Genomic risk prediction of coronary artery disease in nearly 500,000 adults: implications for early screening and primary prevention

Michael Inouye PhD[1,2,3,4,#,*], Gad Abraham PhD[1,2,3,4,#,*], Christopher P. Nelson PhD[5], Angela M. Wood PhD[3], Michael J. Sweeting PhD[3], Frank Dudbridge PhD[3,6], Florence Y. Lai MPhil[5], Stephen Kaptoge PhD[3,7], Marta Brozynska PhD[1,2,3], Tingting Wang PhD[1,2,3], Shu Ye MD, PhD[5], Thomas R Webb PhD[5], Martin K. Rutter MD[8,9], Ioanna Tzoulaki PhD[10,11], Riyaz S. Patel MD[12,13], Ruth J. F. Loos PhD[14], Bernard Keavney MD[15,16], Harry Hemingway MD[17], John Thompson PhD[6], Hugh Watkins MD, PhD[18,19], Panos Deloukas PhD[20], Emanuele Di Angelantonio MD, PhD[3,7], Adam S. Butterworth PhD[3,7], John Danesh FMedSci[3,7,21], Nilesh J. Samani MD[5,#,*] for The UK Biobank CardioMetabolic Consortium CHD Working Group

[1] Cambridge Baker Systems Genomics Initiative
[2] Baker Heart and Diabetes Institute, 75 Commercial Rd, Melbourne 3004, Victoria, Australia
[3] MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, United Kingdom
[4] Department of Clinical Pathology and School of BioSciences, University of Melbourne, Parkville 3010, Victoria, Australia
[5] Department of Cardiovascular Sciences and NIHR Leicester Biomedical Centre, University of Leicester, UK
[6] Department of Health Sciences, University of Leicester, Leicester, UK
[7] National Institute for Health Research Blood and Transplant Research Unit (NIHR BTRU) in Donor Health and Genomics at the University of Cambridge, Cambridge, UK
[8] Division of Diabetes, Endocrinology and Gastroenterology, School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK
[9] Manchester Diabetes Centre, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK
[10] Department of Epidemiology and Biostatistics, Imperial College London, London W2 1PG, UK
[11] Department of Hygiene and Epidemiology, University of Ioannina, 45110, Ioannina, Greece
[12] Institute of Cardiovascular Sciences, University College London, London, UK
[13] Barts Heart Centre, St Bartholomew's Hospital, London, UK
[14] Charles Bronfman Institute for Personalized Medicine, Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA
[15] Division of Cardiovascular Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK
[16] Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK
[17] The Farr Institute of Health Informatics Research and the National Institute for Health Research, Biomedical Research Centre, University College London, London, UK
[18] Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, OX3 9DU, UK
[19] The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK

[20] William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, EC1M 6BQ, UK
[21] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

* These authors contributed equally

[#] Correspondence addressed to:
Michael Inouye PhD
Baker Heart and Diabetes Institute
75 Commercial Road
Melbourne, Victoria 3004, Australia

Department of Public Health and Primary Care
University of Cambridge
Strangeways Research Laboratories
2 Worts' Causeway
Cambridge CB1 8RN, UK
Telephone: +44 (0)1223761950
Fax: +61435482767 (fax)
E-mail: mi336@medschl.cam.ac.uk, minouye@baker.edu.au

And

Gad Abraham PhD
Baker Heart and Diabetes Institute
75 Commercial Road
Melbourne, Victoria 3004, Australia
Telephone: +61 385321522
E-mail: gad.abraham@baker.edu.au

and

Nilesh J. Samani MD
Department of Cardiovascular Sciences
University of Leicester
Cardiovascular Research Centre
Glenfield General Hospital
Leicester LE3 9QP, UK
Telephone: +44 (0)1162044758
Fax: +44 116 2875792
E-mail: njs@leicester.ac.uk

Disclosures
MKR reports receiving honoraria and consulting fees from Novo Nordisk, Ascensia, Cell Catapult, and Roche Diabetes Care. Other authors have nothing to disclose.

Acknowledgements: We are grateful to UK Biobank for access to data to undertake our study. We thank Dr Joanna Howson, Dr Agus Salim, and Dr Brian Ference for their helpful input on the manuscript.

