nature genetics

SUPPLEMENTARY INFORMATION

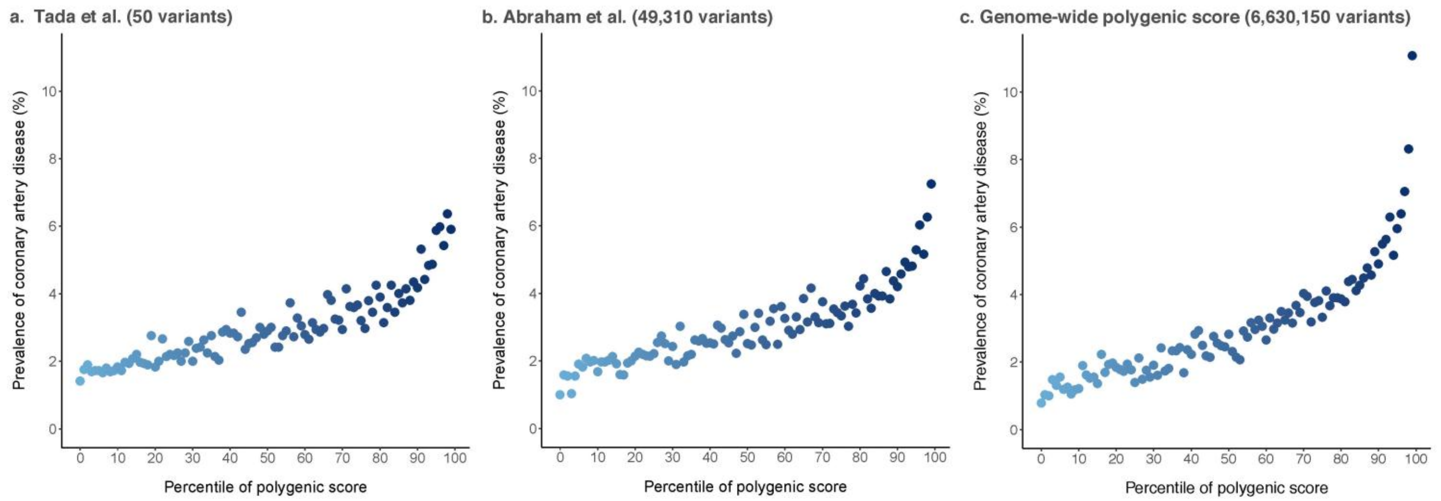
https://doi.org/10.1038/s41588-018-0183-z

In the format provided by the authors and unedited.

Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

Amit V. Khera[1,2,3,4,5], Mark Chaffin[4,5], Krishna G. Aragam[1,2,3,4], Mary E. Haas[4], Carolina Roselli[4], Seung Hoan Choi[4], Pradeep Natarajan[2,3,4], Eric S. Lander[4], Steven A. Lubitz[2,3,4], Patrick T. Ellinor[2,3,4] and Sekar Kathiresan[1,2,3,4]*

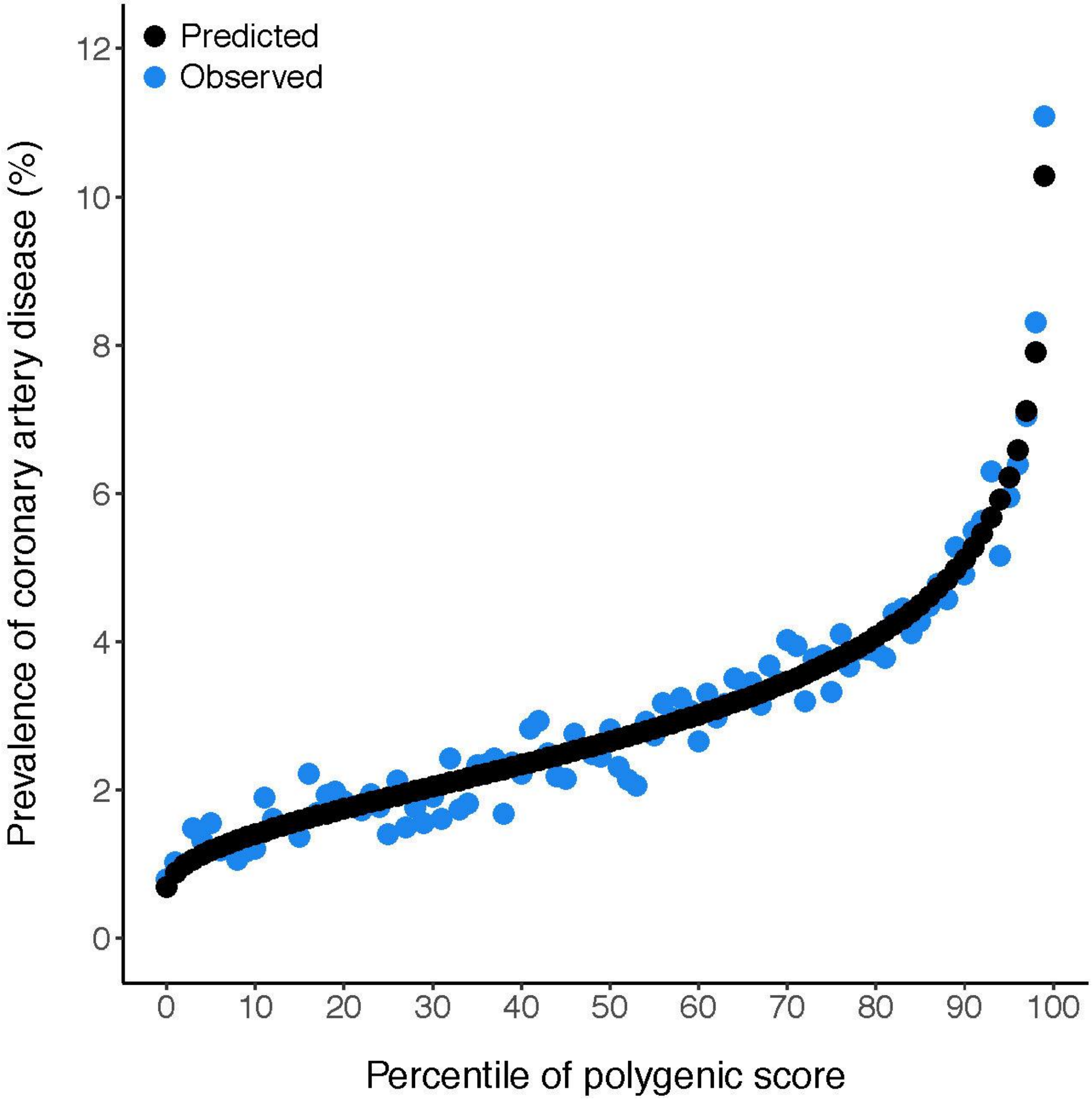
[1]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [2]Cardiology Division of the Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [3]Harvard Medical School, Boston, MA, USA. [4]Cardiovascular Disease Initiative of the Broad Institute of Harvard and MIT, Cambridge, MA, USA. [5]These authors contributed equally: Amit V. Khera, Mark Chaffin. *e-mail: skathiresan1@mgh.harvard.edu



