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]*

A key public health need is to identify individuals at high risk for a given disease to enable enhanced screening or preventive therapies. Because most common diseases have a genetic component, one important approach is to stratify individuals based on inherited DNA variation[1]. Proposed clinical applications have largely focused on finding carriers of rare monogenic mutations at several-fold increased risk. Although most disease risk is polygenic in nature[2–5], it has not yet been possible to use polygenic predictors to identify individuals at risk comparable to monogenic mutations. Here, we develop and validate genome-wide polygenic scores for five common diseases. The approach identifies 8.0, 6.1, 3.5, 3.2, and 1.5% of the population at greater than threefold increased risk for coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer, respectively. For coronary artery disease, this prevalence is 20-fold higher than the carrier frequency of rare monogenic mutations conferring comparable risk[6]. We propose that it is time to contemplate the inclusion of polygenic risk prediction in clinical care, and discuss relevant issues.

For various common diseases, genes have been identified in which rare mutations confer several-fold increased risk in heterozygous carriers. An important example is the presence of a familial hypercholesterolemia mutation in 0.4% of the population, which confers an up to threefold increased risk for coronary artery disease (CAD)[6]. Aggressive treatment to lower circulating cholesterol levels among such carriers can significantly reduce risk[7]. Another example is the p.Glu508Lys missense mutation in *HNF1A*, with a carrier frequency of 0.1% of the general population and 0.7% of Latinos[8], which confers up to fivefold increased risk for type 2 diabetes[9]. Although the ascertainment of monogenic mutations can be highly relevant for carriers and their families, the vast majority of disease occurs in those without such mutations.

For most common diseases, polygenic inheritance, involving many common genetic variants of small effect, plays a greater role than rare monogenic mutations[2–5]. However, it has been unclear whether it is possible to create a genome-wide polygenic score (GPS) to identify individuals at clinically significantly increased risk—for example, comparable to levels conferred by rare monogenic mutations[10,11].

Previous studies to create GPSs had only limited success, providing insufficient risk stratification for clinical utility (for example, identifying 20% of a population at 1.4-fold increased risk relative to the rest of the population)[12]. These initial efforts were hampered by three challenges: (1) the small size of initial genome-wide association studies (GWASs), which affected the precision of the estimated impact of individual variants on disease risk; (2) limited computational methods for creating GPSs; and (3) a lack of large datasets needed to validate and test GPS.

Using much larger studies and improved algorithms, we set out to revisit the question of whether a GPS can identify subgroups of the population with risk approaching or exceeding that of a monogenic mutation. We studied five common diseases with major public health impact: CAD, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer.

For each of the diseases, we created several candidate GPSs based on summary statistics and imputation from recent large GWASs in participants of primarily European ancestry (Table 1). Specifically, we derived 24 predictors based on a pruning and thresholding method, and 7 additional predictors using the recently described LDPred algorithm[13] (Methods, Fig. 1 and Supplementary Tables 1–6). These scores were validated and tested within the UK Biobank, which has aggregated genotype data and extensive phenotypic information on 409,258 participants of British ancestry (average age: 57 years; 55% female)[14,15].

We used an initial validation dataset of the 120,280 participants in the UK Biobank phase 1 genotype data release to select the GPSs with the best performance, defined as the maximum area under the receiver-operator curve (AUC). We then assessed the performance in an independent testing dataset comprised of the 288,978 participants in the UK Biobank phase 2 genotype data release. For each disease, the discriminative capacity within the testing dataset was nearly identical to that observed in the validation dataset.

Taking CAD as an example, our polygenic predictors were derived from a GWAS involving 184,305 participants[16] and evaluated based on their ability to detect the participants in the UK Biobank validation dataset diagnosed with CAD (Table 1). The predictors had AUCs ranging from 0.79–0.81 in the validation set, with the best predictor (GPS$_{CAD}$) involving 6,630,150 variants (Supplementary Table 1). This predictor performed equivalently well in the testing dataset, with an AUC of 0.81.

[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

Table 1 | GPS derivation and testing for five common, complex diseases

Disease	Discovery GWAS (n)	Prevalence in validation dataset	Prevalence in testing dataset	Polymorphisms in GPS	Tuning parameter	AUC (95% CI) in validation dataset	AUC (95% CI) in testing dataset
CAD	60,801 cases; 123,504 controls[16]	3,963/120,280 (3.4%)	8,676/288,978 (3.0%)	6,630,150	LDPred ($\rho = 0.001$)	0.81 (0.80–0.81)	0.81 (0.81–0.81)
Atrial fibrillation	17,931 cases; 115,142 controls[30]	2,024/120,280 (1.7%)	4,576/288,978 (1.6%)	6,730,541	LDPred ($\rho = 0.003$)	0.77 (0.76–0.78)	0.77 (0.76–0.77)
Type 2 diabetes	26,676 cases; 132,532 controls[31]	2,785/120,280 (2.4%)	5,853/288,978 (2.0%)	6,917,436	LDPred ($\rho = 0.01$)	0.72 (0.72–0.73)	0.73 (0.72–0.73)
Inflammatory bowel disease	12,882 cases; 21,770 controls[32]	1,360/120,280 (1.1%)	3,102/288,978 (1.1%)	6,907,112	LDPred ($\rho = 0.1$)	0.63 (0.62–0.65)	0.63 (0.62–0.64)
Breast cancer	122,977 cases; 105,974 controls[33]	2,576/63,347 (4.1%)	6,586/157,895 (4.2%)	5,218	Pruning and thresholding ($r^2 < 0.2$; $P < 5 \times 10^{-4}$)	0.68 (0.67–0.69)	0.69 (0.68–0.69)

