Defining a Contemporary Ischemic Heart Disease Genetic Risk Profile Using Historical Data

Jonathan D. Mosley, MD, PhD; Sara L. van Driest, MD, PhD; Quinn S. Wells, MD, PhamD, MSc; Christian M. Shaffer, BS; Todd L. Edwards, PhD; Lisa Bastarache, MS; Catherine A. McCarty, PhD, MPH; Will Thompson, PhD; Christopher G. Chute, MD, DrPH; Gail P. Jarvik, MD, PhD; David R. Crosslin, PhD; Eric B. Larson, MD, MPH; Iftikhar J. Kullo, MD; Jennifer A. Pacheco, BA; Peggy L. Peissig, PhD, MBA; Murray H. Brilliant, PhD; James G. Linneman, BA; Josh C. Denny, MD, MS; Dan M. Roden, MD

Background—Continued reductions in morbidity and mortality attributable to ischemic heart disease (IHD) require an understanding of the changing epidemiology of this disease. We hypothesized that we could use genetic correlations, which quantify the shared genetic architectures of phenotype pairs and extant risk factors from a historical prospective study to define the risk profile of a contemporary IHD phenotype.

Methods and Results—We used 37 phenotypes measured in the ARIC study (Atherosclerosis Risk in Communities; n=7716, European ancestry subjects) and clinical diagnoses from an electronic health record (EHR) data set (n=19 093). All subjects had genome-wide single-nucleotide polymorphism genotyping. We measured pairwise genetic correlations (rG) between the ARIC and EHR phenotypes using linear mixed models. The genetic correlation estimates between the ARIC risk factors and the EHR IHD were modestly linearly correlated with hazards ratio estimates for incident IHD in ARIC (Pearson correlation [r]=0.62), indicating that the 2 IHD phenotypes had differing risk profiles. For comparison, this correlation was 0.80 when comparing EHR and ARIC type 2 diabetes mellitus phenotypes. The EHR IHD phenotype was most strongly correlated with ARIC metabolic phenotypes, including total:high-density lipoprotein cholesterol ratio (rG=−0.44, P=0.005), high-density lipoprotein (rG=−0.48, P=0.005), systolic blood pressure (rG=0.44, P=0.02), and triglycerides (rG=0.38, P=0.02). EHR phenotypes related to type 2 diabetes mellitus, atherosclerotic, and hypertensive diseases were also genetically correlated with these ARIC risk factors.

Conclusions—The EHR IHD risk profile differed from ARIC and indicates that treatment and prevention efforts in this population should target hypertensive and metabolic disease. **(*Circ Cardiovasc Genet*. 2016;9:521-530. DOI: 10.1161/CIRCGENETICS.116.001530.)**

Key Words: atherosclerosis ■ blood pressure ■ coronary artery disease ■ risk factors ■ triglycerides

There has been a marked decline in mortality and morbidity for ischemic heart disease (IHD) over the past several decades.[1] These gains have come from targeting treatment and prevention strategies toward risk factors identified in landmark longitudinal studies, such as the Framingham Heart Study.[2] However, since the inception of these cohort studies, there have been changes in the prevalence of IHD risk factors such as smoking, type 2 diabetes mellitus (T2D), and obesity, which would be expected to alter the epidemiological risk profile of this disease.[3] To realize continued declines in morbidity, ongoing treatment and prevention efforts must be directed toward contemporary risk profiles.[4] Although initiating new longitudinal studies is 1 approach, it is hampered by long latencies and high costs.[5] Data sources, such as electronic health records (EHRs), offer a contemporary cohort of subjects with prevalent and incident IHD. However, baseline

Clinical Perspective on p 530

Received March 18, 2016; accepted September 28, 2016.

From the Department of Medicine (J.D.M., S.L.v.D., Q.S.W., C.M.S., J.C.D., D.M.R.), Department of Pediatrics (S.L.v.D.), Vanderbilt Epidemiology Center (T.L.E.), Biomedical Informatics (L.B., J.C.D., D.M.R.), and Department of Pharmacology (D.M.R.), Vanderbilt University, Nashville, TN; Essentia Institute of Rural Health, Duluth, MN (C.A.M.); Center for Biomedical Research Informatics, North Shore University Health System, Evanston, IL (W.T.); Department of Health Policy and Management Schools of Medicine, Public Health, and Nursing, Johns Hopkins University, Baltimore, MD (C.G.C.); Departments of Medicine (Medical Genetics) and Genome Sciences (G.P.J.) and Departments of Biomedical Informatics and Medical Education (D.R.C.), University of Washington, Seattle; Group Health Research Institute, Seattle, WA (E.B.L.); Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN (I.J.K.); Center for Genetic Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL (J.A.P.); and Biomedical Informatics Research Center (P.L.P., J.G.L.) and Center for Human Genetics (M.H.B.), Marshfield Clinic Research Foundation, WI.

The Data Supplement is available at http://circgenetics.ahajournals.org/lookup/suppl/doi:10.1161/CIRCGENETICS.116.001530/-/DC1.

Correspondence to Jonathan D. Mosley, MD, PhD, Vanderbilt University School of Medicine, 1285 Medical Research Bldg IV, Nashville, TN 37232. E-mail jonathan.d.mosley@vanderbilt.edu

risk factors are often unavailable or inconsistently measured across subjects in EHR data sets, thereby limiting their use for epidemiological analyses.[6] We propose here an alternative study design that overcomes this limitation using statistical methods that determine the relationship between 2 phenotypes based on their shared genetic risk.

Genetic variation is an important modulator of many known IHD risk factors and biomarkers.[7] Because this genetic variation constitutes a lifelong exposure of disease risk, modeling disease associations based on underlying genetic variation can link risk factors to clinical outcomes, as demonstrated by Mendelian randomization or genetic risk score analyses.[8–11] An alternative genetic association approach used here is based on genetic correlations measured using generalized linear mixed models (GLMMs). GLMMs use common single-nucleotide polymorphisms (SNPs) to quantify the phenotypic variation attributable to additive genetics in unrelated individuals.[12,13] This method is typically more sensitive to capturing the overall additive genetic effects of common SNPs than genome-wide association studies and related approaches, such as genetic risk score analyses.[14] GLMMs can also analyze pairs of phenotypes to measure a genetic correlation, which quantifies the extent to which 2 phenotypes share genetic influences.[15,16] Importantly, a genetic correlation can be calculated using a risk factor measured in one population and an outcome measured in a second, unrelated population.