Supplementary Figure 1

Risk gradient for coronary artery disease across the distribution of the genome-wide polygenic score and two previously published scores.

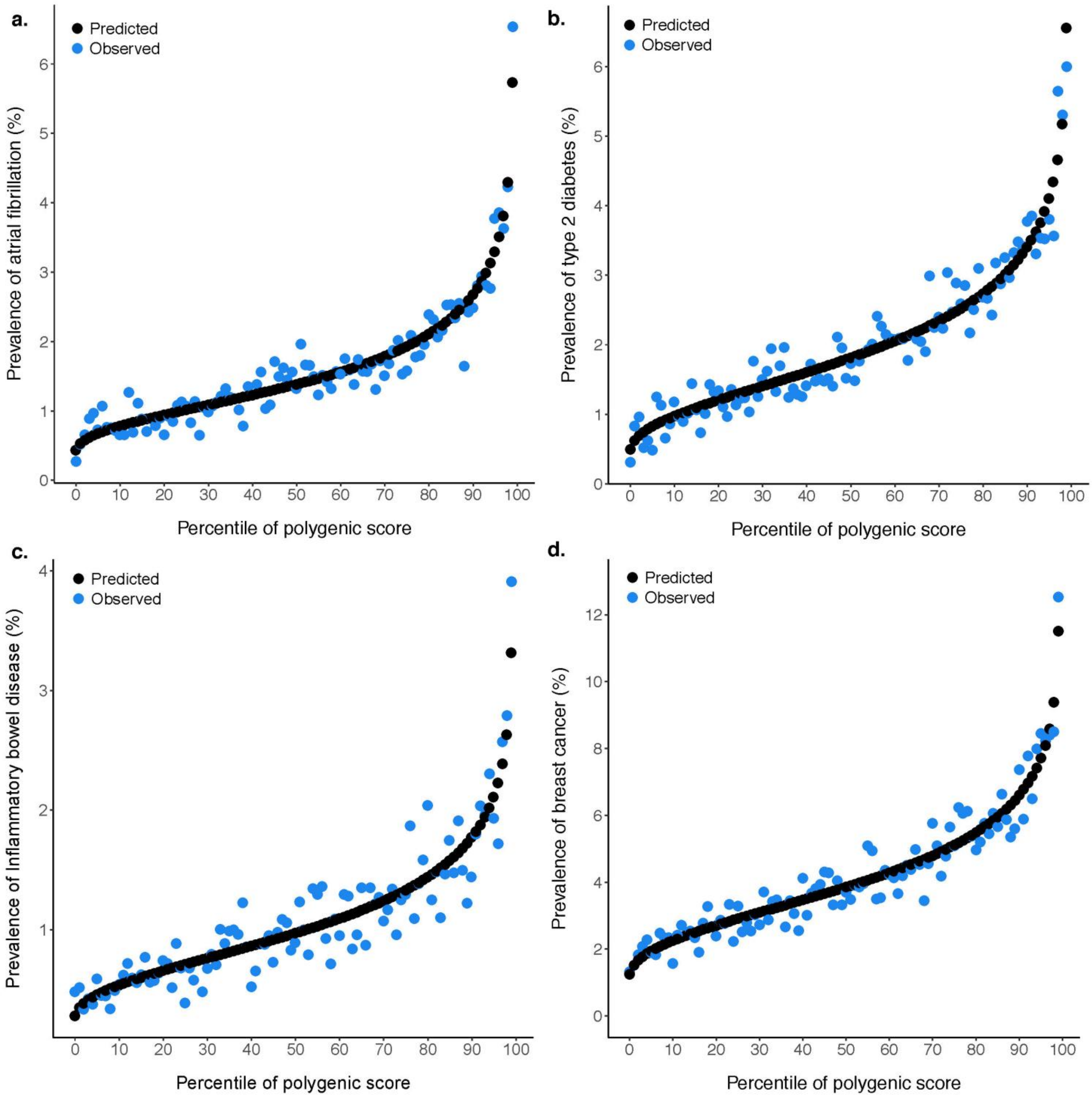
a–**c**, Three polygenic scores for coronary artery disease were calculated within the UK Biobank testing dataset of 288,978 participants: a previously published score comprising 50 variants that had achieved genome-wide levels of statistical significance in previous studies (*Eur. Heart J*. **37**, 561–567, 2016) (**a**); a previously published score comprising 49,310 variants derived from a Metabochip GWAS (*Eur. Heart J*. **37**, 3267–3278, 2016) (**b**); and the newly derived genome-wide polygenic score comprising 6,630,150 variants (**c**). For each score, the population was divided into 100 bins according to percentile of the score and prevalence of coronary artery disease within each bin plotted. The prevalence of coronary artery disease across score percentiles ranged from 1.4% to 5.9% for the 50-variant score, 1.0% to 7.2% for the 49,310-variant score, and 0.8% to 11.1% for the 6,630,150-variant genome-wide polygenic score.



