure and which reference to use. Third, unanswered questions about the content and format of a clinical PRS report include balancing clarity and actionability with full transparency and limitations.

Despite these challenges, PRS assays are under active development by both clinical and research laboratories.\textsuperscript{23–28} Many have reported the aggregate performance of these PRS in a population, including biobanks or customers of direct-to-consumer companies, but none has described the development and validation of a clinical, population structure-adjusted assay for prospectively tested individuals. The Genomic Medicine at VA (GenoVA) Study (ClinicalTrials.gov Identifier: NCT04331535) is an ongoing randomized clinical trial examining the impact of PRS on disease diagnosis and prevention in a prospectively enrolled cohort of adult primary care patients. Here, we report the development and validation of a genotype array-based clinical assay and report for published PRS for six diseases and actionable monogenic findings, for implementation among prospectively genotyped individuals.

\textbf{Materials And Methods}

\textbf{Selection of diseases and PRS for implementation}

Because the GenoVA Study is examining the impact of PRS on disease diagnosis and prevention in adult primary care, we selected PRS for common diseases for which the typical primary care physician likely already has established practice patterns for screening, diagnosis, and prevention: coronary artery disease (CAD), type 2 diabetes mellitus (T2D), atrial fibrillation (AFib), colorectal cancer (CRCa), prostate cancer (PrCa), and female breast cancer (BrCa). We identified large genome-wide association studies (GWAS) for the six target diseases whose summary statistics (base files with alleles and weights) were freely available from the Polygenic Score (PGS) Catalog\textsuperscript{29} (AFib, CAD, T2D, BrCa) or the Cancer-PRSWeb (CRCa, PrCa)\textsuperscript{30} as of December 26, 2019. For the three cardiometabolic diseases (AFib, CAD, and T2D), we chose the PRS derived from the UK Biobank in Khera 2018\textsuperscript{7}: AFib (PGP ID = PGP000006, PGS ID = PGS000016), CAD (PGP ID = PGP000006, PGS ID = PGS000013), and T2D (PGP ID = PGP000006, PGS ID = PGS000014). For the three cancers, we chose PRS derived from the largest published GWAS at the time: BrCa (Michailidou 2017, Mavaddat 2019; PGS ID = PGS000007, PGP ID = PGP000002)\textsuperscript{31,32}, CRCa (Huyghe 2019; PPRSWEB_PHECODE153_CRC-Huyghe_PT_MGI_20191112, PRS tuning parameter:}
Replication of published PRS

Population and sample

As the GenoVA Study is enrolling participants from eastern Massachusetts, USA, we used data from the Mass General Brigham (formerly Partners Healthcare) Biobank (MGBB), described previously, to evaluate the performance of the selected PRSs within a similar population and workflow for our study and assay. MGBB participants were not included in the published derivation and validation studies for the PRS used. In brief, MGBB was launched in 2010 with the initial goal of collecting DNA, plasma, and serum samples from 75,000 patients from Brigham and Women's Hospital, Massachusetts General Hospital, and other MGBB-affiliated healthcare facilities, and obtaining patient consent for the linkage among biospecimen data, medical record data, and survey data. Race, ethnicity, and sex data derive from a combination of participant and healthcare provider report in the MGB electronic health record (EHR). For the present analysis, we collapsed reported race and ethnicity into 4 categories: Asian, Black, white, and other/unknown.

Disease phenotyping

We used validated computed phenotypes from MGBB to define case and control status for each of the 6 diseases (Supplemental Table 1). Validated MGB phenotypes are available for CAD (PPV=95%), AFib (PPV=94%), T2D (PPV=95%), and colorectal (PPV=100%), breast (PPV=95%), and prostate cancer (PPV=100%), as described previously. For each disease, ‘caseness’ was assigned as prevalent disease through December 16, 2019. For subgroup analyses, participant age was determined on December 16, 2019 or at death, if earlier. Only women and men were assigned case or control status for breast and prostate cancer, respectively.

Genotyping and imputation

We used genotype data from the 36,423 MGBB participants whose biospecimens had been genotyped within the MGBB as of December 16, 2019. Genotyping was performed using standard processing described previously on one of three Illumina Infinium genotyping arrays: 1) a pre-release version developed by the Multi-Ethnic Genotyping Array Consortium (Multi-Ethnic Genotyping Array (MEGA), n=4924), 2) an expanded version of this pre-commercial array (Expanded Multi-Ethnic Genotyping Array...
(MEGAEX), n=5345), and 3) the final commercial version (Multi-Ethnic Global (MEG), n=26157). The MEGA, MEGAEX, and MEG arrays consisted of 1.39, 1.74, and 1.78 million probes, respectively. For MEGA and MEGAEX data, only probes found in the commercial version of the array (MEG) were used in the present analysis. Quality control for the genotyping requires samples to have at least a 99% call rate and concordant sex between the EHR and what is computed from the array data. We utilized existing MGBB imputed data generated by batching sets of ~5000 participants and imputing against 1000 Genome Project phase 3 data using the Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html!), with ShapeIT (v2.r790) used for phasing and Minimac3 used for imputation with default settings. Sets of imputed variants were compared to the base files for each PRS to ensure sufficient representation of probes (Supplemental Table 2).