AUC was determined using a logistic regression model adjusted for age, sex, genotyping array, and the first four principal components of ancestry. The breast cancer analysis was restricted to female participants. For the LDPred algorithm, the tuning parameter ρ reflects the proportion of polymorphisms assumed to be causal for the disease. For the pruning and thresholding strategy, r^2 reflects the degree of independence from other variants in the linkage disequilibrium reference panel, and P reflects the P value noted for a given variant in the discovery GWAS. CI, confidence interval.
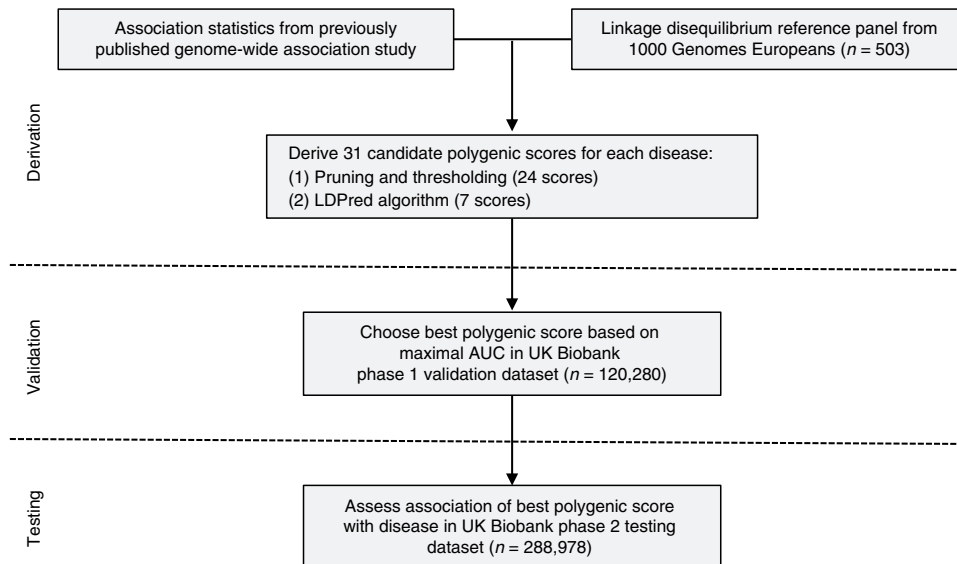

Fig. 1 | Study design and workflow. A GPS for each disease was derived by combining summary association statistics from a recent large GWAS and a linkage disequilibrium reference panel of 503 Europeans[34]. Then, 31 candidate GPSs were derived using two strategies: (1) 'pruning and thresholding' (that is, the aggregation of independent polymorphisms that exceeded a specified level of significance in the discovery GWAS); and (2) the LDPred computational algorithm[13], a Bayesian approach to calculate a posterior mean effect for all variants based on a prior (effect size in the previous GWAS) and subsequent shrinkage based on linkage disequilibrium. The seven candidate LDPred scores vary with respect to the tuning parameter ρ (that is, the proportion of variants assumed to be causal), as previously recommended[13]. The optimal GPS for each disease was chosen based on the AUC in the UK Biobank phase 1 validation dataset ($n = 120,280$ Europeans) and subsequently calculated in an independent UK Biobank phase 2 testing dataset ($n = 288,978$ Europeans).