Sources of Funding: This study was supported by funding from National Health and Medical Research Council (NHMRC) grant APP1062227. Supported in part by the Victorian Government's OIS Program. M.I. was supported by an NHMRC and Australian Heart Foundation Career Development Fellowship (no. 1061435). G.A. was supported by an NHMRC Early Career Fellowship (no. 1090462). N.J.S., C.P.N. and B.K. are supported by the British Heart Foundation and N.J.S. is a NIHR Senior Investigator. R.S.P. is supported by the British Heart Foundation (FS/14/76/30933). The MRC/BHF Cardiovascular Epidemiology Unit is supported by the UK Medical Research Council [MR/L003120/1], British Heart Foundation [RG/13/13/30194], and UK National Institute for Health Research Cambridge Biomedical Research Centre. J.D. is a British Heart Foundation Professor and NIHR Senior Investigator.

Abstract

Background Coronary artery disease (CAD) has substantial heritability and a polygenic architecture. However, the potential of genomic risk scores to help predict CAD outcomes has not been evaluated comprehensively because available studies have involved limited genomic scope and limited sample sizes.

Objectives This study sought to construct a genomic risk score for CAD and to estimate its potential as a screening tool for primary prevention.

Methods Using a meta-analytic approach to combine large-scale genome-wide and targeted genetic association data, we developed a new genomic risk score for CAD (metaGRS), consisting of 1.7 million genetic variants. We externally tested metaGRS, by itself and in combination with available data on conventional risk factors, in 22,242 CAD cases and 460,387 non-cases from UK Biobank.

Results The hazard ratio (HR) for CAD was 1.71 (95% CI 1.68–1.73) per standard deviation increase in metaGRS, an association larger than any other externally tested genetic risk score previously published. The metaGRS stratified individuals into significantly different lifecourse trajectories of CAD risk, with those in the top 20% of metaGRS distribution having a HR of 4.17 (3.97–4.38) compared with those in the bottom 20%. The corresponding HR was 2.83 (2.61–3.07) among individuals on lipid-lowering or anti-hypertensive medications. The metaGRS had a higher C-index (C=0.623, 0.615–0.631) for incident CAD than any of six conventional factors (smoking, diabetes, hypertension, body mass index, self-reported high cholesterol, and family history). For men in the top 20% of metaGRS with >2 conventional factors, 10% cumulative risk of CAD was reach by age 48.

Conclusions The genomic score developed and evaluated here substantially advances the concept of using genomic information to stratify individuals with different trajectories of CAD risk and highlights the potential for genomic screening in early life to complement conventional risk prediction.

Condensed Abstract: Coronary artery disease (CAD) has substantial heritability and a polygenic architecture, thus we sought to construct a genomic risk score (GRS) for CAD and evaluate its potential as an early screening tool. In UK Biobank (N=480,000), we found that a meta-score (metaGRS) outperformed all other genetic risk scores as well as any single conventional risk factor for CAD. Furthermore, metaGRS predicted future CAD risk even in individuals on lipid-lowering or anti-hypertensive medications. As a screening tool in early life, metaGRS makes possible true primary prevention for CAD and shows promise in complementing conventional risk prediction.

Keywords: Genomic risk prediction, Coronary artery disease, Primary prevention

Abbreviations

GRS: genomic risk score
HWE: Hardy-Weinberg equilibrium
LD: linkage disequilibrium
FDR: false discovery rate
APRC: area under precision-recall curve

Introduction

As coronary artery disease (CAD) is the leading cause of morbidity and mortality worldwide, early identification of individuals at high risk of CAD is essential for primary prevention. As the heritability of CAD has been estimated to be 40–60%, comprehensive information on genetic susceptibility could contribute importantly to CAD risk stratification (1,2). Although family history has long been identified as a risk factor for CAD, elucidation of the genetic architecture of CAD has advanced substantially only during the past decade with the advent of genome-wide association studies. Results from these assumption-free surveys across the genome have laid foundations for developing genomic risk scores (GRSs) in the estimation of an individual's underlying genomic risk (3-9). Furthermore, because GRSs are based on germline DNA, they are quantifiable in early life, at or before birth. Hence, they offer the potential for early risk screening and primary prevention, before other conventional risk factors become informative.