**Calculation of PRS and adjustment for population structure**

Unadjusted raw PRS (PRS\textsubscript{raw}) for each disease were calculated using PLINK (version 2.0a) by taking the product of count of risk alleles and the risk allele weight at each locus in the PRS and then summing across available risk loci. The loci included in each PRS, the risk alleles and the corresponding weights were downloaded from the PGS Catalog or Cancer PRSWeb. A population structure-adjusted PRS was calculated for each disease, using a previously described approach implementing principal components analysis (PCA) to compute adjusted residualized PRS for each disease. Principal components were calculated using all genotyped MGBB participants and a set of 16,385 of 16,443 previously reported ancestry-informative SNPs. For each disease, we then fit a linear model for unadjusted PRS (PRS\textsubscript{raw}) as a function of the first four PCs among controls for that disease (PRS\textsubscript{raw} \sim PC1 + PC2 + PC3 + PC4) in R (v4.0.3). We then applied this model to calculate a predicted PRS (PRS\textsubscript{pred}) for each disease among all cases and controls. Residualized, population structure-adjusted PRS (PRS\textsubscript{adj}) were then computed for each individual for each disease as the difference between raw and predicted PRS (PRS\textsubscript{raw} - PRS\textsubscript{pred}). For PRS\textsubscript{raw}, values were standardized (PRS\textsubscript{std-raw}) using the mean and standard deviation (SD) in MGBB of the PRS\textsubscript{raw} values (Supplemental Table 3). Similarly, PRS\textsubscript{std-adj} was computed using the mean and SD values in MGBB of the PRS\textsubscript{adj} values (Supplemental Table 3). The distributions of PRS\textsubscript{std-raw} and PRS\textsubscript{std-adj} by genotype array, sex, age deciles, and reported race were compared among all subjects using the density function in R (v4.0.3).

**PRS-disease association**

The association of PRS\textsubscript{std-adj} with odds of disease was replicated among MGBB participants using the six disease phenotypes described above. For each PRS and disease, odds of disease (\(n_{cases}/n_{controls}\)) were calculated for each of 50 PRS quantiles. For race-stratified analyses, PRS deciles were used if too few cases were available for analysis across 50 quantiles. To visualize PRS-disease associations, we plotted...
the log(odds) of disease against the mean PRS_{std-adj} in each quantile. Correlation was measured with \( R \) coefficients using RStudio (v1.1.383) with R (v4.0.3).

**PRS threshold for high risk**

We set an odds ratio (OR) >2 to indicate high polygenic risk for each disease, mirroring both a common threshold from Mendelian genetics\(^{45,46}\) and the effect sizes for disease risk factors already considered in current clinical care.\(^{47-51}\) We used the OR per standard deviation (SD) reported in the original publication to determine the Z-score threshold corresponding to OR>2 for each disease, where \( Z \) for OR>2 = \( \log(2)/\log(\text{OR change per SD}) \), 2 is the target OR, and OR change per SD is the coefficient from the literature for each disease (Supplemental Table 3).

**Clinical PRS assay for individual samples**

Based on the results of the above methods, we developed and validated a genotype array-based clinical assay for PRS, in addition to secondary findings from the American College of Medical Genetics and Genomics v2.0 list (ACMG SF v2.0, Figure 1).\(^{52}\)

**Validation samples**

Replicates of each of three reference samples from Genome in a Bottle (GIAB)\(^ {53}\) maintained by the National Institute of Standards and Technology were included in the validation assay: NA12878 x 9, NA24631 x 6, and NA24385 x 6. Analytical performance (sensitivity and positive predictive value for presence/absence of variant sites) was determined within the high-confidence regions (v3.3.2). In addition, we included 1) 22 samples with PCR-free genome data (described below) and 2) 9 samples with high-risk PRS for one of the six diseases as determined by the MGBB data, including one individual with high-risk PRS for two diseases. To test the sensitivity of the secondary finding analysis, we genotyped 20 samples with previously identified pathogenic or likely pathogenic variants within the ACMG SF v2.0 list.

**Genotyping and imputation**

Validation samples were genotyped according to manufacturer-standard workflows on either a pre-commercial release of the Illumina Global Diversity Array (GDA-PC) or the final commercial release of the Global Diversity Array (GDA). The GTC files generated by genotype array were converted to VCF format using a custom coding and the gtc2vcf script from Illumina. All samples required an overall call rate
Imputation was performed using updated software, with EAGLE v2.4.1 for phasing and Minimac4 for imputation using the 1000 Genomes Project phase 3 dataset. Importantly, monomorphic sites were not removed during the imputation process due to the small batch sizes used in the prospective assay.

**PRS calculation**

PRS\textsubscript{raw} was calculated for each sample as described above. To determine PRS\textsubscript{adj}, the results from the MGBB PC analysis were used to project each new individual sample onto the MGBB PCs (see Supplemental Methods). The scaled PCs were fitted into the linear model for each disease developed in the MGBB data to obtain PRS\textsubscript{pred}, PRS\textsubscript{adj}, and PRS\textsubscript{std-adj} as above, standardized using the mean and standard deviation for each phenotype from the MGBB data (Supplemental Table 3).

**Genome sequencing**

We selected 22 diverse samples that had previously undergone clinical whole genome sequencing to determine the robustness of PRS across different platforms. Genome sequencing was performed at the Clinical Research Sequencing Platform of the Broad Institute using PCR-free library construction and sequencing on an Illumina NovaSeq with 2 x 150 bp paired-end reads with \( \geq 95\% \) of bases covered at \( \geq 20x \). Reads were aligned to GRCh37 using the Burrows-Wheeler Aligner (BWA version 0.7.15) and variant calls were made using HaplotypeCaller from the Genomic Analysis Tool Kit (GATK version 4.0.3.0)\textsuperscript{56,57,58}. PRS\textsubscript{raw}, PRS\textsubscript{std-raw}, PRS\textsubscript{adj}, and PRS\textsubscript{std-adj} were calculated as above for the other prospective samples. As stated above, these 22 samples were also run on the GDA-PC array to compare PRS between genome sequencing and array. The difference between the sequence-based and array-based PRS were visualized, and dichotomous risk classifications were formally compared with Matthews Correlation Coefficient (MCC)\textsuperscript{59}.