We then investigated whether our polygenic predictor, GPS$_{CAD}$, could identify individuals at similar risk to the threefold increased risk conferred by a familial hypercholesterolemia mutation[6]. Across the population, GPS$_{CAD}$ is normally distributed with the empirical risk of CAD rising sharply in the right tail of the distribution, from 0.8% in the lowest percentile to 11.1% in the highest percentile (Fig. 2). The median GPS$_{CAD}$ percentile score was 69 for individuals with CAD versus 49 for individuals without CAD. By analogy to the
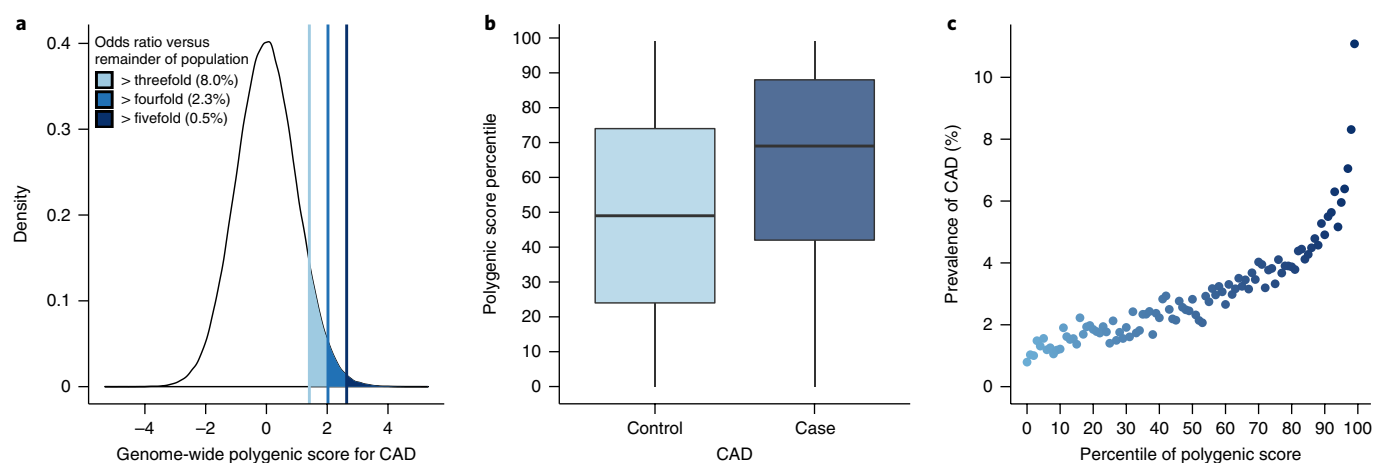
Fig. 2 | Risk for CAD according to GPS. a, Distribution of GPS_{CAD} in the UK Biobank testing dataset ($n = 288{,}978$). The x axis represents GPS_{CAD}, with values scaled to a mean of 0 and a standard deviation of 1 to facilitate interpretation. Shading reflects the proportion of the population with three-, four-, and fivefold increased risk versus the remainder of the population. The odds ratio was assessed in a logistic regression model adjusted for age, sex, genotyping array, and the first four principal components of ancestry. **b**, GPS_{CAD} percentile among CAD cases versus controls in the UK Biobank testing dataset. Within each boxplot, the horizontal lines reflect the median, the top and bottom of each box reflect the interquartile range, and the whiskers reflect the maximum and minimum values within each grouping. **c**, Prevalence of CAD according to 100 groups of the testing dataset binned according to the percentile of the GPS_{CAD}.

traditional analytic strategy for monogenic mutations, we defined 'carriers' as individuals with GPS_{CAD} above a given threshold and 'non-carriers' as all others.

We found that 8% of the population had inherited a genetic predisposition that conferred $\geq$ threefold increased risk for CAD (Table 2). Strikingly, the polygenic score identified 20-fold more people at comparable or greater risk than were found by familial hypercholesterolemia mutations in previous studies[6,7]. Moreover, 2.3% of the population ('carriers') had inherited $\geq$ fourfold increased risk for CAD and 0.5% ('carriers') had inherited $\geq$ fivefold increased risk. GPS_{CAD} performed substantially better than two previously published polygenic scores for CAD that included 50 and 49,310 variants, respectively (Supplementary Table 7 and Supplementary Fig. 1)[17,18].

GPS_{CAD} has the advantage that it can be assessed from the time of birth, well before the discriminative capacity emerges for the risk factors (for example, hypertension or type 2 diabetes) used in clinical practice to predict CAD. Moreover, even for our middle-aged study population, practising clinicians could not identify the 8% of individuals at $\geq$ threefold risk based on GPS_{CAD} using conventional risk factors in the absence of genotype information (Supplementary Table 8). For example, conventional risk factors such as hypercholesterolemia were present in 20% of those with $\geq$ threefold risk based on GPS_{CAD} versus 13% of those in the remainder of the distribution. Hypertension was present in 32 versus 28%, and a family history of heart disease was present in 44 versus 35%, respectively. Making high GPS_{CAD} individuals aware of their inherited susceptibility may facilitate intensive prevention efforts. For example, we previously showed that a high polygenic risk for CAD may be offset by one of two interventions: adherence to a healthy lifestyle or cholesterol-lowering therapy with statin medications[19–21].

Our results for CAD generalized to the four other diseases: risk increased sharply in the right tail of the GPS distribution (Fig. 3). For each disease, the shape of the observed risk gradient was consistent with predicted risk based only on the GPS (Supplementary Figs. 2 and 3).