We hypothesized that we could use genetic correlations to define the epidemiology of IHD in an EHR population using risk factors measured in an unrelated population. We used baseline phenotypes measured in the prospective ARIC study (Atherosclerosis Risk in Communities)[17] and diseases ascertained through the Electronic Medical Records and Genomics (eMERGE) network, a consortium of medical centers with EHR-linked DNA biobanks.[18] We demonstrate that genetic correlation and longitudinal analyses identify similar risk factor associations for T2D. However, this was not the case for IHD, indicating that the IHD genetic risk profile in the EHR cohort differs from the ARIC cohort.

Materials and Methods

An overview of the approaches used is shown in Figure I in the Data Supplement.

Study Populations

ARIC: The ARIC population comprises 13 113 genotyped adult subjects participating in the National Heart, Lung and Blood Institute-funded Atherosclerosis Risk in Communities longitudinal study designed to investigate the natural history of cardiovascular and atherosclerotic diseases.[17] Study subjects were recruited in 1987 to 1989 from 4 US communities: Minneapolis, MN; Washington County, MD; Forsyth County, NC; and Jackson, MS. Genetic and phenotypic data were downloaded from dbGaP (phs000280.v3.p1). EHR subjects: The EHR populations included adult subjects from the eMERGE Phase I Network (n=14 205), a consortium of medical centers using EHRs as a tool for genomic research and the Vanderbilt Electronic Systems for Pharmacogenomic Assessment cohort (n=11 639).[19,20] The eMERGE subjects came from 5 sites (Vanderbilt University [VUMC], Marshfield Clinic, Northwestern University, Mayo Clinic, and Group Health Research Institute), although the nonoverlapping Vanderbilt Electronic Systems for Pharmacogenomic Assessment cohort included additional subjects from VUMC BioVU resource, a deidentified collection of patients whose DNA was extracted from discarded blood and linked to phenotypes through a deidentified EHR.[21] Both the ARIC and EHR data sets were primarily composed of self-reported whites, and so only subjects of European ancestry were included in the final analyses, defined using STRUCTURE[22] in conjunction with ancestry informative markers, with European ancestry defined as >90% (ARIC) or >80% (EHR subjects) probability of being in the HapMap Northern Europeans from Utah cluster. Thresholds were selected based on comparisons of self-reported race with STRUCTURE ancestry assignment using a multiracial population.

The eMERGE study was approved by the Institutional Review Board at each site.[18,21] During the period of study, Vanderbilt BioVU resource operated as nonhuman subjects research according to the provisions of 45 Code of Federal Regulations, part 46, with oversight by Vanderbilt Institutional Review Board, as previously described.[21] This study was approved by Vanderbilt Institutional Review Board.

Genetic Data

SNP genotype data were acquired on the Illumina Human660W-Quadv1_A (eMERGE), Illumina HumanOmni1-Quad (Vanderbilt Electronic Systems for Pharmacogenomic Assessment), HumanOmni5-Quad (Vanderbilt Electronic Systems for Pharmacogenomic Assessment), and Affymetrix 6.0 SNP array (ARIC) platforms. Quality control steps for the EHR data sets were performed according to the published protocols established by the eMERGE Genomics Working Group.[23] For imputation, palindromic alleles were aligned to the positive strand using allele frequency information from the 1000 Genomes Project. SNPs were prephased using SHAPEIT,[24] and data were imputed using IMPUTE2[25] and the October 2014 release of the 1000 Genomes cosmopolitan reference haplotypes. Quality control for the ARIC data set followed the guidelines accompanying the dbGaP release, including removing SNPs with chromosomal anomalies and with >5 discordant calls in replicate samples, and using a predefined subset of unrelated subjects. Quality control analyses used PLINK v1.07.[26] After filtering for a sample missingness rate <2.0%, a SNP missingness rate <2.0%, and a SNP deviation from Hardy–Weinberg <0.001, there were 627 580 SNPs with minor allele frequency >1.0%. The merged intersection of the ARIC and imputed EHR data sets contained 488 525 SNPs with minor allele frequency >1.0%.

Phenotype Data

EHR clinical phenotypes were based on PheCodes, which are collections of related International Classification of Disease-9 diagnosis codes.[27,28] For each phenotype, cases are subjects with 2 or more instances of the code appearing in their medical record on 2 separate dates.[28] Controls are randomly selected for each phenotype among individuals without any closely related PheCodes. There are ≈1600 defined phenotypes, of which 519 had ≥400 cases in the EHR data set. PheCodes used in the primary analyses were T2D (code 250.2), IHD (411), atherosclerosis of the extremities (440.2), and myocardial infarction (411.2). PheCodes and International Classification of Disease-9 mappings are available at https://phewas.mc.vanderbilt.edu/.

The ARIC phenotypes included 37 baseline (visit 1) measurements. A list of these phenotypes, selection criteria, and data transformations are in Table I in the Data Supplement. Longitudinal outcomes and latencies for incident T2D and coronary heart disease (CHD) used phenotypes generated for the ARIC–GENEVA substudy. For the T2D longitudinal analysis, the variables used were incident cases (phv00080468.v1.p1), follow-up time (phv00080469.v1.p1), and prevalent diabetes mellitus at baseline (phv00022845.v1.p1). For CHD, the analyses used incident cases (phv00066400.v1.p1) of myocardial infarction, fatal coronary event, silent infarction, or revascularization procedure by December 31, 2004, follow-up time (phv00022861.v1.p1), and prevalent disease a baseline (phv00022832.v1.p1).

Statistics

GLMMs were used to compute genetic liabilities and genetic correlations, as implemented in the Genome-wide Complex Trait Analysis

program v1.24.7.[29] The GLMM method is described in more detail in Methods in the Data Supplement. First, a genetic relationship matrix was computed using autosomal SNPs with a minor allele frequency >1.0%. One of a pair of subjects with a relatedness score >0.05 was randomly excluded. After exclusions, there were 7716 ARIC and 19093 EHR unrelated subjects. Analyses adjusted for age (ARIC subjects) or birth year (EHR subjects), sex, and 20 principal components. *P* values for genetic liability estimates were computed by a likelihood ratio test comparing a full model with a model excluding the genetic relationship matrix variance component. PheCodes with a positive liability estimate with $P<0.1$ and >400 cases were used for genetic correlation analyses (n=158).