**Identification of actionable variants associated with monogenic disease**

Variants from the original genotyping vcf were annotated and filtered to the 59 genes suggested for screening of secondary findings as recommended by the American College of Medical Genetics and Genomics (ACMG SF v2.0)\textsuperscript{52} to find: 1) variants previously identified as disease causing by the MGB Laboratory for Molecular Medicine, 2) variants classified as P/LP within ClinVar with a minor allele frequency (MAF) <0.1%, 3) variants classified as a disease-causing mutation (DM) in HGMD with a MAF <0.03%, and 4) loss-of-function variants (nonsense, frameshift, canonical splice-site, and initiating methionine variants) with a MAF <0.1% in genes where that is a disease-mechanism. Clinical variant
classification was carried out in accordance with the criteria set by the guidelines by the ACMG and the Association of Molecular Pathology,\textsuperscript{60} with disease specific modifications as recommended by the Clinical Genome Resource Expert Panels.\textsuperscript{61}

**Prospectively enrolled trial participants**

The assay described above is now in use in the ongoing GenoVA Study randomized trial of clinical PRS (ClinicalTrials.gov Identifier: NCT04331535), in which eligible participants are patients of the VA Boston Healthcare System, aged 50-70, without known diagnoses of the 6 target diseases. Enrollees provide a clinical blood or saliva sample for analysis at the LMM and then receive a PRS report along with relevant disease-specific information.

**Results**

**Replication of published PRS**

**Sample characteristics**

Among the 36,423 MGBB participants whose genotype data were used to calculate the 6 PRS, mean (SD) age was 58.8 (17.1) years (range 9-106), 19,719 (54.1\%) were female, and 5706 (15.7\%) were of reported race other than white [30,716 (84.3\%) white, 1,807 (5.0\%) Black, 786 (2.2\%) Asian, and 3,113 (8.5\%) of other/unknown race]. Case counts ranged from 392 CRC\textsubscript{a} cases to 3,554 cases of CAD. Figure 2A shows the counts of participants with one or multiple of the target diseases as determined by the computed phenotypes. The most common disease co-occurrences were the combinations of CAD and T2D (\(n=641\)) and CAD and AFib (\(n=495\)).

**PRS distributions before and after adjustment for population structure**

Supplemental Table 2 shows the numbers of SNPs in the base file for each of the 6 published PRS, ranging from 81 SNPs in the Huyghe CRC\textsubscript{a}\textsuperscript{33} to 6,917,436 in the Khera T2D PRS\textsuperscript{7}, and the subsets of these available as directly genotyped or imputed data from each of the 3 arrays used among MGBB participants, demonstrating minimal loss of information compared to the original published PRS. As shown in Figure 3, using the weights from the publications directly (PRS\textsubscript{std-raw}), we observed marked variation in the distribution of each PRS by race in MGBB, most notably in AFib, CAD, and T2D. For example, only 516/30,716 (1.7\%) of white MGBB participants but almost all (1,606/1,807, 88.9\%) Black MGBB participants had PRS\textsubscript{std-raw} above the threshold associated with OR>2 for T2D in Khera 2018\textsuperscript{7} (Supplemental Table 4). Adjustment of the raw PRS (PRS\textsubscript{std-adj}) reduced this variation (Figure 3), such that, for example, 2,651/30,716 (8.6\%) of white MGBB participants and 75/1,807 (4.2\%) of Black MGBB
participants had aT2D PRS_{std-adj} above the published OR>2 threshold. The distributions of PRS_{std-adj} were well aligned when examined by genotyping batch, decile of age, and sex (Supplemental Figures 1-3).

**Replication of PRS-disease association**

As shown in Figure 4, quantile of PRS_{std-adj} was highly correlated with log(odds) of disease across the 6 phenotypes in MGBB, with correlation coefficients ranging from 0.68 for CRCa to 0.95 for T2D. Supplemental Figures 4-7 show the correlation of PRS_{std-adj} quantile and log(odds) of disease among reported racial groups separately. Our analyses also replicated the published thresholds corresponding to OR>2. As shown in Table 1, at the published PRS_{std-adj} thresholds, we observed OR ranging from 1.75 (95% CI 1.57, 1.95) for T2D to 2.38 (95% CI 2.07, 2.73) for BrCa among MGBB participants overall. With the exception of T2D, the 95% CI of the replicated OR for all diseases either included or, in the case of BrCa and AFib, exceeded a point estimate of 2. Results were consistent in analyses restricted to white MGBB participants but were variable in other groups, in large part because of small numbers of disease cases. In 22 of 24 analyses stratified by reported race, subjects with PRS_{std-adj} above the published OR>2 thresholds had higher odds of disease than those below these thresholds. In the MGBB overall, the prevalence of a high-risk PRS_{std-adj} ranged from 5.4% for CRCa to 13.2% for PrCa (among men). Figure 2B illustrates the number of participants with PRS_{std-adj} above the published OR>2 threshold for one or more of the target diseases. Of note, similar to the disease co-occurrences observed among MGBB participants, the most common co-occurrences of high-risk PRS_{std-adj} were the combinations of CAD and T2D ($n=333$) and CAD and AFib ($n=211$).

**Prospective PRS assay**

**Sensitivity and specificity of GDA and imputation**

To determine the performances of the GDA arrays used in the prospective assay and of the imputation pipeline, we used three reference GIAB samples (NA12878, NA24385 and NA24631) (Supplemental Table 5). Sensitivity and positive predictive value (PPV) for SNVs were > 99.7% on average, with lower performance in indels (sensitivity = 96.3% and PPV = 97.8%). Of note, while sensitivity with the ACMG SF v2.0 region was high (96.2%), PPV was low (63.6%), due to these regions having an excess of poorly performing rare variants.\textsuperscript{62,63}

As expected, sensitivity and PPV decreased for imputed data, especially for indels (SNV sensitivity = 98.0%; SNV PPV = 97.5%; indel sensitivity = 92.8%; indel PPV =90.7%) (Supplemental Table 5). NA12878 was not evaluated for imputation accuracy, as it is present in the imputation reference dataset from the
1000 Genomes Project, and has artificially high imputation accuracy. To further evaluate imputation accuracy, we compared genome sequencing data to array data for 22 diverse samples. Analytical performance was lower in this dataset as compared to the GIAB high-confidence data (~3% reduction in performance for sensitivity and PPV, Supplemental Table 6).