Due to several inter-related factors, however, previous GRSs for CAD have been unable to provide comprehensive assessment of the potential of using genomic information in CAD risk prediction. First, because previously published GRSs have utilized only genetic variants of genome-wide significance (4,5,8) or involved genotyping arrays that focused only on pre-selected loci (3), they have not fully utilised genome-wide variation, preventing accurate estimation of the relative contribution of each genetic variant to CAD risk. Second, because previous studies of GRSs have tended to have moderate statistical power, they have been unable to provide precise effect size estimates (10-12). Third, because previous studies of GRSs have largely lacked external testing in large-scale cohorts that represent a diversity of ancestries (3,13)

and typically have involved only a narrow spectrum of CAD burden (e.g., inclusion of myocardial infarction only) (14,15), their generalizability has been limited.

Here, we report a more powerful and generalizable genome-wide GRS for CAD to provide a more comprehensive evaluation. We utilise a meta-analytic strategy to construct a GRS for CAD (metaGRS) that captures the totality of information from the largest previous genome-wide association studies, and then investigate the external performance of this metaGRS in stratifying CAD risk in >480,000 individuals from the UK Biobank (UKB) (16). Furthermore, we assess the effects of several conventional risk factors (smoking, blood pressure, BMI, diabetes, family history, and high cholesterol) on different genomic risk backgrounds, with the aim of delineating event rates across age, sex, clinical risk factors, and genomic risk score strata to identify individuals more likely to benefit from earlier and more intensive therapies. Finally, to assess the potential therapeutic implications of genomic risk scores, we test the impact of blood pressure and lipid lowering medication on the performance of the metaGRS.

Methods

Study design and participants

The design of this study is shown in **Online Figure 1**. Details of the design of the UKB have been reported previously (16). Participants were members of the UK general population aged between 40–69 years at recruitment, identified through primary care lists, who accepted an invitation to attend one of the 22 assessment centers that were serially established across the UK between 2006 and 2010. At recruitment, detailed information was collected via a standardized questionnaire on socio-demographic characteristics, health status and physician-diagnosed medical conditions, family history, and lifestyle factors. Selected physical and functional measurements were obtained including height, weight, waist-hip ratio, and systolic and diastolic

blood pressures. The UKB data were subsequently linked to Hospital Episode Statistics (HES) data, as well as national death and cancer registries. The HES data available for the current analysis cover all hospital admissions to NHS hospitals in England and Scotland from April 1997 to March 2015, with the Scottish data dating back as early as 1981. HES uses International Classification of Diseases ICD 9 and 10 to record diagnosis information, and OPCS-4 (Office of Population, Censuses and Surveys: Classification of Interventions and Procedures, version 4) to code operative procedures. Death registries include all deaths in the UK up to January 2016, with both primary and contributory causes of death coded in ICD-10.

CAD was defined as fatal or non-fatal myocardial infarction (MI) cases, percutaneous transluminal coronary angioplasty (PTCA), or coronary artery bypass graft (CABG). The age of event in prevalent cases was determined by self-reported age and calculated age based on the earliest hospital record for the event; if both self-reported age and calculated age were available, the smaller value was used. For incident cases, hospital and/or death records were used to determined age of event. Prevalent versus incident status was relative to the UKB enrolment assessment. In UKB self-reported data, cases were defined as having heart attack diagnosed by doctor (data field #6150) or 'non-cancer illnesses that self-reported as heart attack' (data field #20002) or self-reported operation including PTCA, CABG, or triple heart bypass (data field #20004). In HES hospital episodes data and death registry data, MI was defined as hospital admission or cause of death due to ICD9 410–412, ICD10 I21–I24, or I25.2; CABG, PTCA were defined as hospital admission OPCS-4 K40–K46, K49, K50.1, or K75.

We defined risk factors at the first assessment as follows: diabetes diagnosed by doctor (field #2443), body mass index (BMI; field #21001), current smoking (field #20116), hypertension, family history of heart disease, and high cholesterol. For hypertension we used an

expanded definition including self-reported high blood pressure (either on blood pressure medication, data fields #6177, #6153; or systolic blood pressure >140 mmHg, fields #4080, #93; or diastolic blood pressure >90 mmHg, data fields #4079, #94). For family history of heart disease, we considered history in any first degree relative (father, mother, sibling; fields #20107, 20110, and 20111, respectively). For high cholesterol, we considered individuals with self-reported high cholesterol at assessment, as well as diagnoses in the HES/death records (HES/death records (ICD9 272.0; ICD10 E78.0). For the analyses of the number of elevated risk factors, we considered diagnosed diabetes (Y/N), hypertension at assessment (Y/N), BMI >30 kg/m^2, smoking at assessment (Y/N), high cholesterol (Y/N), and family history of heart disease (Y/N).