A bivariate GLMM, adjusting for age, sex, and 20 principal components, was used to compute genetic correlations between the ARIC risk factors and EHR phenotypes. Bivariate GLMMs were constrained to have a genetic correlation between −1 and 1, and estimates from constrained models are noted. These models were not adjusted for comorbidities because adjusting for genetic phenotypes can lead to false-positive associations because of collider bias.[30] *P* values are based on a likelihood ratio test comparing a full model (L1) and a model where the genetic correlation was constrained to be 0 (L0; likelihood ratio test= $-2[L1-L0] \approx \chi^2$ [1 *df*]). Because of the highly correlated nature of the phenotypes, a Benjamini–Hochberg false discovery rate adjustment was applied within each experiment to adjust for multiple testing.[31] Although all association data are shown within each figure and table, only phenotype pairs with false discovery rate q value <0.1 were considered to have a statistically significant genetic correlation.

Cox proportional hazards analysis was used to measure hazard ratios between ARIC risk factor phenotypes and incident T2D or CHD. For each analysis, subjects with prevalent T2D or CHD at baseline were excluded, respectively, and only self-reported whites were analyzed. Each risk factor was standardized to have a SD of 1, so hazard ratios represent the risk per SD increase. The models were adjusted for sex and age. The proportional hazards assumption was evaluated by examining cumulative Martingale residuals and applying Kolmogorov-type supremum tests. Analyses were performed using SAS v9.3 (SAS Institute, Cary, NC).

Results

Study Populations and Phenotypes

We used genetic correlations to associate ARIC phenotypes with clinical diagnoses measured in an EHR data set (Figure 1; Figure I in the Data Supplement). The EHR data set comprised 19093 unrelated European ancestry subjects, of which 50.1% were males and the median birth year was 1945 (interquartile range 1935–1955; Table II in the Data Supplement). There were 158 genetic phenotypes available for analyses in this population (Table III in the Data Supplement). The ARIC data set comprised 7716 unrelated European ancestry subjects. We selected 37 genetically modulated phenotypes from this population that have been associated with IHD and represent a range of anthropometric, laboratory, and physiological biomarkers (Table IV in the Data Supplement).

Validating Genetic Correlation Associations

To assess the specificity and magnitudes of genetic correlations, we computed genetic correlations between 5 continuous ARIC phenotypes and EHR phenotypes whose clinical case definitions are closely related to the continuous phenotypes (eg, diabetes mellitus is defined by elevated blood glucose levels). For each ARIC phenotype, the largest and most significant genetic correlation was seen with its corresponding EHR phenotype (Figure 2A; Table V in the Data Supplement). For instance, glucose levels were significantly positively genetically correlated with a clinical diagnosis of T2D (genetic correlation [rG]=0.79, $P=1.4\times10^{-7}$), indicating that higher genetically determined glucose values are associated with an increased genetic risk of T2D. The genetic correlations for the ARIC phenotypes and the corresponding EHR disease-defining phenotypes ranged from 0.70 (body mass index and obesity) to 1.0 (pack-years of smoking and tobacco use disorder).

We compared cross-sectional phenotype associations based on genetic correlations to longitudinal associations based on hazard ratios for T2D and 37 ARIC phenotypes. The EHR T2D phenotype was significantly genetically correlated with an ARIC phenotype of incident and prevalent T2D (rG=0.76, $P=2.3\times10^{-6}$; 1321 cases and 6395 controls), indicating that the phenotypes have a similar genetic risk profile (Figure 2B). Fourteen ARIC phenotypes were genetically correlated with EHR T2D at false discovery rate q<0.1 (Figure 2C; Table VI in the Data Supplement). Phenotypes with positive correlations included measures of adiposity (eg, waist circumference rG=0.51, $P=8\times10^{-5}$), low-density lipoprotein (LDL) cholesterol (rG=0.79, $P=2.0\times10^{-4}$), SBP (rG=0.35, $P=0.017$), Protein C (rG=0.30, $P=0.001$), and the inflammatory marker C-reactive protein (rG=0.39, $P=0.027$). The negatively correlated phenotypes were activated partial thromboplastin time (rG=−0.44, $P=4.2\times10^{-4}$) and high-density lipoprotein (HDL) cholesterol levels (rG=−0.38, $P=0.006$). We compared the genetic correlation estimates to hazard ratio estimates for incident T2D within the ARIC data set (Table VI in the Data Supplement). There was a strong linear relationship between

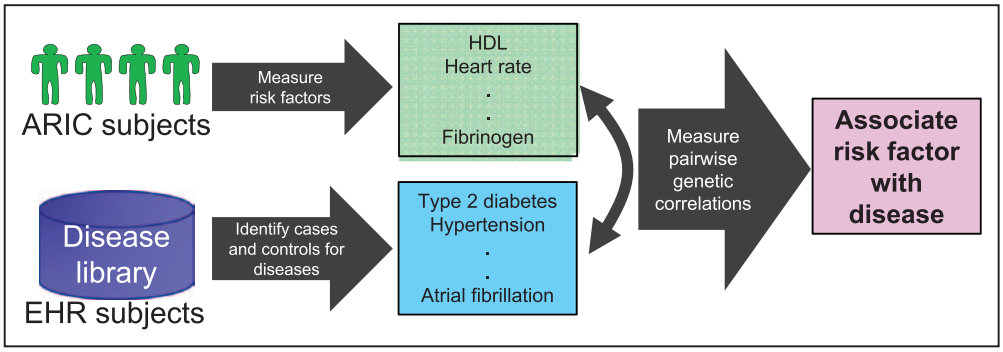

Figure 1. Overview of the study approach. Risk factors and biomarkers were measured from the ARIC study (Atherosclerosis Risk in Communities) population, and clinical diseases were curated from an electronic health record (EHR) data set. Pairwise genetic correlations between the diseases and risk factor were measured to identify genetically correlated disease–risk factor pairs. A more detailed overview is presented in Figure I in the Data Supplement. HDL indicates high-density lipoprotein.