**Performance of prospective PRS assay**

For the GIAB samples, PRS\textsubscript{std-adj} values were robust across different array versions and consistent with results from WGS data; all 3 GIAB samples were below the high risk threshold for all diseases in all methods (Supplemental Table 7). In evaluating the 22 samples with WGS and prospective array data, PRS\textsubscript{std-adj} scores were similarly concordant, particularly for AF, CAD and T2D (Supplemental Fig. 8). Additionally, 108/110 high risk status classifications were concordant among this dataset (98.2% agreement; MCC 0.84, \( p < 0.001 \)), with the two discordant values (1 in CAD and 1 in CRCa) being very close to the high risk threshold (Supplemental Table 8). Finally, we compared 9 individuals with high-risk PRS for 10 diseases identified in the MGBB genotyping data to their PRS risk status using the prospective assay (1 individual at high risk for AFib, 1 individual at high risk for BrCa, 3 individuals at high risk for CAD, 3 individuals at high risk for CRCa, 1 individual at high risk for PrCa, and 1 individual at high risk for T2D). All PRS categories were consistent across the two genotyping approaches (Supplemental Table 9).

**Clinical PRS report**

Based on the above validation, we produced a PRS report consistent in format and content with other clinical genetic test reports (Supplemental File 2).\textsuperscript{64–67} That is, it includes a description of the test performed and a prominently displayed summary of important findings and their interpretations, including any monogenic disease variants identified and any PRS indicating increased polygenic disease risk. A graphic highlights in red the disease(s) for which the patient has increased polygenic disease risk. Subsequent sections of the report give more detail about the results, including, for each disease, a general population prevalence and a brief summary of the GWAS from which the PRS was derived. Sections on methodology and literature references are at the end of the report. The report highlights the European bias of these GWAS and PRS in the initial summary, stating “Polygenic risk calculated using data from predominantly European ancestry individuals. Results are known to be less accurate for individuals of non-European ancestry.” This information is reiterated in the detailed description of each disease and in a limitations section at the end of the report.

**Results from first 141 prospective samples**

As of May 15, 2021, the DNA specimens from 141 GenoVA trial participants (73 blood, 68 saliva) have been assayed with the prospective PRS and secondary findings pipelines. As shown in Table 2, $n=92$ (65%) participants are of non-white reported race/ethnicity, and 49 (35%) currently identify as women. In this preliminary sample of trial enrollees, the proportions of participants whose PRS are above the study threshold for high risk are similar to the proportions expected from the MGBB, ranging from 3.6% for CRCa to 16.7% for BrCa (8 high-risk results among 48 participants of female sex at birth). Two actionable ACMG SF v2.0 variants have been identified and confirmed among the first 141 enrollees ($BRCA1$: NM_007294 c.2748delT (p.Asn916LysfsX84), likely pathogenic; $BRCA2$: NM_000059 c.3545_3546delTT (p.Phe1182X), pathogenic). The reporting of these results to trial participants and their primary care providers is underway.

**Discussion**

Bridging one significant gap between PRS development and clinical PRS implementation, we developed a clinical genotyping array-based assay to calculate published PRS for six common diseases. The scores were robust across multiple genotyping arrays and imputation pipelines. The distributions of these PRS varied by race in a large biobank, impeding clinical validation, but adjustment for population structure enabled the replication of published PRS-disease associations. These results informed the development of a population structure-adjusted pipeline for PRS calculation and reporting among prospectively genotyped patients, now in use in a clinical trial of PRS testing.

The development and implementation of our PRS assay and report illustrates key choices that laboratories must make in such a process. First, across the six target diseases, we had to choose the specific PRS to implement among multiple publicly available options. Considerations here include the performance of the PRS in both the published discovery and replication cohorts in addition to the target population of interest. Guidelines are emerging on what defines high quality in publishing PRS, and this improved transparency should help laboratories select appropriate PRS. Second, we chose to use a genotype array-based approach, instead of genome sequencing. Like genotyping, low coverage genome sequencing technology is also relatively low-cost; we chose the Illumina GDA, in part, because its widespread use in the All of Us Research Program, eMERGE Consortium, and other projects enhances the generalizability of our methods to other institutions looking to implement clinical PRS testing. Third, we chose to impute data from both our development cohort and prospectively enrolled subjects against 1000 Genome Project phase 3 data. Other laboratories may choose to impute against the larger TOPMed population, although issues of genome build discrepancy and regulatory prohibition against sending patient data to external research servers are limitations. Further, the degree to which an imputation pipeline includes stochastic processes may influence robustness, one reason our prospective assay uses EAGLE for imputation. Fourth, once a platform is selected, a clinical laboratory must determine the benchmarks that should define an analytically valid PRS assay. We chose to 1) verify the PRS
performance within our laboratory to determine the appropriate parameters for our assay; 2) calculate the analytical performance of the genotyping array and imputation pipeline using both well-characterized reference samples and individual level genome data; and 3) calculate the robustness and performance of the PRS using genome data and multiple array platforms from both reference and individual samples. This multi-step approach helped ensure the accuracy of the data going into the PRS as well as the final performance of the PRS itself.

Finally, we made numerous choices in how to report PRS results and their interpretation to patients and providers, a full discussion of which is beyond the scope of this report. A key choice we made was to report the PRS interpretation as a categorical result (i.e. high risk vs. average risk) instead of a continuous result (e.g. percentile rank, relative risk, or absolute risk) or as a categorical result (e.g. low, intermediate, or high risk). We have previously described the trade-offs of these approaches.71 For the GenoVA Study, we chose OR>2 to define high categorical polygenic risk, consistent with effect sizes of traditional risk factors considered for our target diseases.47–51 Another laboratory could use the methods we describe to produce measures of continuous risk or of categorical risk at different thresholds thought to be clinically meaningful, which will likely vary between diseases. Modeling the prevailing scenario in clinical medicine in which a treating clinician orders a laboratory test for a patient and then receives the results, the content and format of the PRS report mirror those of a more traditional molecular diagnostic report,64–67 written for clinicians and not explicitly for patients.