Supplementary Figure 2

Predicted versus observed prevalence of coronary artery disease according to genome-wide polygenic score percentile.

For each individual within the UK Biobank testing dataset, the predicted probability of disease was calculated using a logistic regression model with only the genome-wide polygenic score (GPS) as a predictor. The predicted prevalence of disease within each percentile bin of the GPS distribution was calculated as the average predicted probability of all individuals within that bin. The shape of the predicted risk gradient was consistent with the empirically observed risk gradient, reflected by black and blue dots, respectively.



Supplementary Figure 3

Predicted versus observed prevalence of four diseases according to genome-wide polygenic score percentile.

a–**d**, For each individual within the UK Biobank testing dataset, the predicted probability of disease was calculated using a logistic regression model with only the genome-wide polygenic score (GPS) as a predictor. The predicted prevalence of disease within each percentile bin of the GPS distribution was calculated as the average predicted probability of all individuals within that bin. The shape of the predicted risk gradient was consistent with the empirically observed risk gradient, reflected by black and blue dots, respectively, for each of four diseases: atrial fibrillation (**a**), type 2 diabetes (**b**), inflammatory bowel disease (**c**), and breast cancer (**d**). Breast cancer analysis was restricted to female participants.

Supplementary Table 1. Association of candidate polygenic scores with prevalent coronary artery disease

Derivation Strategy	**Tuning Parameter**	**N Variants Available / N Variants in Score (%)**	**OR per SD (95% CI)**	**AUC**
Genome-wide Significant	$p < 5x10^{-8}$ and $r^2 < 0.2$	74/74 (100.0%)	1.39 (1.35-1.44)	0.791
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.4$	100/100 (100.0%)	1.39 (1.35-1.44)	0.791
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.6$	137/137 (100.0%)	1.39 (1.35-1.44)	0.790
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.8$	204/204 (100.0%)	1.37 (1.33-1.42)	0.789
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.2$	192/192 (100.0%)	1.46 (1.42-1.51)	0.794
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.4$	257/257 (100.0%)	1.47 (1.42-1.52)	0.794
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.6$	345/345 (100.0%)	1.45 (1.41-1.50)	0.793
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.8$	505/505 (100.0%)	1.43 (1.38-1.48)	0.792
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.2$	1269/1273 (99.7%)	1.53 (1.48-1.58)	0.797
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.4$	1590/1594 (99.7%)	1.56 (1.51-1.61)	0.798
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.6$	1997/2001 (99.8%)	1.55 (1.50-1.60)	0.797
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.8$	2706/2710 (99.9%)	1.53 (1.48-1.58)	0.797
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.2$	56941/57276 (99.4%)	1.48 (1.44-1.53)	0.794
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.4$	70491/70831 (99.5%)	1.54 (1.49-1.60)	0.797
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.6$	84921/85264 (99.6%)	1.57 (1.52-1.63)	0.798
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.8$	105595/105942 (99.7%)	1.59 (1.54-1.64)	0.799
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.2$	413921/417670 (99.1%)	1.44 (1.39-1.49)	0.792
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.4$	590581/594406 (99.4%)	1.48 (1.43-1.53)	0.794
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.6$	768415/772288 (99.5%)	1.51 (1.46-1.56)	0.795
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.8$	996630/1000544 (99.6%)	1.53 (1.48-1.58)	0.796
Pruning & Thresholding	$p < 1$ and $r^2 < 0.2$	634268/641894 (98.8%)	1.44 (1.39-1.48)	0.792
Pruning & Thresholding	$p < 1$ and $r^2 < 0.4$	973234/981023 (99.2%)	1.48 (1.43-1.52)	0.794
Pruning & Thresholding	$p < 1$ and $r^2 < 0.6$	1349381/1357303 (99.4%)	1.50 (1.46-1.55)	0.795
Pruning & Thresholding	$p < 1$ and $r^2 < 0.8$	1848045/1856048 (99.6%)	1.52 (1.47-1.57)	0.796
LDPred Algorithm	$\rho = 1$	6629369/6630150 (>99.9%)	1.52 (1.47-1.58)	0.796
LDPred Algorithm	$\rho = 0.3$	6629369/6630150 (>99.9%)	1.53 (1.48-1.58)	0.796
LDPred Algorithm	$\rho = 0.1$	6629369/6630150 (>99.9%)	1.54 (1.49-1.59)	0.796
LDPred Algorithm	$\rho = 0.03$	6629369/6630150 (>99.9%)	1.57 (1.52-1.62)	0.798
LDPred Algorithm	$\rho = 0.01$	6629369/6630150 (>99.9%)	1.62 (1.57-1.68)	0.801
LDPred Algorithm	$\rho = 0.003$	6629369/6630150 (>99.9%)	1.69 (1.63-1.75)	0.805
LDPred Algorithm	**$\rho = 0.001$**	**6629369/6630150 (>99.9%)**	**1.72 (1.67-1.78)**	**0.806**

Odds ratio (OR) per standard deviation (SD) and area under the receiver-operator curve (AUC) were calculated using logistic regression in a validation dataset of 120,280 participants in the UK Biobank (adjusted for age, sex, the first four principal components of ancestry and genotyping array) of which 3,963 had been diagnosed with having coronary artery disease.

p – p-value in discovery GWAS study; r^2 – linkage disequilibrium pruning threshold; ρ – tuning parameter to model the proportion of variants assumed to be causal; OR per SD – odds ratio per standard deviation increment; AUC – area under the receiver operator curve.