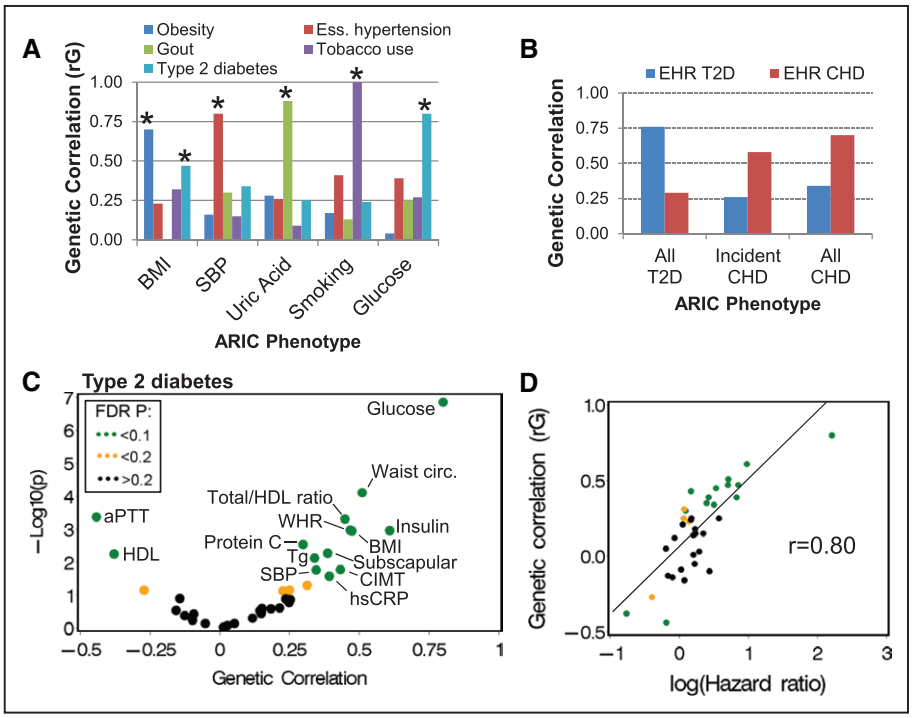

Figure 2. Validating genetic correlations. **A**, Cases and controls for 5 diseases were extracted from the electronic health record (EHR) data set, and its continuous analog was measured in the ARIC study (Atherosclerosis Risk in Communities) population. Genetic correlations for each disease–phenotype pair were computed using a generalized linear mixed model adjusting for age, sex, and 20 PCs. Each bar shows the genetic correlation for a pair. The genetic correlation for the smoking–tobacco user pair was constrained to a value of 1. An asterisk denotes genetic correlations with a nominal $P<0.001$. **B**, Genetic correlations between electronic health record (EHR) and ARIC phenotypes for type 2 diabetes mellitus (T2D) and ischemic heart disease (IHD). All includes both incident and prevalent cases. **C**, Genetic correlations between the EHR T2D phenotype (n=4367 cases; genetic liability=0.15 [SE 0.02]) and 37 ARIC phenotypes. Each point is a pairwise genetic correlation. Color-coding indicates false discovery rate (FDR) significance levels for genetic correlations, as indicated in the key. **D**, Scatter plot comparing the genetic correlations with the EHR T2D phenotype and hazards ratios for incident T2D (n=878 incident cases) computed in the ARIC data set. Hazard ratios were adjusted for age and sex. aPTT indicates activated partial thromboplastin time; BMI, body mass index; CHD, coronary heart disease; CIMT, carotid intimal-medial thickness; CRP, C-reactive protein; HDL, high-density lipoprotein; and SBP, systolic blood pressure.

hazards ratios and the genetic correlations (Pearson r=0.80; Figure 2D).

Risk Factors Associated With IHD

We next examined an EHR IHD phenotype, which comprises myocardial infarction, coronary atherosclerosis, and cardiac angina. IHD cases (n=5114) included a larger proportion of males and had higher rates of comorbid diagnoses related to heart failure, hypertension, kidney disease, T2D, and obesity, when compared with controls (n=8789; Table 1). The genetic correlation between the EHR IHD phenotype and CHD in ARIC (defined as myocardial infarction, fatal coronary artery disease, silent myocardial infarction detected by ECG, or coronary revascularization) was 0.58 (P=0.01) using incident ARIC cases (911 cases and 6272 controls) and 0.70 (P=0.0005) using incident and prevalent cases (n=1314; Figure 2B). The EHR IHD phenotype was positively genetically correlated with carotid intimal-medial thickness (CIMT; rG=0.74, P=0.002), 2 cardiac ECG phenotypes (corrected QT interval [rG=0.59, P=0.003] and Cornell voltage which is used to identify left ventricular hypertrophy[32] [rG=0.52, P=0.007]), total:HDL cholesterol ratio and triglycerides (Figure 3A; Table 2). HDL (rG=–0.48, P=0.005) and apolipoprotein A (rG=–0.45, P=0.016) were negatively correlated. We compared the genetic correlation estimates to hazard ratio estimates for incident CHD in ARIC (Table 2). The estimates were correlated (r=0.62), but more weakly than for T2D (Figure 3B). A similar result was observed when comparing odds ratios based on incident and prevalent disease in ARIC (r=0.66; Figure II in the Data Supplement). Several risk factors were associated with the ARIC IHD phenotype but not the EHR phenotype, including C-reactive protein, fibrinogen, and lipid measures related to LDL (Figure 3B). LDL cholesterol had a highly significant association with the ARIC phenotype (hazard ratio [HR], 1.4; 95% confidence interval [CI], 1.3–1.5; $P=1\times10^{-27}$) but a nonsignificant genetic correlation (rG=0.07, P=0.7) with the EHR phenotype (Table 2). The genetic correlation between LDL cholesterol and incident CHD, both measured in ARIC, was larger than with the EHR phenotype but was not significant (rG=0.33, P=0.16).

We examined 2 additional EHR phenotypes associated with atherosclerotic disease to see whether they had LDL associations. Peripheral vascular disease was positively genetically correlated with LDL (rG=0.79, $P=2\times10^{-4}$; Figure 3C; Table VII in the Data Supplement). There were also significant positive correlations with smoking, SBP, and glucose levels and negative correlations with HDL and apolipoprotein A. Myocardial infarction had a different pattern of associations and was most strongly positively correlated with triglycerides, total:HDL cholesterol ratio, coagulation Factor VII levels,

Table 1. Characteristics of the Electronic Health Record Ischemic Heart Disease Cases and Controls

Characteristic	Cases, n=5114	Controls, n=8789	*P* Value*
Sex, n (%)			
Males	3206 (62.7)	4126 (47.0)	<0.0001
Females	1908 (37.3)	4663 (53.0)	
Birth decade			
Median (IQR)	1935 (1925–1945)	1945 (1935–1955)	<0.0001
Comorbidity, n (%)†			
Congestive heart failure	2013 (39.4)	427 (4.9)	<0.0001
Atrial fibrillation	1738 (34.0)	596 (6.8)	<0.0001
Hyperlipidemia	3881 (75.9)	3718 (42.3)	<0.0001
Essential hypertension	4276 (83.6)	4402 (50.1)	<0.0001
Chronic renal failure	1292 (25.3)	800 (9.1)	<0.0001
Type 2 diabetes mellitus	2057 (40.2)	1739 (19.8)	<0.0001
Chronic airway obstruction	1222 (23.9)	689 (7.8)	<0.0001
Tobacco use disorder	965 (18.9)	885 (10.1)	<0.0001
Gout	591 (11.6)	294 (3.3)	<0.0001
Sleep apnea	805 (15.7)	667 (7.6)	<0.0001
Obesity	1168 (22.8)	1490 (17.0)	<0.0001

IQR indicates interquartile range.
*P values were based on χ^2 analysis (sex) or logistic regression analyses, adjusting for sex and birth year (comorbidities).
†Comorbidities were defined based on International Classification of Disease-9 derived PheCodes.