Genotyping of UK Biobank participants was undertaken using a custom-built genome-wide array (the UK Biobank Axiom array: http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UK-Biobank-Axiom-Array-Datasheet-2014.pdf) of ~826,000 markers. Genotyping was done in two phases. 50,000 subjects were initially typed as part of the UK BiLEVE project (17). The rest of the participants were genotyped using a slightly modified array. Imputation to ~92 million markers was subsequently carried out using the Haplotype Reference Consortium (HRC) (18) and UK10K/1000Genomes haplotype resource panels, however at the time of analysis, known issues existed with the imputation using the latter panel.

Data processing and quality control

A detailed description is available in the **Online Appendix**. Briefly, we adapted appropriate quality control procedures to the set of GWAS summary statistics being utilised, filtering genetic variants for minor allele frequency, Hardy-Weinberg equilibrium, and imputation quality using PLINK (19). Population structure was controlled using the genetic

principal components supplied by UKB (17). Individuals from UKB were removed if they were diagnosed with coronary aneurysm or had no CAD event date information.

Construction of the metaGRS

A detailed description is available in the **Online Appendix**. Briefly, we built a meta-score (metaGRS) based on three genetic risk scores: (i) a previously published score (GRS46K) of 46,000 SNPs derived from a genetic association study using Metabochip, a genotyping array with a focus on cardiometabolic genetic loci (3), (ii) a score of 202 genetic variants significantly associated with CAD at FDR<0.05 (FDR202) in a recent GWAS from CARDIoGRAMplusC4D (20), and (iii) a genome-wide polygenic score (1000Genomes) based on the same GWAS (20). To derive the 1000Genomes score and weight the three genetic risk scores for the metaGRS, we used a small training set from UKB (N=3,000 individuals). The remaining 482,629 UKB individuals not in the training set comprised the external validation set.

Statistical analysis

All scores were standardized to zero-mean and unit-variance. All scores were evaluated using logistic regression or age-as-time-scale Cox proportional hazards regression, with censoring at 75y, as well as with Kaplan-Meier estimates of cumulative incidence (censored at 75y). Unless otherwise noted, analyses using only genetic risk scores include both prevalent and incident CAD cases (germline DNA variation being determined prior to any disease); to avoid reverse causation, analyses that included conventional risk factors (measured at the UKB assessment) used only incident CAD. The Cox models were stratified by sex and adjusted for genotyping array (BiLEVE vs UKB) and 10 genetic PCs. C-indices for the Cox models were sex stratified, using age as time scale. A competing risk analysis, using the Aalen-Johansen estimator (three states: CAD, non-CAD death, and censored), was conducted using the R package

'survival' (21). The precision-recall curves (equivalent to the positive-predictive-value vs sensitivity curve) were computed in the R package 'ROCR' (22), and the area under the curve (APRC) was computed using numerical integration.

Results

The characteristics of the UKB subjects in the external validation set (n=482,629) are shown in **Table 1**, comprising 22,242 CAD cases before age 75y and 460,387 non-cases in total. There were 9,729 prevalent cases of CAD at the time of recruitment and a further 12,513 incident cases of CAD during a mean follow-up of 6.2 years, at the censoring age of 75 years in 2017. Our meta-analysis approach resulted in a 'metaGRS' comprising 1,745,180 genetic variants, themselves explaining 26.8% of CAD heritability (**Online Appendix**). A comparison of the metaGRS with its individual components and previously published GRSs from Tikkanen et al (6) and Tada et al (8) in the UKB external validation set is given in **Figure 1**, showing the metaGRS had substantially greater association with CAD risk, in terms of hazard ratio as well as positive predictive value (PPV) at any given sensitivity.

In the external UKB validation set, the metaGRS was accurate at classifying CAD cases versus non-cases with an area under the ROC curve (AUC) of 0.79 (+2.8% over the reference logistic model consisting of sex, age at assessment, genotyping array, and 10 PCs). The metaGRS offered greater PPV at any given sensitivity and thus greater Area under the Precision-Recall Curve (APRC; recall is also known as sensitivity) compared to the reference model (0.161 vs 0.123; **Figure 2A**). The distributions of the metaGRS amongst prevalent CAD cases, incident CAD cases and non-CAD were each approximately Gaussian and revealed a trend of increasing genomic risk (**Online Figure 2**), with prevalent cases more easily differentiable as they likely comprise individuals at higher genomic risk who have thus had earlier CAD events.