Atrial fibrillation is an underdiagnosed and often asymptomatic disorder in which an irregular heart rhythm predisposes to blood clots and is a leading cause of ischemic stroke[22]. The polygenic

Table 2 | Proportion of the population at three-, four- and fivefold increased risk for each of the five common diseases

High GPS definition	Individuals in testing dataset (n)	% of individuals
Odds ratio ≥3.0		
CAD	23,119/288,978	8.0
Atrial fibrillation	17,627/288,978	6.1
Type 2 diabetes	10,099 288,978	3.5
Inflammatory bowel disease	9,209 288,978	3.2
Breast cancer	2,369/157,895	1.5
Any of the five diseases	57,115/288,978	19.8
Odds ratio ≥4.0		
CAD	6,631/288,978	2.3
Atrial fibrillation	4,335/288,978	1.5
Type 2 diabetes	578/288,978	0.2
Inflammatory bowel disease	2,297/288,978	0.8
Breast cancer	474/157,895	0.3
Any of the five diseases	14,029/288,978	4.9
Odds ratio ≥5.0		
CAD	1,443/288,978	0.5
Atrial fibrillation	2,020 288,978	0.7
Type 2 diabetes	144/288,978	0.05
Inflammatory bowel disease	571/288,978	0.2
Breast cancer	158/157,895	0.1
Any of the five diseases	4,305/288,978	1.5

For each disease, progressively more extreme tails of the GPS distribution were compared with the remainder of the population in a logistic regression model with disease status as the outcome, and age, sex, the first four principal components of ancestry, and genotyping array as predictors. The breast cancer analysis was restricted to female participants.
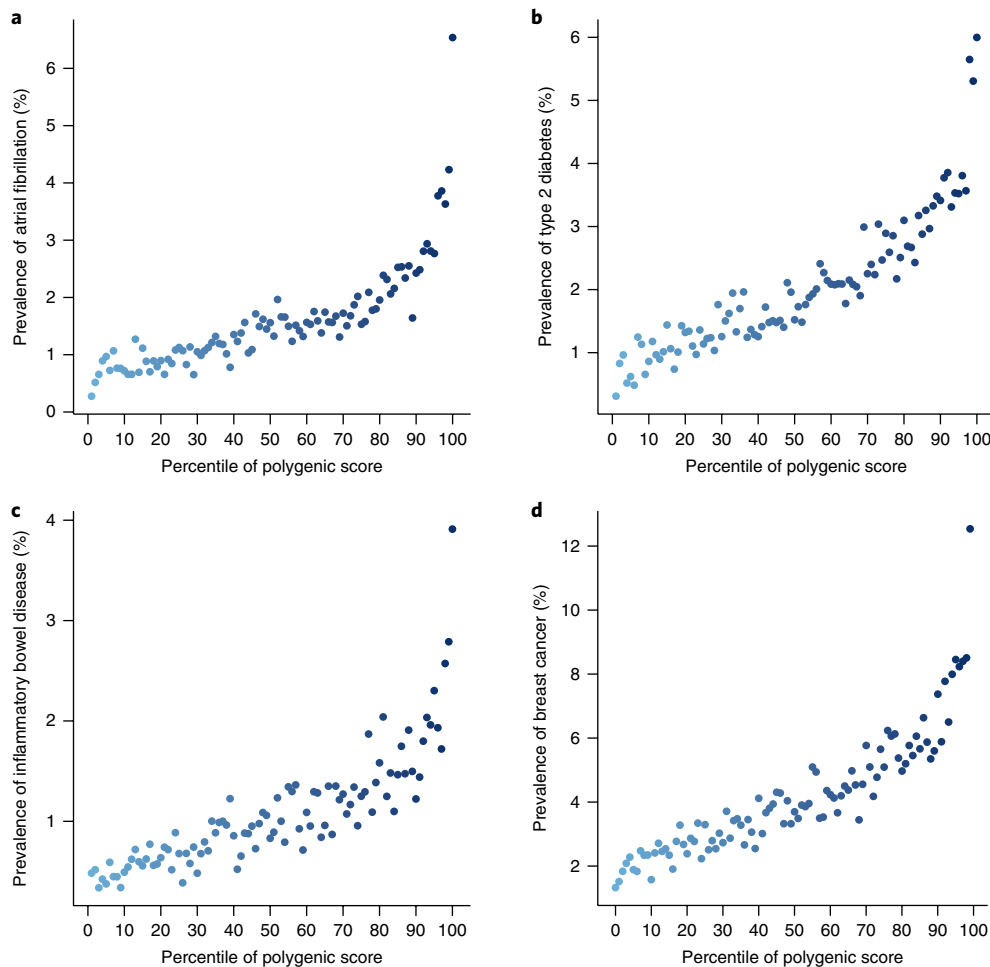
Fig. 3 | Risk gradient for disease according to the GPS percentile. 100 groups of the testing dataset were derived according to the percentile of the disease-specific GPS. **a–d**, Prevalence of disease displayed for the risk of atrial fibrillation (**a**), type 2 diabetes (**b**), inflammatory bowel disease (**c**), and breast cancer (**d**) according to the GPS percentile.

predictor identified 6.1% of the population at ≥ threefold risk and the top 1% had 4.63-fold risk (Tables 2 and 3). Screening for atrial fibrillation has become increasingly feasible owing to the development of 'wearable' device technology; these efforts to increase detection may have maximal utility in those with high GPS$_{AF}$.

Type 2 diabetes is a key driver of cardiovascular and renal disease, with rapidly increasing global prevalence[23]. The polygenic predictor identified 3.5% of the population at ≥ threefold risk and the top 1% had 3.30-fold risk (Tables 2 and 3). Both medications and an intensive lifestyle intervention have been proven to prevent progression to type 2 diabetes[24], but widespread implementation has been limited by side effects and cost, respectively. Ascertainment of those with high GPS$_{T2D}$ may provide an opportunity to target such interventions with increased precision.