Supplementary Table 2. Association of candidate polygenic scores with prevalent atrial fibrillation

Derivation Strategy	Tuning Parameter	N Variants Available / N Variants in Score (%)	OR per SD (95% CI)	AUC
Genome-wide Significant	$p < 5x10^{-8}$ and $r^2 < 0.2$	55/55 (100.0%)	1.48 (1.43-1.54)	0.766
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.4$	78/78 (100.0%)	1.52 (1.46-1.58)	0.768
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.6$	106/106 (100.0%)	1.53 (1.47-1.60)	0.768
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.8$	149/149 (100.0%)	1.55 (1.49-1.62)	0.768
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.2$	161/161 (100.0%)	1.51 (1.45-1.58)	0.767
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.4$	218/218 (100.0%)	1.56 (1.50-1.62)	0.769
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.6$	288/288 (100.0%)	1.58 (1.51-1.64)	0.770
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.8$	383/383 (100.0%)	1.60 (1.53-1.67)	0.770
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.2$	2304/2327 (99.0%)	1.35 (1.29-1.41)	0.754
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.4$	2558/2580 (99.1%)	1.45 (1.38-1.51)	0.759
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.6$	2919/2941 (99.3%)	1.51 (1.44-1.58)	0.763
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.8$	3445/3474 (99.2%)	1.54 (1.47-1.61)	0.765
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.2$	122196/123113 (99.3%)	1.20 (1.15-1.26)	0.748
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.4$	138395/139383 (99.3%)	1.26 (1.20-1.31)	0.750
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.6$	156473/157515 (99.3%)	1.31 (1.25-1.37)	0.753
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.8$	180571/181743 (99.4%)	1.33 (1.27-1.39)	0.754
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.2$	872572/880291 (99.1%)	1.18 (1.13-1.23)	0.747
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.4$	1067307/1075829 (99.2%)	1.23 (1.17-1.28)	0.749
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.6$	1272661/1282064 (99.3%)	1.26 (1.21-1.32)	0.750
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.8$	1522420/1532899 (99.3%)	1.28 (1.22-1.33)	0.751
Pruning & Thresholding	$p < 1$ and $r^2 < 0.2$	1491900/1506103 (99.1%)	1.17 (1.12-1.23)	0.747
Pruning & Thresholding	$p < 1$ and $r^2 < 0.4$	1842010/1857685 (99.2%)	1.22 (1.17-1.28)	0.749
Pruning & Thresholding	$p < 1$and $r^2 < 0.6$	2246065/2263436 (99.2%)	1.26 (1.20-1.32)	0.750
Pruning & Thresholding	$p < 1$ and $r^2 < 0.8$	2765175/2784693 (99.3%)	1.27 (1.22-1.33)	0.751
LDPred Algorithm	$\rho = 1$	6705798/6730541 (99.6%)	1.33 (1.27-1.39)	0.754
LDPred Algorithm	$\rho = 0.3$	6705798/6730541 (99.6%)	1.34 (1.28-1.40)	0.755
LDPred Algorithm	$\rho = 0.1$	6705798/6730541 (99.6%)	1.39 (1.32-1.45)	0.757
LDPred Algorithm	$\rho = 0.03$	6705798/6730541 (99.6%)	1.45 (1.39-1.51)	0.761
LDPred Algorithm	$\rho = 0.01$	6705798/6730541 (99.6%)	1.53 (1.47-1.60)	0.767
LDPred Algorithm	**$\rho = 0.003$**	**6705798/6730541 (99.6%)**	**1.63 (1.56-1.70)**	**0.773**
LDPred Algorithm*	$\rho = 0.001$	6705798/6730541 (99.6%)	1.04 (0.99-1.08)	0.743

*LDPred Algorithm failed to converge

Odds ratio (OR) per standard deviation (SD) and area under the receiver operator curve (AUC) were calculated using logistic regression in a validation dataset of 120,280 participants in the UK Biobank (adjusted for age, sex, the first four principal components of ancestry and genotyping array) of which 2,024 had been diagnosed with atrial fibrillation.

p – p-value in discovery GWAS study; r^2 – linkage disequilibrium pruning threshold; ρ – tuning parameter to model the proportion of variants assumed to be causal; OR per SD – odds ratio per standard deviation increment; AUC – area under the receiver-operator curve.

Supplementary Table 3. Association of candidate polygenic scores with prevalent type 2 diabetes