ECG measures, and CIMT (Figure 3D; Table VIII in the Data Supplement).

Diagnoses Associated With Risk Factors

The genetic correlation approach can test associations between a risk factor and a collection of EHR diagnosis to identify EHR diagnoses associated with the risk factor. We used this approach to further define 4 ARIC phenotypes positively correlated with the EHR IHD phenotype (Figure 3C). For each ARIC phenotype, we measured pairwise genetic correlations between 158 EHR phenotypes. Non-HDL cholesterols directly contribute to atherosclerotic disease, although CIMT is a measure of accumulated disease burden within the carotid artery.[33] Total:HDL cholesterol ratio was significantly associated with a 19 EHR phenotypes, including atherosclerotic diseases, such as peripheral vascular disease and coronary atherosclerosis (rG=0.42, P=0.008), and T2D (rG=0.45, P=4.9×10^{-4}; Figure 4A; Table IX in the Data Supplement). The number and strength of the significant genetic correlations for the total:HDL ratio phenotype was greater than either LDL (n=0) or HDL (n=2) cholesterol phenotypes (Figure III in the Data Supplement). CIMT was significantly correlated with 4 diagnoses related to hypertension and IHD and showed weaker associations (false discovery rate q>0.1) with atherosclerotic diseases (Figure 4B; Table X in the Data Supplement). We next examined systolic blood pressure (SBP) and Cornell voltage. SBP was significantly positively genetically correlated with 4 diagnoses related to hypertension and hypertensive heart and kidney disease (rG=0.75, P=0.0004; Figure 4C; Table XI in the Data Supplement). There were no significant genetic correlations with Cornell voltage (Figure 4D; Table XII in the Data Supplement). The strongest nonsignificant correlations were seen with IHD and hypertension-related diagnoses. In sum, the pattern of diseases associated with these risk factors point to a contribution of hypertensive, metabolic, and atherosclerotic disease processes to the EHR IHD risk.

Discussion

We used genetic correlations to associate baseline risk factors in the ARIC prospective study with clinical diagnoses from an EHR data set. We show that, for T2D, the genetic correlations from cross-sectional analyses were linearly related to hazard ratio estimates from a longitudinal analysis. The correlation was weaker when comparing these values for an IHD phenotype. The EHR IHD phenotype was genetically correlated with CIMT, total/HDL cholesterol ratio, HDL, SBP, and triglycerides. An analysis of additional EHR phenotypes genetically correlated with the IHD risk factors indicates that a genetic predisposition toward hypertension, atherosclerosis, and metabolic syndrome is an important contributor to IHD risk in the EHR population.

The marked reductions in morbidity and mortality for stroke and myocardial infarction are largely attributable to effective treatment and prevention strategies targeting epidemiologically significant risk factors.[34] Over the past decades, there have also been changes in societal behaviors and norms that have altered the prevalence of IHD risk factors and, potentially, their contribution to the risk of this disease. We hypothesized that we could directly evaluate whether the epidemiological profile for IHD has changed by comparing associations between both a

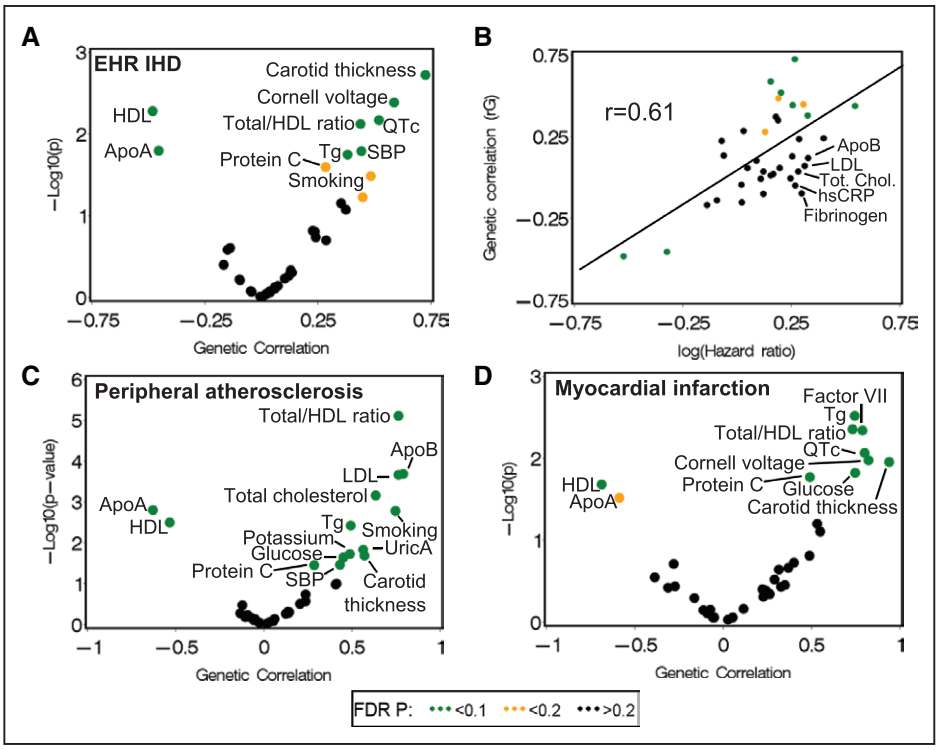

Figure 3. ARIC study (Atherosclerosis Risk in Communities) phenotypes genetically correlated with ischemic heart disease (IHD). Each point is a pairwise genetic correlation. Color-coding indicates FDR significance levels for genetic correlations, as indicated in the key at the bottom of the figure. **A**, Genetic correlations with the electronic health record (EHR) IHD phenotype (n=5114 cases, liability=0.09[0.02]). **B**, Scatter plot comparing the IHD genetic correlations to hazard ratios for incident coronary heart disease (n=1314 cases) in the ARIC data set. Risk factors strongly associated with the ARIC phenotype, but not with the EHR phenotype, are labeled. Genetic correlations for (**C**) peripheral atherosclerosis (n=1604 cases, genetic liability=0.15[0.04]) and (**D**) myocardial infarction (n=1700 cases, liability=0.06[0.04]). Apo apolipoprotein; CRP, C-reactive protein; FDR, false discovery rate; HDL, high-density lipoprotein; LDL, low-density lipoprotein; SBT, systolic blood pressure; and Tg, triglycerides.

contemporary or historical IHD phenotype and a common set of risk factors. Testing this hypothesis was feasible because associations can now be modeled based on underlying genetic risk. Furthermore, by testing associations based on underling genetics, we demonstrate that it is possible to rapidly delineate a contemporary risk profile for an EHR IHD phenotype. This contemporary risk profile can be used to redirect treatment and prevention strategies toward areas of unmet need.