Inflammatory bowel disease involves chronic intestinal inflammation and often requires lifelong anti-inflammatory medications or surgery to remove afflicted segments of the intestines[25]. The polygenic predictor identified 3.2% of the population at ≥ threefold risk and the top 1% had 3.87-fold risk (Tables 2 and 3). Although no therapies to prevent inflammatory bowel disease are currently available, ascertainment of those with increased GPS$_{IBD}$ may enable enrichment of a clinical trial population to assess a novel preventive therapy.

Breast cancer is the leading cause of malignancy-related death in women. The polygenic predictor identified 1.5% of the population at ≥ threefold risk (Tables 2 and 3). Moreover, 0.1% of women had

≥ fivefold risk of breast cancer, corresponding to a breast cancer prevalence of 19.0% in this group versus 4.2% in the remaining 99.9% of the distribution. The role of screening mammograms for asymptomatic middle-aged women has remained controversial owing to a low incidence of breast cancer in this age group and a high false positive rate. Knowledge of GPS$_{BC}$ may inform clinical decision making about the appropriate age to recommend screening[26].

These results show that, for a number of common diseases, polygenic risk scores can now identify a substantially larger fraction of the population than is found by rare monogenic mutations, at comparable or greater disease risk. Our validation and testing were performed in the UK Biobank population. Individuals who volunteered for the UK Biobank tended to be more healthy than the general population[27]; although this non-random ascertainment is likely to deflate disease prevalence, we expect the relative impact of genetic risk strata to be generalizable across study populations. Additional studies are warranted to develop polygenic risk scores for many other common diseases with large GWAS data and validate risk estimates within population biobanks and clinical health systems.

Polygenic risk scores differ in important ways from the identification of rare monogenic risk factors. Whereas identifying carriers of rare monogenic mutations requires sequencing of specific genes and careful interpretation of the functional effects of the mutations found, polygenic scores can be readily calculated for many diseases simultaneously, based on data from a single genotyping array.

Table 3 | Prevalence and clinical impact of a high GPS

High GPS definition	Reference group	Odds ratio	95% CI	P value
CAD				
Top 20% of distribution	Remaining 80%	2.55	2.43–2.67	$<1\times10^{-300}$
Top 10% of distribution	Remaining 90%	2.89	2.74–3.05	$<1\times10^{-300}$
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}
Atrial fibrillation				
Top 20% of distribution	Remaining 80%	2.43	2.29–2.59	2.1×10^{-177}
Top 10% of distribution	Remaining 90%	2.74	2.55–2.94	7.0×10^{-169}
Top 5% of distribution	Remaining 95%	3.22	2.95–3.51	1.1×10^{-152}
Top 1% of distribution	Remaining 99%	4.63	3.96–5.39	2.9×10^{-84}
Top 0.5% of distribution	Remaining 99.5%	5.23	4.24–6.39	3.5×10^{-56}
Type 2 diabetes				
Top 20% of distribution	Remaining 80%	2.33	2.20–2.46	3.1×10^{-201}
Top 10% of distribution	Remaining 90%	2.49	2.34–2.66	1.2×10^{-167}
Top 5% of distribution	Remaining 95%	2.75	2.53–2.98	1.7×10^{-130}
Top 1% of distribution	Remaining 99%	3.30	2.81–3.85	1.4×10^{-49}
Top 0.5% of distribution	Remaining 99.5%	3.48	2.79–4.29	4.3×10^{-30}
Inflammatory bowel disease				
Top 20% of distribution	Remaining 80%	2.19	2.03–2.36	7.7×10^{-95}
Top 10% of distribution	Remaining 90%	2.43	2.22–2.65	8.8×10^{-88}
Top 5% of distribution	Remaining 95%	2.66	2.38–2.96	3.0×10^{-68}
Top 1% of distribution	Remaining 99%	3.87	3.18–4.66	1.4×10^{-43}
Top 0.5% of distribution	Remaining 99.5%	4.81	3.74–6.08	9.0×10^{-37}
Breast cancer				
Top 20% of distribution	Remaining 80%	2.07	1.97–2.19	3.4×10^{-159}
Top 10% of distribution	Remaining 90%	2.32	2.18–2.48	2.3×10^{-148}
Top 5% of distribution	Remaining 95%	2.55	2.35–2.76	2.1×10^{-112}
Top 1% of distribution	Remaining 99%	3.36	2.88–3.91	1.3×10^{-54}
Top 0.5% of distribution	Remaining 99.5%	3.83	3.11–4.68	8.2×10^{-38}

Odds ratios were calculated by comparing those with high GPS with the remainder of the population in a logistic regression model adjusted for age, sex, genotyping array, and the first four principal components of ancestry. The breast cancer analysis was restricted to female participants. CI, confidence interval.

In our testing dataset, 19.8% of participants were at ≥ threefold increased risk for at least 1 of the 5 diseases studied (Table 2).