Derivation Strategy	**Tuning Parameter**	**N Variants Available / N Variants in Score (%)**	**OR per SD (95% CI)**	**AUC**
Genome-wide Significant	$p < 5x10^{-8}$ and $r^2 < 0.2$	72/72 (100.0%)	1.34 (1.30-1.39)	0.700
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.4$	98/98 (100.0%)	1.33 (1.28-1.38)	0.698
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.6$	133/133 (100.0%)	1.31 (1.26-1.36)	0.697
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.8$	201/201 (100.0%)	1.29 (1.25-1.34)	0.695
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.2$	209/209 (100.0%)	1.40 (1.35-1.46)	0.704
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.4$	274/274 (100.0%)	1.40 (1.34-1.45)	0.703
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.6$	388/388 (100.0%)	1.37 (1.32-1.42)	0.701
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.8$	550/551 (99.8%)	1.36 (1.31-1.41)	0.700
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.2$	2838/2913 (97.4%)	1.36 (1.31-1.41)	0.701
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.4$	3269/3346 (97.7%)	1.40 (1.34-1.45)	0.704
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.6$	3858/3937 (98.0%)	1.43 (1.37-1.48)	0.706
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.8$	4832/4912 (98.4%)	1.43 (1.37-1.48)	0.705
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.2$	145622/151854 (95.9%)	1.37 (1.32-1.42)	0.701
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.4$	169289/175728 (96.3%)	1.43 (1.38-1.49)	0.705
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.6$	193703/200323 (96.7%)	1.48 (1.42-1.53)	0.708
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.8$	226545/233313 (97.1%)	1.47 (1.41-1.53)	0.707
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.2$	1049001/1107833 (94.7%)	1.32 (1.27-1.37)	0.697
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.4$	1353005/1414886 (95.6%)	1.38 (1.33-1.44)	0.701
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.6$	1634296/1698631 (96.2%)	1.42 (1.37-1.48)	0.704
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.8$	1959214/2025081 (96.7%)	1.45 (1.39-1.50)	0.705
Pruning & Thresholding	$p < 1$ and $r^2 < 0.2$	1682488/1794860 (93.7%)	1.31 (1.26-1.36)	0.696
Pruning & Thresholding	$p < 1$ and $r^2 < 0.4$	2280565/2399906 (95.0%)	1.37 (1.32-1.42)	0.700
Pruning & Thresholding	$p < 1$and $r^2 < 0.6$	2881225/3006278 (95.8%)	1.42 (1.36-1.47)	0.703
Pruning & Thresholding	$p < 1$ and $r^2 < 0.8$	3575137/3703499 (96.5%)	1.44 (1.39-1.50)	0.706
LDPred Algorithm	$\rho = 1$	6893037/6917436 (99.6%)	1.52 (1.47-1.58)	0.714
LDPred Algorithm	$\rho = 0.3$	6893037/6917436 (99.6%)	1.53 (1.47-1.59)	0.714
LDPred Algorithm	$\rho = 0.1$	6893037/6917436 (99.6%)	1.55 (1.49-1.61)	0.716
LDPred Algorithm	$\rho = 0.03$	6893037/6917436 (99.6%)	1.59 (1.53-1.65)	0.720
LDPred Algorithm	**$\rho = 0.01$**	**6893037/6917436 (99.6%)**	**1.65 (1.59-1.71)**	**0.725**
LDPred Algorithm	$\rho = 0.003$	6893037/6917436 (99.6%)	1.15 (1.11-1.20)	0.687
LDPred Algorithm*	$\rho = 0.001$	6893037/6917436 (99.6%)	1.05 (1.02-1.10)	0.683

*LDPred Algorithm failed to converge

Odds ratio (OR) per standard deviation (SD) and area under the receiver-operator curve (AUC) were calculated using logistic regression in a validation dataset of 120,280 participants in the UK Biobank (adjusted for age, sex, the first four principal components of ancestry and genotyping array) of which 2,785 had been diagnosed with type 2 diabetes.

p – p-value in discovery GWAS study; r^2 – linkage disequilibrium pruning threshold; ρ – tuning parameter to model the proportion of variants assumed to be causal. OR per SD – odds ratio per standard deviation increment; AUC – area under the receiver-operator curve.

Supplementary Table 4. Association of candidate polygenic scores with prevalent inflammatory bowel disease

Derivation Strategy	**Tuning Parameter**	**N Variants Available / N Variants in Score (%)**	**OR per SD (95% CI)**	**AUC**
Genome-wide Significant	$p < 5x10^{-8}$ and $r^2 < 0.2$	288/292 (98.6%)	1.40 (1.34-1.47)	0.614
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.4$	475/484 (98.1%)	1.31 (1.24-1.38)	0.582
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.6$	800/812 (98.5%)	1.23 (1.17-1.30)	0.567
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.8$	1529/1545 (99.0%)	1.18 (1.11-1.24)	0.557
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.2$	520/533 (97.6%)	1.43 (1.37-1.50)	0.625
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.4$	857/875 (97.9%)	1.36 (1.29-1.43)	0.591
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.6$	1334/1356 (98.4%)	1.26 (1.19-1.33)	0.572
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.8$	2391/2418 (98.9%)	1.19 (1.13-1.26)	0.560
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.2$	2979/3028 (98.4%)	1.54 (1.46-1.62)	0.631
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.4$	3817/3875 (98.5%)	1.45 (1.38-1.53)	0.610
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.6$	4949/5013 (98.7%)	1.34 (1.27-1.42)	0.587
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.8$	7111/7185 (99.0%)	1.24 (1.17-1.30)	0.569
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.2$	118775/121914 (97.4%)	1.53 (1.44-1.61)	0.616
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.4$	140825/144087 (97.7%)	1.58 (1.50-1.67)	0.629
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.6$	163967/167349 (98.0%)	1.54 (1.46-1.63)	0.623
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.8$	195815/199334 (98.2%)	1.39 (1.31-1.46)	0.597
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.2$	812741/842603 (96.5%)	1.46 (1.37-1.55)	0.598
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.4$	1066545/1098071 (97.1%)	1.50 (1.42-1.59)	0.608
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.6$	1308728/1341631 (97.5%)	1.53 (1.44-1.61)	0.616
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.8$	1602425/1636580 (97.9%)	1.46 (1.39-1.55)	0.610
Pruning & Thresholding	$p < 1$ and $r^2 < 0.2$	1291770/1349599 (95.7%)	1.45 (1.36-1.54)	0.597
Pruning & Thresholding	$p < 1$ and $r^2 < 0.4$	1783031/1844513 (96.7%)	1.49 (1.41-1.58)	0.607
Pruning & Thresholding	$p < 1$and $r^2 < 0.6$	2291513/2356075 (97.3%)	1.52 (1.44-1.61)	0.615
Pruning & Thresholding	$p < 1$ and $r^2 < 0.8$	2917090/2984351 (97.7%)	1.47 (1.39-1.55)	0.610
LDPred Algorithm	$\rho = 1$	6882324/6907112 (99.6%)	1.58 (1.49-1.66)	0.628
LDPred Algorithm	$\rho = 0.3$	6882324/6907112 (99.6%)	1.58 (1.50-1.67)	0.629
LDPred Algorithm	**$\rho = 0.1$**	**6882324/6907112 (99.6%)**	**1.61 (1.52-1.70)**	**0.633**
LDPred Algorithm	$\rho = 0.03$	6882324/6907112 (99.6%)	1.55 (1.47-1.64)	0.625
LDPred Algorithm	$\rho = 0.01$	6882324/6907112 (99.6%)	1.28 (1.22-1.35)	0.580
LDPred Algorithm*	$\rho = 0.003$	6882324/6907112 (99.6%)	1.21 (1.15-1.27)	0.563
LDPred Algorithm*	$\rho = 0.001$	6882324/6907112 (99.6%)	1.16 (1.10-1.23)	0.556