In sex-stratified Cox regression models for all CAD (prevalent and incident), the metaGRS had an HR of 1.71 (95% CI 1.68–1.73) per s.d. of metaGRS (P<0.0001) (**Figure 1**). The metaGRS was significantly but weakly associated with body mass index (BMI) at assessment (0.0044 log(kg/m^2) per s.d., 95% CI 0.0039–0.0049, P<0.0001), diagnosed diabetes (OR=1.14 per s.d., 95% CI 1.13–1.16, P<0.0001), hypertension at assessment (OR=1.19 per s.d., 95% CI 1.18–1.20, P<0.0001), current smoking at assessment (OR=1.06 per s.d., 95% CI 1.04–1.07, P<0.0001), family history of heart disease (OR=1.21 per s.d., 95% CI 1.199–1.214, P<0.0001), and self-reported high cholesterol at/before assessment (OR=1.27 per s.d., 95% CI 1.26–1.28, P<0.0001). No evidence for competing risk effects was observed (**Online Figure 3**). In Cox regression of incident CAD (**Figure 2b**), models based on the metaGRS had higher C-index (C=0.623, 95% CI 0.615–0.630) than any of the individual conventional risk factors, with the second-best factor being self-reported high cholesterol at assessment (C=0.594, 95% CI 0.587–0.601). A model combining the six conventional risk factors had only slightly better performance (C=0.670, 95% CI 0.663–0.678) than the metaGRS individually. Combining the metaGRS with all six conventional risk factors led to a model with C-index of 0.696 (95% CI 0.688–0.703), an increase of 2.6% over the model consisting of the six conventional risk factors. When adjusting for conventional risk factors only incident CAD cases could be considered; however the HR for metaGRS was only modestly attenuated (HR=1.58 per s.d., 95% CI 1.55–1.61 not adjusting for risk factors; HR=1.55 per s.d., 95% CI 1.52–1.58 adjusting for family history; HR=1.48 per s.d., 95% CI 1.45–1.51 after adjustment for six other risk factors).

To investigate the potential role of the metaGRS in earlier life genetic screening, we compared the sex-stratified cumulative incidence of CAD across quintiles of the metaGRS (**Figure 3**). In UKB men, we observed that CAD risk in the highest metaGRS quintile began

exponentially increasing shortly after age 40, reaching a threshold of 10% cumulative risk by 61 years of age (**Figure 3**). By comparison, CAD risk for men in the lowest metaGRS quintile did not begin increasing until age 50 and on average did not reach 10% by the censoring age of 75. In UKB women, the metaGRS results were similar but delayed given the lower absolute CAD risk overall compared to men. For women in the highest metaGRS quintile, CAD risk began increasing at age 49 and reached 10% at age 75; while women in the lowest metaGRS quintile were at extremely low levels of risk, reaching 2.5% CAD risk by the censoring age of 75. There was no evidence for a statistical interaction of the metaGRS with sex. Overall, on average UKB individuals in the top metaGRS quintile were at 4.17-fold (95% CI 3.97–4.38) higher hazard of CAD than those in the bottom metaGRS quintile (**Figure 3**).

We next assessed the differences in incident CAD risk across metaGRS quintiles when combined with conventional risk factors (current smoking, diagnosed diabetes, high blood pressure, high BMI, family history of heart disease, and high cholesterol) individually (**Online Figures 4–9**) or as an unweighted score, the number (0–6) of conventional risk factors per individual (**Figure 4**). Broadly, the patterns were similar across all the analyses. Genomic risk and lifestyle/clinical factors combined to be associated with higher risk in both men and women; however, in most instances this was additive rather than interactive. In Cox regression models of incident CAD, adjusting for current smoking, diagnosed diabetes, hypertension, log BMI, family history, high cholesterol, genotyping array, and 10 genetic PCs, there was no strong evidence of statistical interactions between the metaGRS and either diabetes (P=0.074 for interaction), smoking (P=0.13 for interaction), hypertension (P=0.93 for interaction), family history (P=0.51 for interaction), or high cholesterol (P=0.14 for interaction), but there was some evidence for interaction with log BMI (HR=0.85, 95% CI 0.76–0.95, P=0.0052). From a clinical perspective,