Although other laboratories are developing PRS assays in both clinical and research settings,23–28 few have described their clinical interpretation pipeline in general or their approach to ancestry in particular. Myriad Genetics uses a next-generation sequencing panel backbone to measure a PRS of >80 SNPs identified from European GWAS72 and validated in women of European descent.23 This test had been available only to European ancestry patients, but a multi-ancestry version is in validation and will be commercially available in 2022.73–75 Ambry Genetics developed similar clinical PRS assays for breast and prostate cancer but announced that it would cease offering these tests as of May 25, 2021, citing lack of validation across ancestries and National Comprehensive Cancer Network advisement against the use of PRS in routine clinical care.24,76 The direct-to-consumer genetic testing company 23andMe offers customers a proprietary T2D PRS comprised of >1000 loci, using data derived from its own customers.25 23andMe customers can also share their genotype data with the MyGeneRank research study, which, as of December 2019, calculates a 163-SNP PRS for CAD and reports an ancestry-specific percentile rank to each participant.26 Color Genomics has developed a low-coverage whole-genome sequencing assay for PRS for CAD, AFib, and BrCa and shown that coverage depth as low as x0.5 achieved high correlation with PRS from genotype array-based PRS from UK Biobank and with PRS from sequence data from 120 ancestrally diverse samples from the 1000 Genomes Project.27 For a period of time, the laboratory
offered this test through a research protocol that has since closed.\textsuperscript{17} Although the eMERGE IV consortium and other studies are actively developing trans-ancestry PRS for a number of common diseases,\textsuperscript{28,77} to our knowledge ours is the first report of a single clinical assay for population structure-adjusted PRS for multiple diseases.

Much has been written about the reduced validity of most PRS among populations of non-European ancestry, due to their use of non-causal loci and effect sizes from GWAS in predominantly European discovery cohorts\textsuperscript{19,20,78,79}. As the genomics community awaits larger data sets from more diverse populations and develops improved methods for deriving trans-ancestry PRS,\textsuperscript{10,15,80} a laboratory aiming to develop a clinical PRS assay for a given disease has a few options: 1) postpone implementation; 2) implement separate ancestry-specific published PRS in those ancestries where it is available; or 3) implement a single PRS and report transparently any applicable limitations in the underlying evidence and its interpretation for specific individuals. We chose the last of these approaches and implemented a single method of adjustment for population structure. After doing so, we observed that the chosen OR>2 threshold generally identified subjects at higher risk of disease across race/ethnicity, although the magnitude of the increase varied across race/ethnicity. We contend that we have developed a clinically valid PRS assay for use in patients of diverse ancestry, including admixed individuals, whose results nonetheless have limitations which must be contextualized for each individual. Moving forward, our approach can be adapted to incorporate the rapid developments in this area to implement improved PRS with greater validity across populations. Similarly, our approach can include additional variants identified by the ACMG or organizations as important secondary findings.\textsuperscript{81}

The strengths of our approach include the inclusion of validated PRS for multiple common diseases and the use of a cost-effective, widely used genotyping platform that will be well supported for future improvements. We and others can use these methods to update the clinical interpretative pipeline as improved imputation methods are available and as improved PRS for these six and additional diseases become available. Limitations include the predominance of self-reported white race in the MGBB population used to replicate the published PRS and the small number of disease cases for certain diseases among certain racial groups.

In conclusion, data from increasingly larger and more diverse populations, coupled with computational advances, are propelling PRS into consideration for clinical implementation. We have shown that implementing these advancements in a laboratory assay and clinical report is a feasible but non-trivial next step in realizing the potential of PRS for improved patient health.
Declarations

Ethics declaration

Analysis of the genomic and MGB Biobank samples and data has been reviewed and approved by the Mass General Brigham IRB. All individuals with genome sequencing data gave consent for clinical genomic screening and all individual data were de-identified. The GenoVA Study is approved by the VA Boston Healthcare System and Harvard Medical School IRB.

Data availability

Individual-level MGBB data are available from https://personalizedmedicine.partners.org/Biobank/Default.aspx; however, there are restrictions on this data which was accessed under IRB protocol for this current study so are not publicly available. The majority of the MGBB genotyped samples are deposited in dbGAP as part of the eMERGE consortium https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001584.v2.p2. Data from the GenoVA Study trial will be made publicly available after study completion. Code used to adjust the PRS for population structure are available for download here: [URL will be provided prior to publication].

Acknowledgements

This work was supported by the National Human Genome Research Institute (R35HG010706). Dr. Lubitz is supported by NIH grant 1R01HL139731 and American Heart Association 18SFRN34250007.

Author contributions

LH, PK, CAB, MKG, RCG, ACFL, SAL, PN, JLV, and MSL conceived the study and contributed to its design. LH, GB, EDH, CK, PKVK, SSP, MS, WY, and CAB analyzed the data from the MGBB and GenoVA samples. AAA, CAB, MD, NEJ, and JLV collected GenoVA data. LH, JLV, and MSL drafted the manuscript, and all authors reviewed the scientific content of the manuscript prior to submission.

Competing interests statement
ACFL owns stock in Fabric Genomics. SAL receives sponsored research support from Bristol Myer Squibb / Pfizer, Bayer AG, Boehringer Ingelheim, Fitbit, and IBM, and has consulted for Bristol Myers Squibb / Pfizer, Bayer AG, and Blackstone Life Sciences. PN reports investigator-initiated grants from Amgen, Apple, Boston Scientific, AstraZeneca, and Novartis, personal fees from Apple, Genentech, AstraZeneca, Blackstone Life Science, Foresite Labs, and Novartis, and spousal employment at Vertex, all unrelated to the present work. RCG has received compensation for advising the following companies: AlA, Genomic Life, Grail, Humanity, Kneed Media, OptumLabs, Plumcare, Verily, and VibrentHealth and is co-founder of Genome Medical, Inc. CK now works at Novartis Institutes for BioMedical Research.

References

1. Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic Medicine-Progress, Pitfalls, and Promise. *Cell* **177**, 45–57 (2019).

2. GWAS Catalog. https://www.ebi.ac.uk/gwas/. Accessed July 21, 2021.

3. Meigs, J. B. *et al.* Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N. Engl. J. Med.* **359**, 2208–2219 (2008).

4. Ripatti, S. *et al.* A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *The Lancet* **376**, 1393–1400 (2010).

5. Zheng, S. L. *et al.* Cumulative Association of Five Genetic Variants with Prostate Cancer. *N. Engl. J. Med.* **358**, 910–919 (2008).

6. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).

7. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219–1224 (2018).

8. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).

9. Inouye, M. *et al.* Genomic risk prediction of coronary artery disease in nearly 500,000 adults: implications for early screening and primary prevention. *BioRxiv* (2018) doi:10.1101/250712.

10. Conti, D. V. *et al.* Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nat. Genet.* **53**, 65–75 (2021).

11. Klarin, D. *et al.* Genome-wide association analysis of venous thromboembolism identifies new risk loci and genetic overlap with arterial vascular disease. *Nat. Genet.* **51**, 1574–1579 (2019).
12. Mosley, J. D. et al. Predictive Accuracy of a Polygenic Risk Score Compared With a Clinical Risk Score for Incident Coronary Heart Disease. *Jama* **323**, 627–635 (2020).

13. Vassy, J. L. et al. Polygenic type 2 diabetes prediction at the limit of common variant detection. *Diabetes* (2014) doi:10.2337/db13-1663.

14. Seibert, T. M. et al. Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. *BMJ* **360**, j5757 (2018).

15. National Human Genome Research Institute (NHGRI). Electronic Medical Records and Genomics (eMERGE) Network. (2020).

16. Shieh, Y. et al. Breast Cancer Screening in the Precision Medicine Era: Risk-Based Screening in a Population-Based Trial. *J. Natl. Cancer Inst.* **109**, (2017).

17. Brockman, D. G. et al. Design and user experience testing of a polygenic score report: a qualitative study of prospective users. *medRxiv* 2021.04.14.21255397 (2021) doi:10.1101/2021.04.14.21255397.

18. https://clinicaltrials.gov/ct2/show/NCT04331535 (2020). Accessed July 21, 2021.

19. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584–591 (2019).

20. Lewis, C. M. & Vassos, E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* **12**, 44 (2020).

21. Khera, A. V. et al. Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. *Circulation* **139**, 1593–1602 (2019).

22. Dikilitas, O. et al. Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups. *Am. J. Hum. Genet.* **106**, 707–716 (2020).

23. Hughes, E. et al. Development and Validation of a Clinical Polygenic Risk Score to Predict Breast Cancer Risk. *JCO Precis. Oncol.* **585–592** (2020) doi:10.1200/PO.19.00360.

24. Cancer Genetic Testing | AmbryScore | Health Risk Tests | Ambry Genetics. https://www.ambrygen.com/providers/ambryscore. Accessed July 21, 2021.

25. 23andMe. Our Health + Ancestry DNA Service - 23andMe. https://www.23andme.com/dna-health-ancestry/. Accessed July 21, 2021.

26. Chen, S.-F. et al. Genotype imputation and variability in polygenic risk score estimation. *Genome Med.* **12**, 100 (2020).
27. Homburger, J. R. et al. Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. *Genome Med.* 11, 74 (2019).

28. NIH funds centers to improve the role of genomics in assessing and managing disease risk. https://www.genome.gov/news/news-release/NIH-funds-centers-to-improve-role-of-genomics-in-assessing-and-managing-disease-risk. Accessed July 21, 2021.

29. Lambert, S. A. et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* 53, 420–425 (2021).

30. Fritsche, L. G. et al. Cancer PRSweb: An Online Repository with Polygenic Risk Scores for Major Cancer Traits and Their Evaluation in Two Independent Biobanks. *Am. J. Hum. Genet.* 107, 815–836 (2020).

31. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* 551, 92–94 (2017).

32. Mavaddat, N. et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am. J. Hum. Genet.* 104, 21–34 (2019).

33. Huyghe, J. R. et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet* 51, 76–87 (2019).

34. Schumacher, F. R. et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet* 50, 928–936 (2018).

35. Boutin, N. et al. The Information Technology Infrastructure for the Translational Genomics Core and the Partners Biobank at Partners Personalized Medicine. *J. Pers. Med.* 6, (2016).

36. Karlson, E. W., Boutin, N. T., Hoffnagle, A. G. & Allen, N. L. Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations. *J. Pers. Med.* 6, (2016).

37. Yu, S. et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J. Am. Med. Inform. Assoc. JAMIA* 22, 993–1000 (2015).

38. Zou, H. The Adaptive Lasso and Its Oracle Properties. *J. Am. Stat. Assoc.* 101, 1418–1429 (2006).

39. Yu, S. et al. Enabling phenotypic big data with PheNorm. *J. Am. Med. Inform. Assoc. JAMIA* 25, 54–60 (2018).

40. Gainer, V. S. et al. The Biobank Portal for Partners Personalized Medicine: A Query Tool for Working with Consented Biobank Samples, Genotypes, and Phenotypes Using i2b2. *J. Pers. Med.* 6, (2016).
41. Blau, A., Brown, A., Mahanta, L. & Amr, S. S. The Translational Genomics Core at Partners Personalized Medicine: Facilitating the Transition of Research towards Personalized Medicine. *J. Pers. Med.* **6**, (2016).

42. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

43. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).

44. Libiger, O. & Schork, N. J. A Method for Inferring an Individual’s Genetic Ancestry and Degree of Admixture Associated with Six Major Continental Populations. *Front. Genet.* **3**, (2013).

45. Senol-Cosar, O. *et al.* Considerations for clinical curation, classification, and reporting of low-penetrance and low effect size variants associated with disease risk. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **21**, 2765–2773 (2019).

46. Roberts, N. J. *et al.* The predictive capacity of personal genome sequencing. *Sci. Transl. Med.* **4**, 133ra58 (2012).

47. Goff, D. C. *et al.* 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* (2013) doi:10.1161/01.cir.0000437741.48606.98.

48. American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2021. *Diabetes Care* **44**, S15–S33 (2021).

49. Grossman, D. C. *et al.* Screening for Prostate Cancer: US Preventive Services Task Force Recommendation Statement. *Jama* **319**, 1901–1913 (2018).