*LDPred Algorithm failed to converge

Odds ratio (OR) per standard deviation (SD) and area under the receiver-operator curve (AUC) were calculated using logistic regression in a validation dataset of 120,280 participants in the UK Biobank (adjusted for age, sex, the first four principal components of ancestry and genotyping array) of which 1,360 had been diagnosed with inflammatory bowel disease.

p – p-value in discovery GWAS study; r^2 – linkage disequilibrium pruning threshold; ρ – tuning parameter to model the proportion of variants assumed to be causal; OR per SD – odds ratio per standard deviation increment; AUC – area under the receiver-operator curve.

Supplementary Table 5. Association of candidate polygenic scores with prevalent breast cancer

Derivation Strategy	**Tuning Parameter**	**N Variants Available / N Variants in Score (%)**	**OR per SD (95% CI)**	**AUC**
Genome-wide Significant	$p < 5x10^{-8}$ and $r^2 < 0.2$	572/577 (99.1%)	1.47 (1.42-1.53)	0.677
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.4$	878/884 (99.3%)	1.44 (1.39-1.50)	0.673
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.6$	1284/1292 (99.4%)	1.39 (1.34-1.45)	0.666
Pruning & Thresholding	$p < 5x10^{-8}$ and $r^2 < 0.8$	1959/1971 (99.4%)	1.39 (1.33-1.45)	0.666
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.2$	1151/1165 (98.8%)	1.51 (1.45-1.57)	0.681
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.4$	1692/1712 (98.8%)	1.48 (1.42-1.54)	0.677
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.6$	2382/2411 (98.8%)	1.43 (1.38-1.49)	0.671
Pruning & Thresholding	$p < 5x10^{-6}$ and $r^2 < 0.8$	3588/3624 (99.0%)	1.43 (1.37-1.49)	0.671
Pruning & Thresholding	**$p < 5x10^{-4}$ and $r^2 < 0.2$**	**5158/5218 (98.9%)**	**1.56 (1.49-1.62)**	**0.685**
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.4$	6868/6942 (98.9%)	1.55 (1.49-1.61)	0.684
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.6$	8945/9036 (99.0%)	1.51 (1.45-1.57)	0.679
Pruning & Thresholding	$p < 5x10^{-4}$ and $r^2 < 0.8$	12352/12461 (99.1%)	1.50 (1.44-1.56)	0.678
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.2$	114421/115503 (99.1%)	1.45 (1.39-1.50)	0.672
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.4$	143235/144508 (99.1%)	1.49 (1.43-1.55)	0.677
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.6$	173750/175238 (99.2%)	1.50 (1.44-1.56)	0.678
Pruning & Thresholding	$p < 5x10^{-2}$ and $r^2 < 0.8$	217554/219334 (99.2%)	1.51 (1.45-1.57)	0.678
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.2$	657758/663879 (99.1%)	1.38 (1.33-1.44)	0.665
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.4$	910344/918115 (99.2%)	1.41 (1.36-1.47)	0.668
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.6$	1157487/1166909 (99.2%)	1.43 (1.38-1.49)	0.670
Pruning & Thresholding	$p < 5x10^{-1}$ and $r^2 < 0.8$	1471670/1483324 (99.2%)	1.45 (1.39-1.51)	0.671
Pruning & Thresholding	$p < 1$ and $r^2 < 0.2$	997491/1007125 (99.0%)	1.38 (1.32-1.43)	0.664
Pruning & Thresholding	$p < 1$ and $r^2 < 0.4$	1469656/1482406 (99.1%)	1.41 (1.35-1.47)	0.668
Pruning & Thresholding	$p < 1$and $r^2 < 0.6$	1968975/1984988 (99.2%)	1.43 (1.37-1.49)	0.669
Pruning & Thresholding	$p < 1$ and $r^2 < 0.8$	2612769/2633156 (99.2%)	1.44 (1.38-1.50)	0.670
LDPred Algorithm	$\rho = 1$	7227160/7261712 (99.5%)	1.47 (1.41-1.53)	0.674
LDPred Algorithm	$\rho = 0.3$	7227160/7261712 (99.5%)	1.51 (1.45-1.57)	0.678
LDPred Algorithm	$\rho = 0.1$	7227160/7261712 (99.5%)	1.52 (1.46-1.59)	0.679
LDPred Algorithm	$\rho = 0.03$	7227160/7261712 (99.5%)	1.30 (1.25-1.35)	0.657
LDPred Algorithm*	$\rho = 0.01$	7227160/7261712 (99.5%)	1.18 (1.14-1.23)	0.646
LDPred Algorithm*	$\rho = 0.003$	7227160/7261712 (99.5%)	1.12 (1.08-1.17)	0.642
LDPred Algorithm*	$\rho = 0.001$	7227160/7261712 (99.5%)	1.13 (1.08-1.17)	0.642

*LDPred Algorithm failed to converge

Odds ratio (OR) per standard deviation (SD) and area under the curve (AUC) were calculated using logistic regression in a validation dataset of 63,347 female participants in the UK Biobank (adjusted for age, the first four principal components of ancestry and genotyping array) of which 2,576 had been diagnosed with having breast cancer.

p – p-value in discovery GWAS study; r^2 – linkage disequilibrium pruning threshold; ρ – tuning parameter to model the proportion of variants assumed to be causal; OR per SD – odds ratio per standard deviation increment; AUC – area under the receiver-operator curve.

Supplementary Table 6. Genome-wide polygenic score characteristics for five diseases across derivation strategies.