it was notable that men in the highest metaGRS quintile who had no conventional risk factors still reached 10% cumulative incidence of CAD by age 69, with a similar cumulative incidence as men in the lowest metaGRS quintile who had 2 elevated conventional risk factors (**Figure 4**). Men in the highest metaGRS quintile and with 3 or more conventional risk factors were at extremely high levels of CAD risk, reaching the 10% threshold by age 48**.** Approximately 79% of women did not reach 10% CAD risk before age 75, even if they had 2 conventional risk factors, due to compensation by low or moderate metaGRS risk. Even amongst women in the highest metaGRS quintile, only those with 2 or more conventional risk factors achieved 10% risk before age 75 (**Figure 4**).

To assess the impact of use of treatments (lipid lowering and anti-hypertensive medication) that have been proven to lower CAD risk on the performance of the metaGRS, we analyzed the association of the metaGRS with incident CAD in those taking one or both of these classes of drugs at baseline. The hazard ratios for each s.d. in GRS were reduced but not negated by these therapies, with HRs of 1.44 (95% CI 1.40–1.48), 1.46 (95% CI 1.42–1.50) and 1.42 (95% CI 1.37–1.47) for those individuals on lipid lowering, anti-hypertensives treatments or both treatments, respectively. Accordingly, the HRs between those in the top versus bottom metaGRS quintiles were also reduced but remained substantial with HRs of 2.71 (95% CI 2.47–2.98), 2.81 (95% CI 2.56–3.09), and 2.55 (95% CI 2.28–2.86), for those individuals on lipid lowering, anti-hypertensives treatments, or both treatments, respectively (**Figure 5**).

Discussion

In an analysis of almost 500,000 people in a prospective nationwide cohort study, we evaluated a combined genomic risk score (metaGRS) built from summary statistics of the largest previous genome-wide association studies of CAD. We report a series of findings that

substantially advance the concept of using genomic information to help stratify individuals for CAD risk in general populations, an approach that leverages the fixed nature of germline DNA over the lifecourse to anticipate different lifelong trajectories of CAD risk.

First, our metaGRS achieved greater risk discrimination than have previously published genomic risk scores based on selected SNPs (3-9). For example, we found metaGRS had greater hazard ratio and positive predictive value at any given sensitivity, as well as a four-fold hazard ratio for CAD in a comparison of individuals in the top fifth versus those in bottom fifth of the risk score distribution.

Second, we found that the predictive ability of the metaGRS was largely independent of established risk factors for CAD, implying that genetic information complements (rather than replaces) conventional risk factors. As our data have suggested that higher genetic risk can at least partly be attenuated by lipid-lowering and/or anti-hypertensive therapies, it implies that individuals at high genetic risk may gain most from early initiation of these therapies and, therefore, constitute a subpopulation for which primary prevention may be particularly cost-effective (7). However, as our results have suggested that the metaGRS predicts CAD risk even among individuals taking CAD therapies at baseline, it also underscores the need to develop new therapies to address residual disease risk.

Third, we found that the metaGRS identified individuals at high risk of premature CAD as well as those unlikely ever to reach a life-long risk level requiring intervention. For example, our findings have suggested that because men in the highest metaGRS quintile are at such high risk, they are likely benefit from more intensive preventative interventions regardless of levels of traditional clinical risk factors. By contrast, the present findings suggest that about 80% of women in general populations (i.e., those not in the top 20% of the metaGRS) may not benefit

from intensive preventive interventions, in the absence of other compelling indications, before age 75 years. This finding underscores the potential value of using genomic information to optimise use of scarce resources for disease prevention; however further health economic studies would be necessary.

Although applied health studies will be needed to evaluate properly the clinical utility of CAD genomic risk scores, elements of potential clinical implementation can now be foreseen. For example, genome-wide array genotyping has a one-time cost (approximately US$50 at current prices) and can be used to calculate updated genomic risk scores for CAD as further more powerful association data emerge. Indeed, data from a genome-wide genotyping array can be utilised to calculate GRSs for a wide range of common diseases. To calculate genomic risk for individuals, simple algorithms can draw on information from such arrays, as well as from large reference groups from similar populations, such as UK Biobank. In translating genomic risk scores, standardization in assay and data processing will be necessary but achievable, including in imputation (e.g. reference panel and quality control) and handling of population stratification (e.g. using a population-specific GRS distribution and/or adjustment of GRS directly). We have made the metaGRS algorithm freely available (23) to facilitate development and translation of the concept of genomic risk as an early screening tool.