50. Siu, A. L. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann. Intern. Med.* **164**, 279–296 (2016).

51. US Preventive Services Task Force *et al.* Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **325**, 1965 (2021).

52. Kalia, S. S. *et al.* Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).

53. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).
54. Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

55. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).

57. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. http://biorxiv.org/lookup/doi/10.1101/201178 (2017) doi:10.1101/201178.

58. Van der Auwera, G. A. & O’Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition).* (O’Reilly Media).

59. MAQC Consortium. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol.* **28**, 827-838 (2010).

60. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* (2015) doi:10.1038/gim.2015.30.

61. Rivera-Muñoz, E. A. et al. ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. *Hum. Mutat.* **39**, 1614–1622 (2018).

62. Bowling, K. M. et al. Identifying rare, medically relevant variation via population-based genomic screening in Alabama: opportunities and pitfalls. *Genet. Med.* **23**, 280–288 (2021).

63. Weedon, MN, W. et al. Use of SNP chips to detect rare pathogenic variants: retrospective, population based diagnostic evaluation. *BMJ* **372**, n214 (2021).

64. Scheuner, M. T., Edelen, M. O., Hilborne, L. H. & Lubin, I. M. Effective communication of molecular genetic test results to primary care providers. *Genet. Med.* **15**, 444–9 (2013).

65. McLaughlin, H. M. et al. A systematic approach to the reporting of medically relevant findings from whole genome sequencing. *BMC Med. Genet.* **15**, 134 (2014).

66. Davis, K. W., Hamby Erby, L., Fiallos, K., Martin, M. & Wassman, E. R. A comparison of genomic laboratory reports and observations that may enhance their clinical utility for providers and patients. *Mol. Genet. Genomic Med.* **7**, e00551 (2019).

67. Farmer, G. D., Gray, H., Chandratillake, G., Raymond, F. L. & Freeman, A. L. J. Recommendations for designing genetic test reports to be understood by patients and non-specialists. *Eur. J. Hum. Genet.*
68. Wand, H. et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature* **591**, 211–219 (2021).

69. Denny, J. C. et al. The ‘All of Us’ Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).

70. Kowalski, M. H. et al. Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet* **15**, e1008500 (2019).

71. Lewis, A., Green, R. & Vassy, J. Polygenic risk scores in the clinic: Translating risk into action. *HGG Adv.* (in press).

72. Michailidou, K. et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47**, 373–380 (2015).

73. Hughes, E. R. et al. Abstract P5-10-02: Development and validation of a polygenic score to predict breast cancer risk in unaffected Hispanic women negative for mutations on a multi-gene hereditary cancer panel. *Cancer Res.* **79**, P5-P5-10–02 (2019).

74. Kurian, A. W. et al. Performance of the IBIS/Tyrer-Cuzick (TC) Model by race/ethnicity in the Women’s Health Initiative. *J. Clin. Oncol.* **38**, 1503–1503 (2020).

75. Myriad Genetics Recalibrates Breast Cancer PRS for All Ancestries in Anticipation of Broader Launch. https://www.genomeweb.com/molecular-diagnostics/myriad-genetics-recalibrates-breast-cancer-prs-all-ancestries-anticipation (2021). Accessed July 21, 2021.

76. AmbryScore discontinuation notice. https://info.ambrygenetics.com/take-a-brief-survey-for-entry-into-amazon-gift-card-drawing. Accessed July 21, 2021.

77. National Human Genome Research Institute. NIH awards $38 million to improve utility of polygenic risk scores in diverse populations. https://www.genome.gov/news/news-release/nih-awards-38-million-dollars-to-improve-utility-of-polygenic-risk-scores-in-diverse-populations. Accessed July 21, 2021.

78. Manolio, T. A. Using the Data We Have: Improving Diversity in Genomic Research. *Am. J. Hum. Genet.* **105**, 233–236 (2019).

79. Lewis, A. C. F. & Green, R. C. Polygenic risk scores in the clinic: new perspectives needed on familiar ethical issues. *Genome Med.* **13**, 14 (2021).

80. Polygenic Risk Score (PRS) Diversity Consortium. *Genome.gov* https://www.genome.gov/Funded-Programs-Projects/Polygenic-Risk-Score-Diversity-Consortium. Accessed July 21, 2021.
Tables

Table 1. Prevalence and disease associations of high-risk PRS for 6 diseases in MGB Biobank overall and by reported race
| Disease | High risk (%) | OR Overall | OR White | OR Black | OR Asian | OR Other/Unknown |
|---------|--------------|------------|----------|----------|----------|-----------------|
| BrCa    | 8.6          | 2.38       | 2.39     | 2.24     | 0.51     | 2.35            |
|         |              | [2.07, 2.73] | [2.07, 2.76] | [0.97, 5.15] | [0.07, 3.9] | [1.08, 5.1] |
|         |              | (286/1400, 1427/16606) | (270/1156, 1318/13495) | (7/73, 43/1004) | (1/33, 24/405) | (8/138, 42/1702) |
| CRCa    | 5.4          | 2.37       | 2.29     | 4.11     | 0        | 3.30            |
|         |              | [1.74, 3.24] | [1.65, 3.19] | [1.17, 14.48] | [0, NaN] | [0.73, 14.88] |
|         |              | (46/1913, 346/34117) | (41/1646, 312/28717) | (3/83, 15/1706) | (0/35, 7/744) | (2/149, 12/2950) |
| PrCa    | 13.1         | 2.22       | 2.31     | 1.39     | 2.58     | 1.41            |
|         |              | [1.98, 2.48] | [2.05, 2.59] | [0.74, 2.59] | [0.5, 13.28] | [0.78, 2.58] |
|         |              | (498/1698, 1693/12813) | (468/1448, 1544/11017) | (14/71, 74/521) | (2/36, 6/279) | (14/143, 69/996) |
| AFib    | 8.3          | 2.37       | 2.40     | 1.47     | 2.00     | 2.28            |
|         |              | [2.12, 2.64] | [2.14, 2.69] | [0.72, 3.01] | [0.57, 7.03] | [1.32, 3.94] |
|         |              | (450/2589, 2282/31101) | (422/2179, 2101/26014) | (9/137, 71/1590) | (3/62, 17/704) | (16/211, 93/2793) |
| CAD     | 9.8          | 1.86       | 1.91     | 1.41     | 3.96     | 1.47            |
|         |              | [1.69, 2.05] | [1.73, 2.12] | [0.86, 2.29] | [1.79, 8.76] | [0.97, 2.22] |
|         |              | (562/3018, 2991/29851) | (503/2459, 2680/25074) | (21/177, 125/1484) | (9/51, 31/695) | (29/331, 155/2598) |
| T2D     | 8.4          | 1.75       | 1.93     | 1.21     | 1.07     | 1.58            |
|         |              | [1.57, 1.95] | [1.71, 2.17] | [0.7, 2.09] | [0.37, 3.08] | [1.14, 2.19] |
|         |              | (439/2612, 2924/30447) | (367/2284, 2159/25906) | (18/57, 358/1374) | (4/49, 52/681) | (50/222, 355/2486) |