For each disease, characteristics of genome-wide polygenic scores (GPSs) are displayed according to derivation strategy of GWAS significant variants only (pruning and thresholding with $p < 5x10^{-8}$ and $r^2 < 0.2$), the best of the remaining 23 pruning and thresholding GPSs, and the best of 7 LDPred GPSs. The score with the highest area under the receiver-operator curve (denoted by bolded font) was carried forward to the testing dataset.

Disease	**Derivation strategy**	**N variants available / N variants in score (%)**	**Tuning parameters**	**AUC (95%CI)**
Coronary artery disease	GWAS significant variants	74 / 74 (100%)	$p < 5x10^{-8}$, $r^2 < 0.2$	0.791 (0.785 – 0.798)
Coronary artery disease	Pruning and thresholding	105,942 / 105,595 (99.67%)	$p < 0.05$, $r^2 < 0.8$	0.799 (0.793 – 0.806)
Coronary artery disease	**LDPred**	**6,629,369 / 6,630,150 (99.99%)**	**$\rho = 0.001$**	**0.806 (0.800 – 0.813)**
Atrial fibrillation	GWAS significant variants	55 / 55 (100%)	$p < 5x10^{-8}$, $r^2 < 0.2$	0.766 (0.757 – 0.776)
Atrial fibrillation	Pruning and thresholding	383 / 383 (100%)	$p < 5x10^{-6}$, $r^2 < 0.8$	0.770 (0.760 – 0.780)
Atrial fibrillation	**LDPred**	**6,705,798 / 6,730,541 (99.63%)**	**$\rho = 0.003$**	**0.773 (0.763 – 0.782)**
Type 2 diabetes	GWAS significant variants	72 / 72 (100%)	$p < 5x10^{-8}$, $r^2 < 0.2$	0.700 (0.690 – 0.709)
Type 2 diabetes	Pruning and thresholding	193,703 / 200,323 (96.7%)	$p < 0.05$, $r^2 < 0.6$	0.708 (0.699 – 0.717)
Type 2 diabetes	**LDPred**	**6,893,037 / 6,917,436 (99.65%)**	**$\rho = 0.01$**	**0.725 (0.716 – 0.734)**
Inflammatory bowel disease	GWAS significant variants	288 / 292 (98.6%)	$p < 5x10^{-8}$, $r^2 < 0.2$	0.614 (0.600 – 0.629)
Inflammatory bowel disease	Pruning and thresholding	2979 / 3028 (98.4%)	$p < 5x10^{-4}$, $r^2 < 0.2$	0.631 (0.619 – 0.645)
Inflammatory bowel disease	**LDPred**	**6,882,324 / 6,907,112 (99.64%)**	**$\rho = 0.1$**	**0.633 (0.619 – 0.648)**
Breast cancer	GWAS significant variants	572 / 577 (99.1%)	$p < 5x10^{-8}$, $r^2 < 0.2$	0.677 (0.667 – 0.687)
Breast cancer	**Pruning and thresholding**	**5158 / 5218 (98.85%)**	**$p < 5x10^{-4}$, $r^2 < 0.2$**	**0.685 (0.675 – 0.695)**
Breast cancer	LDPred	7,227,160 / 7,261,712 (99.5%)	$\rho = 0.1$	0.679 (0.669 – 0.689)

Supplementary Table 7. Comparison of GPS_{CAD} to two previously published polygenic scores for coronary artery disease

High GPS definition	Reference group	Odds ratio	95% Confidence interval	P-value
Tada et al.[1] (50 variants)				
Top 20% of distribution	Remaining 80%	1.86	1.78 – 1.95	2.1×10^{-143}
Top 10% of distribution	Remaining 90%	2.09	1.97 – 2.22	4.5×10^{-136}
Top 5% of distribution	Remaining 95%	2.26	2.09 – 2.43	8.6×10^{-100}
Top 1% of distribution	Remaining 99%	2.24	1.90 – 2.62	1.7×10^{-22}
Top 0.5% of distribution	Remaining 99.5%	2.31	1.83 – 2.88	3.7×10^{-13}
Abraham et al.[2] (49,310 variants)				
Top 20% of distribution	Remaining 80%	1.94	1.85 – 2.03	3.2×10^{-163}
Top 10% of distribution	Remaining 90%	2.07	1.95 – 2.19	4.5×10^{-132}
Top 5% of distribution	Remaining 95%	2.28	2.12 – 2.46	1.8×10^{-103}
Top 1% of distribution	Remaining 99%	2.71	2.33 – 3.14	2.1×10^{-39}
Top 0.5% of distribution	Remaining 99.5%	2.55	2.04 – 3.14	1.7×10^{-17}
GPS (6,630,100 variants)				
Top 20% of distribution	Remaining 80%	2.55	2.43 – 2.67	$< 1 \times 10^{-300}$
Top 10% of distribution	Remaining 90%	2.89	2.74 – 3.05	$< 1 \times 10^{-300}$
Top 5% of distribution	Remaining 95%	3.34	3.12 – 3.58	6.5×10^{-264}
Top 1% of distribution	Remaining 99%	4.83	4.25 – 5.46	1.0×10^{-132}
Top 0.5% of distribution	Remaining 99.5%	5.17	4.34 – 6.12	7.9×10^{-78}

GPS – genome-wide polygenic score
50 of 50 (100%) of the variants included in the Tada et al.[1] score were available in the UK Biobank testing dataset. 49,297 of 49,310 (99.97%) of the variants included in the Abraham et al.[2] score were available in the UK Biobank testing dataset. 6,630,100 / 6,630,150 (>99.9%) of the variants included in the GPS were available in the UK Biobank testing dataset. Odds ratios calculated by comparing those with high GPS to the remainder of the population in a logistic regression model adjusted for age, sex, genotyping array, and the first four principal components of ancestry.