For T2D, the genetic correlations were strongly linearly correlated with longitudinal hazard ratio estimates. Consistently, the ARIC and EHR T2D phenotypes were also strongly genetically correlated. Together, these results indicate that the T2D cases in the 2 data sets had similar epidemiological and genetic profiles. This could be expected, as the EHR and ARIC case definitions are based on standardized measures of serum glucose levels.[35] Our findings also show that genetic factors, which predispose to T2D, modulate a wide range of physiological measures beyond serum glucose, including lipid levels, blood pressure, hematologic parameter, inflammatory markers, and adiposity, consistent with findings from epidemiological and genetic studies.[36,37]

The correlation between the genetic correlation and hazard ratio estimates was weaker for the IHD phenotype, consistent with the weaker genetic correlation between the EHR and ARIC incident IHD phenotypes. The EHR IHD phenotype was genetically correlated with several phenotypes comprising the metabolic syndrome, including low HDL levels, elevated triglycerides, and elevated SBP.[38] In ARIC, LDL cholesterol and apolipoprotein B were strongly associated with incident IHD but were not significantly genetically correlated with the EHR IHD or myocardial infarction phenotypes. The genetic correlation between LDL and coronary artery disease has been previously reported to be 0.25,[39] which is similar to the genetic correlation between the ARIC LDL and ARIC IHD phenotypes (rG=0.33). Although our study was not powered to detect correlations this small (we had only 20% power to detect a genetic correlation of <0.20[40]), our point estimate was considerably smaller than this (rG=0.07), suggesting that the correlation in our data set may be lower than these estimates. One explanation for this discrepancy is that the LDL-associated risk is strongly driven by environmental LDL modulators, such as diet.[41] Some ARIC subjects were also taking lipid-lowering medications, which attenuate genetic heritability estimates as their LDL levels do not accurately reflect their genetically determined levels. We found that both LDL and apolipoprotein B were strongly genetically correlated with an EHR peripheral vascular disease phenotype,[42] however, indicating the LDL could be associated with other atherosclerotic EHR phenotypes. Another explanation is that subjects in the EHR data set were effectively treated with LDL-lowering medications, which attenuated the contribution of LDL to IHD risk in this population. C-reactive protein and fibrinogen were also more strongly associated with IHD in

Table 2. Associations Between ARIC Risk Factors and the EHR IHD (Based on Genetic Correlations) and ARIC IHD (Based on Hazard Ratios) Phenotypes

	Genetic Correlation With EHR IHD*		Hazard Ratio for ARIC IHD†	
ARIC Risk Factor	rG (SE)	*P* Value	Hazard Ratio (95% CI)	*P* Value
CIMT	0.73 (0.28)	0.002	1.31 (1.2–1.4)	6.3E-23
Cornell voltage	0.59 (0.23)	0.004	1.17 (1.1–1.2)	2.3E-07
HDL cholesterol	–0.48 (0.18)	0.005	0.59 (0.5–0.6)	3.1E-37
QTc	0.52 (0.21)	0.007	1.20 (1.1–1.3)	3.6E-09
Total:HDL cholesterol ratio	0.44 (0.17)	0.008	1.72 (1.6–1.8)	3.1E-86
Serum protein C	0.28 (0.13)	0.01	1.15 (1.1–1.2)	7.6E-06
Apolipoprotein A	–0.45 (0.20)	0.02	0.73 (0.7–0.8)	6.7E-20
Systolic BP	0.44 (0.20)	0.02	1.34 (1.3–1.4)	1.4E-21
Serum triglycerides	0.38 (0.17)	0.02	1.38 (1.3–1.5)	6.8E-33
Smoking	0.48 (0.25)	0.03	1.21 (1.2–1.3)	5.3E-15
Serum insulin	0.45 (0.26)	0.06	1.37 (1.3–1.5)	1.4E-25
Serum glucose	0.35 (0.20)	0.07	1.27 (1.2–1.3)	2.8E-21
Serum uric acid	0.37 (0.23)	0.08	1.14 (1.1–1.2)	3.5E-05
aPTT	0.23 (0.16)	0.15	0.94 (0.9–1.0)	3.7E-02
White blood cell count	0.24 (0.17)	0.16	1.36 (1.3–1.4)	1.4E-24
Waist–hip ratio	0.24 (0.18)	0.18	1.45 (1.3–1.6)	1.4E-16
Eosinophil percent	0.29 (0.23)	0.20	1.05 (1.0–1.1)	1.7E-01
Height	–0.14 (0.12)	0.25	0.87 (0.8–1.0)	1.6E-03
Platelet count	–0.15 (0.14)	0.26	1.02 (1.0–1.1)	4.4E-01
Mean corpuscular volume	–0.17 (0.20)	0.40	0.86 (0.8–0.9)	8.7E-06
Subscapular skinfold	0.13 (0.18)	0.46	1.39 (1.3–1.5)	5.5E-15
Serum potassium	0.13 (0.20)	0.49	0.93 (0.9–1.0)	2.1E-02
Apolipoprotein B	0.12 (0.20)	0.55	1.42 (1.3–1.5)	5.7E-38
Serum creatinine	0.10 (0.19)	0.59	1.05 (1.0–1.1)	2.3E-01
Diastolic BP	–0.10 (0.19)	0.60	1.11 (1.0–1.2)	9.9E-04
Fibrinogen level	–0.10 (0.19)	0.61	1.33 (1.3–1.4)	1.0E-26
LDL cholesterol	0.07 (0.20)	0.72	1.38 (1.3–1.5)	1.1E-27
Body mass index	0.06 (0.18)	0.75	1.23 (1.1–1.3)	1.6E-08
Total protein	0.06 (0.20)	0.77	1.05 (1.0–1.1)	1.2E-01
Total cholesterol	0.04 (0.17)	0.83	1.35 (1.3–1.4)	2.1E-25
Hs C-reactive protein	–0.05 (0.22)	0.83	1.32 (1.2–1.4)	4.8E-16
Serum calcium	–0.04 (0.23)	0.85	1.00 (0.9–1.1)	9.5E-01
Heart rate	0.04 (0.20)	0.86	1.12 (1.1–1.2)	7.4E-05
Factor VII level	0.02 (0.17)	0.90	1.18 (1.1–1.3)	1.7E-07
von Willebrand factor	0.01 (0.16)	0.94	1.16 (1.1–1.2)	1.6E-06
Waist circumference	0.00 (0.16)	0.97	1.29 (1.2–1.4)	6.5E-12
Hemoglobin level	–0.01 (0.22)	0.97	1.15 (1.1–1.2)	6.0E-04