Our study has several limitations. First, while previous studies have shown the added value of a GRS to clinical risk scores, such as Framingham Risk Score and ACC/AHA13 Risk Score (23), UK Biobank does not yet have measurements of lipids and other biochemical factors available, thus relationships of the metaGRS with lipids or traditional clinical risk scores (e.g. Framingham Risk Score, QRISK, etc) could not be assessed. Second, the UK Biobank has a minimum enrolment age of 40 years and participants have been shown to be healthier than the

UK general population (24,25), thus our study may have underestimated population-level lifetime CAD risk. Third, people of non-European ancestry make up a small proportion (< 5%) of the UK Biobank, suggesting the need for studies in people of other ancestries. Similarly, future studies that externally validate the metaGRS in large multi-ethnic cohorts would maximize generalisability and minimize risk of overfitting to any single dataset or population (26). Fourth, current GWAS sample sizes and imputation efficiencies are also limiting in that they introduce noise into GRS estimates; our meta score approach here addresses this to some extent, however future large-scale cohorts will offer more powerful genomic scores. Lastly, despite the metaGRS showing substantial CAD risk discrimination in individuals already on medication, we were also unable to assess the effect of medication versus non-medication in individuals at high metaGRS risk; as without blind randomisation, this analysis would be susceptible to reverse causation, with those on medication likely already at higher CAD risk.

In conclusion, the genomic score developed and evaluated in the present study strengthens the concept of using genomic information to stratify individuals for CAD risk in general populations and demonstrates the potential for genomic screening in early life to complement conventional risk prediction.

Perspectives

Competency in Medical Knowledge: Genetics plays a major part in coronary artery disease (CAD) risk. Genomic risk of CAD is largely independent of conventional risk factors, such as lipids, blood pressure and smoking.

Translational Outlook 1: Because germline DNA does not vary with time, genomic risk can be quantified in early life, and thus play a role in primary prevention before conventional CAD risk factors become informative.

Translational Outlook 2: Genomic risk of CAD can complement conventional risk factors, providing improved risk prediction models.

References

1. Khera AV, Kathiresan S. Genetics of coronary artery disease: discovery, biology and clinical translation. Nat Rev Genet 2017;18:331-344.
2. Watkins H, Farrall M. Genetic susceptibility to coronary artery disease: from promise to progress. Nat Rev Genet 2006;7:163-73.
3. Abraham G, Havulinna AS, Bhalala OG et al. Genomic prediction of coronary heart disease. Eur Heart J 2016;37:3267-3278.
4. Khera AV, Emdin CA, Drake I et al. Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. N Engl J Med 2016;375:2349-2358.
5. Ripatti S, Tikkanen E, Orho-Melander M et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. Lancet 2010;376:1393-400.
6. Tikkanen E, Havulinna AS, Palotie A, Salomaa V, Ripatti S. Genetic risk prediction and a 2-stage risk screening strategy for coronary heart disease. Arterioscler Thromb Vasc Biol 2013;33:2261-6.
7. Mega JL, Stitziel NO, Smith JG et al. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. Lancet 2015;385:2264-71.
8. Tada H, Melander O, Louie JZ et al. Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history. Eur Heart J 2016;37:561-7.
9. Natarajan P, Young R, Stitziel NO 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 2017;135:2091-2101.

10. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet 2013;9:e1003348.

11. Chatterjee N, Shi J, Garcia-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet 2016;17:392-406.

12. Abraham G, Inouye M. Genomic risk prediction of complex human disease and its clinical application. Curr Opin Genet Dev 2015;33:10-6.

13. Khera AV, Chaffin M, Aragam K et al. Genome-wide polygenic score to identify a monogenic risk-equivalent for coronary disease. bioRxiv 2017.

14. Krarup NT, Borglykke A, Allin KH et al. A genetic risk score of 45 coronary artery disease risk variants associates with increased risk of myocardial infarction in 6041 Danish individuals. Atherosclerosis 2015;240:305-10.

15. Qi L, Ma J, Qi Q, Hartiala J, Allayee H, Campos H. Genetic risk score and risk of myocardial infarction in Hispanics. Circulation 2011;123:374-80.