Supplementary Table 8. Baseline characteristics according to high genome-wide polygenic score for coronary artery disease

Baseline characteristics according to high coronary artery disease polygenic score status, defined as the top 8% of the distribution empirically shown to be at ≥3-fold risk of CAD. Values displayed are mean (standard deviation) for continuous variables and N (%) for categorical variables.
GPS_{CAD} – genome-wide polygenic score for coronary artery disease

	Remainder of population	**Top 8% of GPS_{CAD} distribution**	**P-value**
Number of individuals	265,859	23,119	
Coronary artery disease	7,061 (2.7%)	1,615 (7.0%)	< 0.001
Age, years	56.9 (8.0)	56.7 (8.1)	< 0.001
Male sex	120,673 (45%)	10,410 (45%)	0.29
Hypertension	73,982 (28%)	7,477 (32%)	< 0.001
Type 2 diabetes	5,240 (2.0%)	613 (2.7%)	< 0.001
Hypercholesterolemia	35,042 (13%)	4,559 (20%)	< 0.001
Current smoking	24,399 (9.2%)	2,200 (9.5%)	0.09
Family history of heart disease	94,117 (35%)	10,101 (44%)	< 0.001
Body mass index, kg/m^2	27.3 (4.7)	27.6 (4.8)	< 0.001
Lipid-lowering therapy	43,923 (17%)	5,589 (24%)	< 0.001

Supplementary Table 9. Assessment of genome-wide polygenic scores in the testing dataset.

Disease	N variants available / N variants in score (%)	Proportion of variance explained (%)
Coronary artery disease	6,630,100 / 6,630,150 (> 99.9%)	4.0%
Atrial fibrillation	6,722,280 / 6,730,541 (99.9%)	2.9%
Type 2 diabetes	6,909,367 / 6,917,436 (99.9%)	2.9%
Inflammatory bowel disease	6,899,007/6,907,112 (99.9%)	2.1%
Breast cancer	5,186 / 5,218 (99.4%)	2.7%

Proportion of variance explained was calculated for each disease using the Nagelkerke's pseudo-R^2 metric. The R^2 was calculated for the full model inclusive of the genome-wide polygenic score plus the covariates minus R^2 for the covariates alone, thus yielding an estimate of the explained variance attributable to the polygenic score. Covariates in the model included age, gender, genotyping array, and the first four principal components of ancestry.

Supplementary Table 10. Prevalence and clinical impact of a high genome-wide polygenic score in unrelated individuals

High GPS definition	Reference group	Odds ratio	95% Confidence interval	P-value
Coronary artery disease				
Top 20% of distribution	Remaining 80%	2.53	2.42 – 2.66	< 1 x 10^{-300}
Top 10% of distribution	Remaining 90%	2.90	2.74 – 3.07	< 1 x 10^{-300}
Top 5% of distribution	Remaining 95%	3.34	3.11 – 3.58	1.6 x 10^{-244}
Top 1% of distribution	Remaining 99%	4.53	3.95 – 5.17	5.2 x 10^{-108}
Top 0.5% of distribution	Remaining 99.5%	5.18	4.31 – 6.20	1.6 x 10^{-70}
Atrial fibrillation				
Top 20% of distribution	Remaining 80%	2.47	2.31 – 2.65	6.7 x 10^{-150}
Top 10% of distribution	Remaining 90%	2.74	2.52 – 2.96	7.2 x 10^{-136}
Top 5% of distribution	Remaining 95%	3.17	2.87 – 3.49	5.4 x 10^{-119}
Top 1% of distribution	Remaining 99%	4.42	3.78 – 5.36	1.4 x 10^{-64}
Top 0.5% of distribution	Remaining 99.5%	5.27	4.15 – 6.60	4.4 x 10^{-45}
Type 2 diabetes				
Top 20% of distribution	Remaining 80%	2.37	2.23 – 2.52	4.2 x 10^{-168}
Top 10% of distribution	Remaining 90%	2.52	2.35 – 2.71	2.3 x 10^{-138}
Top 5% of distribution	Remaining 95%	2.77	2.53 – 3.03	1.5 x 10^{-106}
Top 1% of distribution	Remaining 99%	3.36	2.81 – 3.99	1.8 x 10^{-41}
Top 0.5% of distribution	Remaining 99.5%	3.42	2.67 – 4.33	2.5 x 10^{-23}
Inflammatory bowel disease				
Top 20% of distribution	Remaining 80%	2.19	2.01 – 2.38	9.1 x 10^{-73}
Top 10% of distribution	Remaining 90%	2.51	2.27 – 2.77	4.1 x 10^{-74}
Top 5% of distribution	Remaining 95%	2.75	2.42 – 3.10	1.9 x 10^{-57}
Top 1% of distribution	Remaining 99%	3.72	2.96 – 4.62	8.4 x 10^{-31}
Top 0.5% of distribution	Remaining 99.5%	4.47	3.31 – 5.89	1.4 x 10^{-24}
Breast cancer				
Top 20% of distribution	Remaining 80%	2.08	1.96 – 2.21	3.2 x 10^{-122}
Top 10% of distribution	Remaining 90%	2.36	2.20 – 2.54	6.8 x 10^{-118}
Top 5% of distribution	Remaining 95%	2.59	2.36 – 2.84	1.5 x 10^{-89}
Top 1% of distribution	Remaining 99%	3.47	2.91 – 4.12	4.4 x 10^{-45}
Top 0.5% of distribution	Remaining 99.5%	3.78	2.97 – 4.75	9.7 x 10^{-29}

GPS – genome-wide polygenic score
A sensitivity analysis was performed in 222,529 of 288,978 (77%) of the testing dataset after excluding one of each pair of related individuals (third-degree or closer). Odds ratios calculated by comparing those with high GPS to the remainder of the population in a logistic regression model adjusted for age, sex, genotyping array, and the first four principal components of ancestry. Breast cancer analysis was restricted to female participants.

Supplementary References

1. Tada H, *et al*. Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history. *Eur Heart J*. **37**, 561-7 (2016).
2. Abraham G., *et al*. Genomic prediction of coronary heart disease. *Eur Heart J*. **37**, 3267-3278 (2016).