The potential to identify individuals at significantly higher genetic risk, across a wide range of common diseases and at any age, poses a number of opportunities and challenges for clinical medicine.

Where effective prevention or early detection strategies are available, key issues will include the allocation of attention and resources across individuals with different levels of genetic risk and integration of genetic risk stratification with other risk factors, including rare monogenic mutations, and clinical, and environmental factors. Where such strategies do not exist or are suboptimal, the identification of individuals at high risk should facilitate the design of efficient natural-history studies to discover early markers of disease onset and clinical trials to test prevention strategies. In both cases, it is important to recognize that the risk associated with a high polygenic score may not reflect a single underlying mechanism, but rather the combined influence of multiple pathways[28]. Nonetheless, prevention and detection strategies may have utility regardless of the underlying mechanism, as is the case for statin therapy for CAD, blood-thinning medications to prevent stroke in those with atrial fibrillation, or intensified mammography screening for breast cancer.

Risk communication will require serious consideration. While polygenic risk scores can be simultaneously calculated at birth for all common diseases, the usefulness of the knowledge and the potential harms to the individual may vary with the disease and stage of life—from juvenile diabetes to Alzheimer's disease. Yet, it may not be feasible or appropriate to withhold information that can be readily calculated from genetic data. Moreover, it will be important to consider how to assess both absolute and relative risks and how to communicate these risks to best serve each patient; for example, to encourage the adoption of lifestyle modifications or disease screening.

Finally, we highlight a crucial equity issue. The polygenic risk scores described here were derived and tested in individuals of primarily European ancestry—the group in which most genetic studies have been undertaken to date. Because allele frequencies, linkage disequilibrium patterns, and effect sizes of common polymorphisms vary with ancestry, the specific GPS here will not have optimal predictive power for other ethnic groups[29]. It will be important for the biomedical community to ensure that all ethnic groups have access to genetic risk prediction of comparable quality, which will require undertaking or expanding GWAS in non-European ethnic groups.

References

1. Green, E. D. & Guyer, M. S., National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**, 204–213 (2011).
2. Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance. *Proc. R. Soc. Edinb.* **52**, 99–433 (1918).
3. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2012).
4. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl Acad. Sci. USA* **111**, E5272–E5281 (2014).
5. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
6. Abul-Husn, N. S. et al. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science* **354**, pii: aaf7000 (2016).
7. Nordestgaard, B. G. et al. Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease: consensus statement of the European Atherosclerosis Society. *Eur. Heart J.* **34**, 3478–3490a (2013).
8. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
9. Estrada, K. et al. Association of a low-frequency variant in *HNF1A* with type 2 diabetes in a Latino population. *JAMA* **311**, 2305–2314 (2014).
10. Chatterjee, N. et al. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45**, 400–405 (2013).
11. Zhang, Y. et al. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits and implications for the future. Preprint at https://www.biorxiv.org/content/early/2017/08/11/175406 (2017).
12. Ripatti, S. et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet* **376**, 1393–1400 (2010).
13. Vilhjálmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
14. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
15. Bycroft, C. et al. Genome-wide genetic data on ~500,000 UK Biobank participants. Preprint at https://www.biorxiv.org/content/early/2017/07/20/166298 (2017).
16. Nikpay, M. et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
17. 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**, 561–567 (2016).
18. Abraham, G. et al. Genomic prediction of coronary heart disease. *Eur. Heart J.* **37**, 3267–3278 (2016).
19. Khera, A. V. et al. Genetic risk, adherence to a healthy lifestyle, and coronary disease. *N. Engl. J. Med.* **375**, 2349–2358 (2016).
20. Mega, J. L. et al. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *Lancet* **385**, 2264–2271 (2015).
21. Natarajan, P. et al. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation* **135**, 2091–2101 (2017).
22. January, C. T. et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on practice guidelines and the Heart Rhythm Society. *Circulation* **130**, e199–e267 (2014).
23. GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* **388**, 1545–1602 (2016).
24. Knowler, W. C. et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N. Engl. J. Med.* **346**, 393–403 (2002).
25. Abraham, C. & Cho, J. H. Inflammatory bowel disease. *N. Engl. J. Med.* **361**, 2066–2078 (2009).
26. Pharoah, P. D., Antoniou, A. C., Easton, D. F. & Ponder, B. A. Polygenes, risk prediction, and targeted prevention of breast cancer. *N. Engl. J. Med.* **358**, 2796–2803 (2008).
27. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
28. Khera, A. V. & Kathiresan, S. Is coronary atherosclerosis one disease or many? Setting realistic expectations for precision medicine. *Circulation* **135**, 1005–1007 (2017).
29. Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
30. Christophersen, I. E. et al. Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. *Nat. Genet.* **49**, 946–952 (2017).
31. Scott, R. A. et al. An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* **66**, 2888–2902 (2017).
32. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
33. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
34. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

Author contributions

A.V.K., M.C., and S.K. conceived and designed the study. A.V.K., M.C., K.G.A., M.E.H., C.R., S.H.C., and S.A.L. acquired, analyzed, and interpreted the data. A.V.K., M.C., E.S.L., and S.K. drafted the manuscript. A.V.K., M.C., P.N., E.S.L., P.T.E., and S.K. critically revised the manuscript for important intellectual content.