High-risk PRS, defined here as a standardized, adjusted PRS (PRS\textsubscript{std-adj}) associated with OR>2 for disease in the original publication. OR shown are the observed OR [95% CI] among up to 36,423 MGBB participants in the overall cohort and by race reported in MGBB. Data below each OR take the format (n(n-1))
cases with high-risk PRS / n controls with high-risk PRS, n cases without high-risk PRS / n controls without high-risk PRS). Second column shows the proportion of MGBB participants exceeding the literature-derived OR>2 threshold for each disease. Abbreviations: AFib, atrial fibrillation; BrCa, breast cancer; CAD, coronary artery disease; CRCa, colorectal cancer; MGBB, Mass General Brigham Biobank; NaN, not a number; OR, odds ratio; PrCa, prostate cancer; PRS, polygenic risk score; T2D, type 2 diabetes.

Table 2. Summary of PRS results from first four batches of clinical samples in GenoVA Study

|                | BrCa | CRCa | PrCa | AFib | CAD | T2D |
|----------------|------|------|------|------|-----|-----|
| Total analyzed, n | 48   | 141  | 93*  | 141  | 141 | 141 |
| Average-risk results, n (%) | 40 (83.3) | 136 (96.5) | 82 (88.2) | 127 (90.1) | 131 (92.9) | 130 (92.2) |
| High-risk results, n (%)      | 8 (16.7) | 5 (3.5) | 11 (11.8) | 14 (9.9) | 10 (7.1) | 11 (7.8) |

Results from the first 141 GenoVA participants. High-risk PRS, defined here as a standardized, adjusted PRS (PRS<sub>std-adj</sub>) associated with OR>2 for disease in the original publication. All other results are considered average-risk. *One participant with male sex identifies as female. Abbreviations: AFib, atrial fibrillation; BrCa, breast cancer; CAD, coronary artery disease; CRCa, colorectal cancer; GenoVA, Genomic Medicine at VA; OR, odds ratio; PrCa, prostate cancer; PRS, polygenic risk score; T2D, type 2 diabetes.

Figures
Figure 1

Clinical PRS pipeline for prospectively collected samples. Genotype array data are imputed against 1000 Genomes Project data and used to calculate published PRS (PRSraw, bottom of figure). PRSraw is adjusted for population structure and standardized as described in the text (PRSstd-adj). High-risk status for each disease is determined using published thresholds for OR>2. A parallel pipeline (top of figure) annotates and filters variants for potentially returnable variants in the ACMG SF v2.0 gene list. Variants are manually classified by genetic counselors and laboratory geneticists and confirmed using Sanger sequencing. Results from both components of the assay are included on the laboratory report which clearly details the high-level findings. Abbreviations: ACMG-AMP, American College of Medical Genetics and Genomics - Association for Molecular Pathology; OR, odds ratio; PRS, polygenic risk score; VCF, variant call format.
Figure 2

Frequency of disease and high-risk PRS results by race in MGB Biobank. (A) UpSet plot of total cases of each of 6 phenotypes among 36,423 biobank participants (horizontal bars) and the counts of participants with one or more diseases (vertical bars), by reported race. (B) UpSet plot of total counts of high-risk PRS results (population structure-adjusted PRS corresponding to OR>2) for each of 6 diseases (horizontal bars) and the counts of participants with one or more high-risk PRSstd-adj results, by reported race. Abbreviations: AFib, atrial fibrillation; BrCa, breast cancer; CAD, coronary artery disease; CRCa, colorectal cancer; MGB, Mass General Brigham; OR, odds ratio; PrCa, prostate cancer; PRS, polygenic risk score; T2D, type 2 diabetes.
Figure 3

PRS distributions by reported race before and after adjustment for population structure. Plots to the left of each arrow show the distributions of unadjusted published PRS (PRSstd-raw) by race for each of 6 diseases among up to 36,423 MGB Biobank participants. Plots to the right of each arrow show these distributions after adjustment for population structure (PRSstd-adj), as described in the text. Red vertical lines indicate the standardized PRS threshold corresponding to OR>2 for each disease, based on the OR per standard deviation from the original publication. Abbreviations: AFib, atrial fibrillation; BrCa, breast cancer; CAD, coronary artery disease; CRCa, colorectal cancer; MGB, Mass General Brigham; OR, odds ratio; PrCa, prostate cancer; PRS, polygenic risk score; T2D, type 2 diabetes.
Figure 4

Correlation between adjusted PRS and odds of disease. Plots show log(odds) of each of six diseases versus quantile (n=50) of standardized population structure-adjusted PRS (PRSstd-adj) among up to 36,423 MGB Biobank participants. The correlation coefficient, r, is shown in each panel. Abbreviations: AFib, atrial fibrillation; BrCa, breast cancer; CAD, coronary artery disease; CRCa, colorectal cancer; MGB, Mass General Brigham; OR, odds ratio; PrCa, prostate cancer; PRS, polygenic risk score; T2D, type 2 diabetes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplement20210722.pdf
- HMSGenoVAProtocolCLEAN20210610.pdf
- flatVassyepc.pdf
- flatVassysrs.pdf