aPTT indicates activated partial thromboplastin time; ARIC, Atherosclerosis Risk in Communities; BP, blood pressure; CI, confidence interval; CIMT, carotid intimal-medial thickness; EHR, electronic health record; GLMM, generalized linear mixed model; Hs, high sensitivity; IHD, ischemic heart disease; PC, principal component; rG, genetic correlation; and QTc, corrected QT interval.

*Genetic correlations between the EHR IHD phenotype and each ARIC risk factor were computed using a bivariate GLMM adjusting for age, sex, and 20 PCs.

†Hazard ratios units for incident IHD in the ARIC set are per SD increase in the risk factor and are adjusted for sex and age.

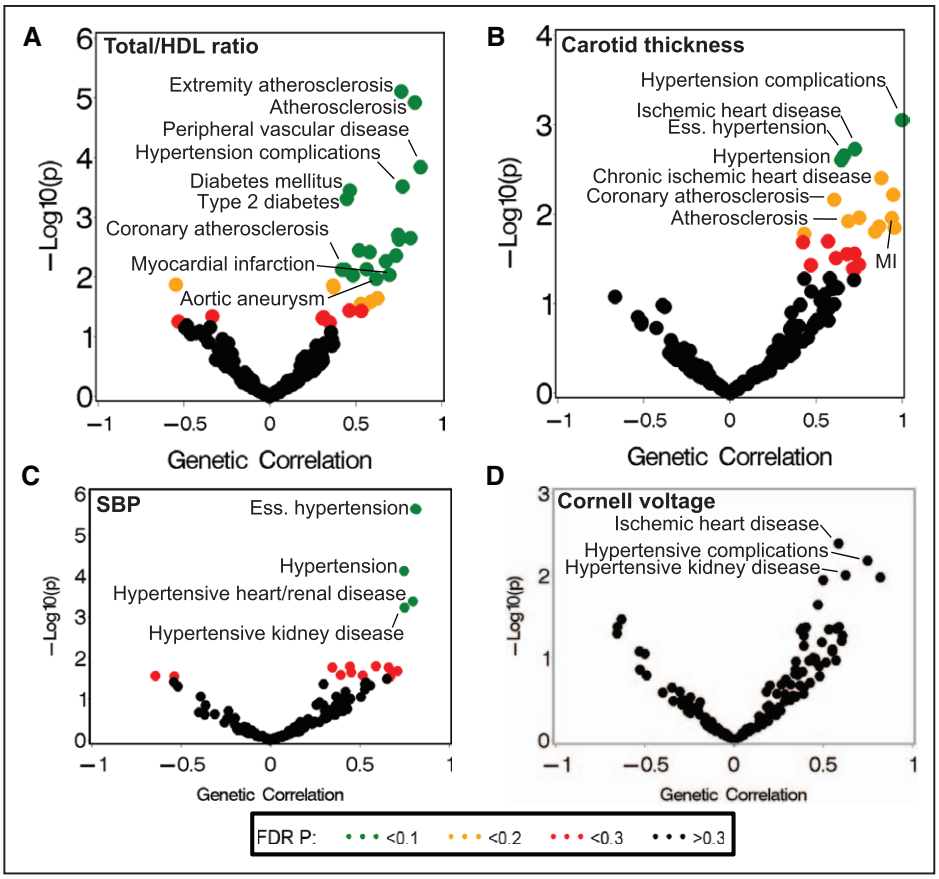

Figure 4. Electronic health record (EHR) phenotypes genetically correlated with ARIC study (Atherosclerosis Risk in Communities) risk factors and biomarkers. Pairwise genetic correlations between 158 EHR phenotypes and the ARIC phenotypes of (**A**) total:HDL cholesterol ratio (n=7701, genetic heritability [h2]=0.20[0.04]), (**B**) log carotid intima-medial arterial wall thickness (n=7315, h2=0.10[0.04]), (**C**) systolic blood pressure (SBP; n=7710, h2=0.16[0.04], and (**D**) Cornell voltage from cardiac ECG (n=7432, h2=0.13[0.04]). HDL indicates high-density lipoprotein; and FDR, false discovery rate.

ARIC. However, a difference in a genetic versus epidemiological association for these factors would be expected because genetic risk scores for these traits do not predict IHD.[43,44]

A strength of a prospective study is that it can delineate the extended set of outcomes associated with a biomarker measured at baseline.[5] To emulate this feature of prospective study, we used a large set of clinical phenotypes extracted from the EHR data set to generate a library of genetically modulated diseases. We then used genetic correlation analyses to identify diseases from this library associated with an ARIC risk factor. Using this discovery approach, we showed that 3 IHD risk factors (CIMT, SBP, and total:HDL cholesterol ratio) had significant correlations with atherosclerotic diseases, hypertension-related diseases, and T2D. Thus, these analyses highlight the contributions of these disease processes to IHD risk and indicate that interventions and therapies directed toward these processes would benefit this EHR population.

An advantage of measuring associations using genetic correlations is that associations are not confounded by environment because the association is based solely on underlying genetics.[45] However, genetic factors can modulate exposures to environmental factors, which can result in genetic correlations with environmental risk factors.[46] For instance, we found that pack-years of smoking measured in the ARIC study was genetically correlated with tobacco use disorder, peripheral artery disease, and IHD. While smoking is an environmental toxin and would not be expected to be an intrinsically heritable biomarker, smoking behaviors are genetically influenced.[47] Hence, a genetic predisposition toward high cumulative exposures to cigarettes is associated with clinically significant morbidity.