Competing interests

A.V.K. and S.K. are listed as co-inventors on a patent application for the use of genetic risk scores to determine risk and guide therapy. S.K. and P.T.E. are supported by a grant from Bayer AG to the Broad Institute focused on the genetics and therapeutics of myocardial infarction and atrial fibrillation.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0183-z.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.K.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Polygenic score derivation. Polygenic scores provide a quantitative metric of an individual's inherited risk based on the cumulative impact of many common polymorphisms. Weights are generally assigned to each genetic variant according to the strength of their association with disease risk (effect estimate). Individuals are scored based on how many risk alleles they have for each variant (for example, zero, one, or two copies) included in the polygenic score.

For our score derivation, we used summary statistics from recent GWAS studies conducted primarily among participants of European ancestry for 5 diseases[16,30–33] and a linkage disequilibrium reference panel of 503 European samples from 1000 Genomes phase 3 version 5 (ref. [34]). UK Biobank samples were not included in any of the five discovery GWAS studies. DNA polymorphisms with ambiguous strands (A/T or C/G) were removed from the score derivation. For each disease, we computed a set of candidate GPSs using the LDPred algorithm and pruning and threshold derivation strategies.

The LDPred computational algorithm was used to generate seven candidate GPSs for each disease[13]. This Bayesian approach calculates a posterior mean effect size for each variant based on a prior and subsequent shrinkage based on the extent to which this variant is correlated with similarly associated variants in the reference population. The underlying Gaussian distribution additionally considers the fraction of causal (for example, non-zero effect size) markers via a tuning parameter, ρ. Because ρ is unknown for any given disease, a range of ρ values (the fraction of causal variants) were used—1.0, 0.3, 0.1, 0.03, 0.01, 0.003, and 0.001.

A second approach—pruning and thresholding—was used to build an additional 24 candidate GPSs. Pruning and thresholding scores were built using a P value and linkage disequilibrium-driven clumping procedure in PLINK version 1.90b (--clump)[35]. In brief, the algorithm forms clumps around SNPs with association P values less than a provided threshold. Each clump contains all SNPs within 250 kilobases of the index SNP that are also in linkage disequilibrium with the index SNP as determined by a provided pairwise correlation (r^2) threshold in the linkage disequilibrium reference. The algorithm iteratively cycles through all index SNPs, beginning with the smallest P value, only allowing each SNP to appear in one clump. The final output should contain the most significantly disease-associated SNP for each linkage disequilibrium-based clump across the genome. A GPS was built containing the index SNPs of each clump with association estimate betas (log-odds) as weights. GPSs were created over a range of P value (1.0, 0.5, 0.05, 5×10^{-4}, 5×10^{-6}, and 5×10^{-8}) and r^2 (0.2, 0.4, 0.6, and 0.8) thresholds, for a total of 24 pruning and thresholding-based candidate scores for each disease. The resulting GPS for a P value threshold of 5×10^{-8} and an r^2 of < 0.2 was denoted the 'GWAS significant variant' derivation strategy.

Polygenic score calculation in the validation dataset. For each disease, the 31 candidate GPSs were calculated in a validation dataset of 120,280 participants of European ancestry derived from the UK Biobank phase 1 release. The UK Biobank is a large prospective cohort study that enrolled individuals from across the United Kingdom, aged 40–69 years at the time of recruitment, starting in 2006[14]. Individuals underwent a series of anthropometric measurements and surveys, including medical history review with a trained nurse.

Scores were generated by multiplying the genotype dosage of each risk allele for each variant by its respective weight, and then summing across all variants in the score using PLINK2 software[35]. Incorporating genotype dosages accounts for uncertainty in the genotype imputation. The vast majority of variants in the GPSs were available for scoring purposes in the validation dataset with sufficient imputation quality (INFO > 0.3) (Supplementary Tables 1–6).

For each of the five diseases, the score with the best discriminative capacity was determined based on the maximal AUC in a logistic regression model with the disease as the outcome and the disease-specific candidate GPS, age, sex, first four principal components of ancestry, and an indicator variable for genotyping array used (Supplementary Tables 1–6). AUC confidence intervals were calculated using the 'pROC' package within R.

Testing dataset. The testing dataset was comprised of 288,978 UK Biobank phase 2 genotype data release participants distinct from those in the validation dataset described above. Individuals in the UK Biobank underwent genotyping with one of two closely related custom arrays (UK BiLEVE Axiom Array or UK Biobank Axiom Array) consisting of over 800,000 genetic markers scattered across the genome[15]. Additional genotypes were imputed centrally using the Haplotype Reference Consortium resource, the UK10K panel, and the 1000 Genomes panel. To analyze individuals with a relatively homogenous ancestry and owing to small percentages of non-British individuals, the present analysis was restricted to white British ancestry individuals. This subpopulation was constructed centrally using a combination of self-reported ancestry and genetically confirmed ancestry using principal components. Additional exclusion criteria included outliers for

heterozygosity or genotype missing rates, discordant reported versus genotypic sex, putative sex chromosome aneuploidy, or withdrawal of informed consent, derived centrally as previously reported[15].