16. Sudlow C, Gallacher J, Allen N 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 2015;12:e1001779.

17. Bycroft C, Freeman C, Petkova D et al. Genome-wide genetic data on~ 500,000 UK Biobank participants. bioRxiv 2017:166298.

18. McCarthy S, Das S, Kretzschmar W et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 2016;48:1279-83.

19. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4:7.

20. Nikpay M, Goel A, Won HH et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet 2015;47:1121-1130.

21. Therneau T. A Package for Survival Analysis in S. R package version 2.41-3., 2017.

22. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics 2005;21:3940-1.

23. Abraham G, Inouye M. Coronary Artery Disease (CAD) MetaGRS. v2 ed: figshare.com, 2018.

24. Fry A, Littlejohns TJ, Sudlow C et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. American Journal of Epidemiology 2017;186:1026-1034.

25. Ganna A, Ingelsson E. 5 year mortality predictors in 498,103 UK Biobank participants: a prospective population-based study. Lancet 2015;386:533-40.

26. Alpaydin E. Introduction to Machine Learning. 1st edition. Cambridge, MA: MIT Press, 2009:76-79.

Figure Legends

Central Illustration: Genomic risk score for coronary artery disease. The genomic score provides potential for risk screening early in life as well as complements conventional risk factors for coronary artery disease.

Figure 1: Relative performance of individual genomic risk scores for CAD compared to the metaGRS. In the UKB validation set (n=482,629), (a) hazard ratios per s.d. of each score for all CAD (n=22,242), censored at 75y, from Cox regression stratified by sex and adjusted for genotyping array (BiLEVE/UKB) and 10 genetic PCs; (b) Positive predictive value vs sensitivity for a logistic regression for each GRS, adjusted for sex, age, genotyping array (BiLEVE/UKB) and 10 genetic PCs.

Figure 2: Predictive measures of CAD using the metaGRS and conventional risk factors. (a) Positive predictive values vs sensitivity for the reference model (sex + age + array + 10 genetic PCs) and when adding the metaGRS to the model for all CAD in the UKB testing set. APRC is Area under the Precision-Recall Curve. (b) C-index for sex-stratified age-as-time-scale Cox regression of incident CAD for conventional risk factors individually and in combination with the metaGRS, including genotyping array and 10 genetic PCs as covariates.

Figure 3: Cumulative risk of CAD by quintiles of metaGRS in men and women. Dotted lines represent 95% confidence intervals. For subgroup sample sizes, see **Supplementary Table 1**.

Figure 4: Cumulative risk of incident CAD for increasing numbers of conventional risk factors stratified by metaGRS quintile. Dotted lines represent 95% confidence intervals.

Figure 5: Cumulative risk of incident CAD within individuals on lipid-lowering or BP-lowering medication at assessment. Dotted lines represent 95% confidence intervals.

Table 1: Study characteristics

	UK Biobank (n=482,629)	**Male** n=220,284 (45.6%)	**Female** n=262,345 (54.4%)
Age at assessment, years [mean (sd)]	56.5 (8.1)	56.7 (8.2)	56.4 (8.0)
Current smoker (%)	50,664 (10.5%)	27,391 (12.4%)	23,273 (8.9%)
Blood pressure, systolic, mm Hg [mean (sd)]	139.8 (19.7)	142.8 (18.5)	137.3 (20.3)
Diabetes diagnosed by doctor (%)	24,920 (5.2%)	15,336 (7.0%)	9,887 (4.5%)
Hypertension (%)	254,564 (52.7%)	133,013 (60.4%)	121,533 (46.3)
Family history, 1st degree relative (%)	206,363 (42.8%)	87,946 (39.9%)	118,417 (45.1%)
High cholesterol	65,829 (13.6%)	37,801 (17.2%)	28,028 (10.7%)
Prevalent CAD events before age 75y (%)	9,729 (2.0%)	7950 (3.6%)	1779 (0.7%)
Incident CAD events before age 75y (%)	12,513 (2.6%)	9320 (4.2%)	3193 (1.2%)
On blood-pressure lowering medication	99,454 (20.6%)	53,535 (24.3%)	45,939 (17.5%)
On lipid-lowering medication	82,493 (17.1%)	49,459 (22.5%)	33,028 (12.6%)
Follow-up time, years [mean (sd)]	6.2 (2.1)	5.9 (2.6)	6.4 (1.4)