For each of the five diseases, the proportion of variance explained was calculated for each disease using the Nagelkerke's pseudo-R^2 metric (Supplementary Table 9). The R^2 was calculated for the full model inclusive of the GPS plus the covariates minus R^2 for the covariates alone, thus yielding an estimate of the explained variance. Covariates in the model included age, gender, genotyping array, and the first four principal components of ancestry.

A sensitivity analysis was performed by removing one individual from each pair of related individuals (third-degree or closer; kinship coefficient > 0.0442), confirming similar results within this subpopulation comprised of 222,529 (77%) of the 288,978 testing dataset participants (Supplementary Table 10).

Diagnosis of prevalent disease was based on a composite of data from self-report in an interview with a trained nurse, electronic health record information including inpatient International Classification of Diseases (ICD-10) diagnosis codes, and Office of Population Censuses and Surveys (OPCS-4) procedure codes.

CAD ascertainment was based on a composite of myocardial infarction or coronary revascularization. Myocardial infarction was based on self-report or hospital admission diagnosis, as performed centrally. This included individuals with ICD-9 codes of 410.X, 411.0, 412.X, or 429.79, or ICD-10 codes of I21.X, I22.X, I23.X, I24.1, or I25.2 in hospitalization records. Coronary revascularization was assessed based on an OPCS-4 coded procedure for coronary artery bypass grafting (K40.1–40.4, K41.1–41.4, or K45.1–45.5), or coronary angioplasty with or without stenting (K49.1–49.2, K49.8–49.9, K50.2, K75.1–75.4, or K75.8–75.9).

Atrial fibrillation ascertainment was based on self-report of atrial fibrillation, atrial flutter, or cardioversion in an interview with a trained nurse, an ICD-9 code of 427.3 or ICD-10 code of I48.X in hospitalization records, or a history of a percutaneous ablation or cardioversion based on the OPCS-4 coded procedure (K57.1, K62.1, K62.2, K62.3, or K 62.4), as performed previously[30].

Type 2 diabetes ascertainment was based on self-report in an interview with a trained nurse or an ICD-10 code of E11.X in hospitalization records. Inflammatory bowel disease ascertainment was based on report in an interview with a trained nurse, or an ICD-9 code of 555.X or ICD-10 code of K51.X in hospitalization records.

Breast cancer ascertainment was based on self-report in an interview with a trained nurse, ICD-9 codes (174 or 174.9) or ICD-10 codes (C50.X) in hospitalization records, or a breast cancer diagnosis reported to the national registry before the date of enrollment.

Statistical analysis within the testing dataset. For each disease, the GPS with the best discriminative capacity in the testing dataset was calculated in the testing dataset of 288,278 participants using genotyped and imputed variants and the Hail software package[36]. The proportion of the population and of diseased individuals with a given magnitude of increased risk was determined by comparing progressively more extreme tails of the distribution with the remainder of the population in a logistic regression model predicting disease status and adjusted for age, gender, four principal components of ancestry, and genotyping array. Individuals were next binned into 100 groupings according to the percentile of the GPS, and the unadjusted prevalence of disease within each bin was determined. We next compared the observed risk gradient across percentile bins with that which would be predicted by the GPS. For each individual, the predicted probability of disease was calculated using a logistic regression model with only the 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 for each of the five diseases (Supplementary Figs. 2 and 3).

Statistical analyses were conducted using R version 3.4.3 software (The R Foundation).

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability. GPSs for each of the five diseases are available for research uses at http://www.broadcvdi.org/informational/data.

References

35. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
36. Ganna, A. et al. Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nat. Neurosci.* **19**, 1563–1565 (2016).

nature research

Corresponding author(s):   Sekar Kathiresan (NG-LE48246)

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

▸ Experimental design

1. Sample size

   Describe how sample size was determined.

   > Sample based on the UK Biobank cohort study. All eligible participants with genetic and phenotypic data available included in the present analyses.

2. Data exclusions

   Describe any data exclusions.

   > Participants were excluded based on non-British ancestry, included outliers for heterozygosity or genotype missing rates, discordant reported versus genotypic sex, putative sex chromosome aneuploidy, or withdrawal of informed consent. These exclusions were prespecified prior to the analysis.

3. Replication

   Describe the measures taken to verify the reproducibility of the experimental findings.

   > Associations with disease status confirmed in a validation and testing dataset within the UK Biobank cohort study of >400,000 participants.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   > No randomization.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   > Polygenic scores extracted blinded to phenotype status of participants.

   Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided _Only common tests should be described solely by name; describe more complex techniques in the Methods section._
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	Test values indicating whether an effect is present _Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted._
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

▸ Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Statistical analyses were conducted using R version 3.4.3 software (The R Foundation).

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

▸ Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	No unique materials.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines used.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No eukaryotic cell lines used.

▸ Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	No animals used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	UK Biobank dataset included 409,258 participants of British ancestry (average age 57 years; 55% female). Disease specific phenotypes are included in Table 1.