There are several limitations to this study. EHR PheCode phenotypes rely on clinical disease assignments that are often not concisely defined. Hence, even though the ARIC and EHR phenotypes evaluated in these analyses had similar clinical definitions, case status in the ARIC study was assigned based on active surveillance and standardized case definitions, whereas the EHR phenotypes were based on observational data and represent incident and prevalent cases detected through routine clinical care. These differences in ascertainment can contribute to unexpected differences between the phenotypes that could alter the patterns of risk factor associations between them. Because the EHR data set comprised incident and prevalent disease, factors associated with their acute presentation could not be ascertained. Consequently, we could not adjust for medication use at the time of the diagnosis or the presenting illness. Not factoring in medication use at the time of diagnosis could attenuate associations. The phenotypes were derived from billing code data, which can be incomplete, leading to misclassification in control groups. However, the phenotyping methodologies used in these analyses have been extensively studied and perform

well in genetic studies.[28,48] The cases and controls for the EHR IHD phenotype differed with respect to many baseline characteristics. These differences, such as the younger age of controls, can contribute to phenotype misclassification, which would attenuate associations with risk factors. Furthermore, prevalent IHD, which cannot easily be diagnosed based on routine clinical testing, may not be effectively captured in an EHR data set, which would lead to further misclassification. In the future, because ongoing longitudinal epidemiological studies grow and accumulate sufficiently large numbers of clinical end points, it will be informative to apply this analytic approach to determine whether the patterns of associations that we observed in the EHR population are seen with comparisons across other well-defined longitudinal cohorts. For many of the EHR phenotypes, there were relatively few cases, which decreases the power to detect significant genetic correlations and can result in false-negative findings. The ARIC phenotypes used in this study were limited to risk factors, which have previously been associated with IHD. Hence, we were not able to identify new associations in these analyses. It is also important to note that a genetic correlation does not always indicate that a pair of phenotypes has a shared mechanism because it is possible that SNPs used to compute the correlations may be simultaneously tagging disparate causative genetic variants.[49] Finally, these analyses were limited to subjects of European ancestry because the numbers of subjects of other ancestries were insufficient for analysis. Further validation of this approach in other ancestral groups is needed.

In summary, we demonstrate that genetic correlations using cross-sectional data derived from separate data sets can identify clinical phenotypes associated with biomarkers and can recapitulate longitudinal associations. We also demonstrate the use of EHR data sets as a source of clinically relevant disease outcomes that can be linked to genetically modulated preclinical markers ascertained in epidemiological cohorts. We anticipate that this analytic paradigm will prove useful because new putative biomarkers are identified through large-scale proteomics studies and other discovery approaches because it enables rapid identification of clinically relevant phenotypes associated with these markers and overcomes several limitations inherent to longitudinal studies designs.

Acknowledgments

We thank the staff and participants of the ARIC study (Atherosclerosis Risk in Communities) for their contributions.

Sources of Funding

This work was supported by a career development award from the Vanderbilt Faculty Research Scholars Fund (Dr Mosley), American Heart Association (15MCPRP25620006 and 16FTF30130005; Dr Mosley), and PGRN (P50-GM115305) and R01 LM010685. BioVU is supported by institutional funding and by Clinical and Translational Science Awards grant UL1 TR000445 from National Center for Advancing Translational Sciences/National Institutes of Health. The Electronic Medical Records and Genomics (eMERGE) Network is funded by National Human Genome Research Institute and National Institute of General Medical Sciences: U01-HG8672 and U01-HG006378 (VUMC); U01-HG-004610 (Group Health Cooperative/University of Washington); U01-HG-004608 (Marshfield Clinic Research Foundation and VUMC); U01-HG-04599 (Mayo Clinic); U01HG004609 (Northwestern University); U01-HG-006378 and U01-HG-04603 (VUMC Coordinating Center); U01HG004438 (Center for Inherited Disease Research); and U01HG004424 (the Broad Institute) serving as Genotyping Centers. ARIC study (Atherosclerosis Risk in Communities) is supported by National Heart, Lung and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C,HHSN268201100011C,andHHSN268201100012C). Funding for GENEVA was provided by National Human Genome Research Institute grant U01HG004402 (E. Boerwinkle).

Disclosures

None.

References

1. Lloyd-Jones D, Adams RJ, Brown TM, Carnethon M, Dai S, De Simone G, et al; WRITING GROUP MEMBERS; American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics–2010 update: a report from the American Heart Association. *Circulation*. 2010;121:e46–e215. doi: 10.1161/CIRCULATIONAHA.109.192667.
2. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet*. 2014;383:999–1008. doi: 10.1016/S0140-6736(13)61752-3.
3. Wong ND. Epidemiological studies of CHD and the evolution of preventive cardiology. *Nat Rev Cardiol*. 2014;11:276–289. doi: 10.1038/nrcardio.2014.26.
4. Ruff CT, Braunwald E. The evolving epidemiology of acute coronary syndromes. *Nat Rev Cardiol*. 2011;8:140–147. doi: 10.1038/nrcardio.2010.199.
5. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. *Lancet*. 2002;359:341–345. doi: 10.1016/S0140-6736(02)07500-1.
6. Denny JC. Chapter 13: Mining electronic health records in the genomics era. *PLoS Comput Biol*. 2012;8:e1002823. doi: 10.1371/journal.pcbi.1002823.
7. McPherson R, Tybjaerg-Hansen A. Genetics of coronary artery disease. *Circ Res*. 2016;118:564–578. doi: 10.1161/CIRCRESAHA.115.306566.
8. Maher BS. Polygenic scores in epidemiology: risk prediction, etiology, and clinical utility. *Curr Epidemiol Rep*. 2015;2:239–244. doi: 10.1007/s40471-015-0055-3.
9. Smith JA, Ware EB, Middha P, Beacher L, Kardia SL. Current applications of genetic risk scores to cardiovascular outcomes and subclinical phenotypes. *Curr Epidemiol Rep*. 2015;2:180–190. doi: 10.1007/s40471-015-0046-4.
10. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. 2003;32:1–22. doi: 10.1093/ije/dyg070.
11. Holmes MV, Lange LA, Palmer T, Lanktree MB, North KE, Almoguera B, et al. Causal effects of body mass index on cardiometabolic traits and events: a Mendelian randomization analysis. *Am J Hum Genet*. 2014;94:198–208. doi: 10.1016/j.ajhg.2013.12.014.
12. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet*. 2011;43:519–525. doi: 10.1038/ng.823.
13. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 2010;42:565–569. doi: 10.1038/ng.608.
14. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet*. 2014;46:100–106. doi: 10.1038/ng.2876.
15. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*. 2012;28:2540–2542. doi: 10.1093/bioinformatics/bts474.
16. Vattikuti S, Guo J, Chow CC. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet*. 2012;8:e1002637. doi: 10.1371/journal.pgen.1002637.
17. The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am J Epidemiol*. 1989;129:687–702. doi: 10.1093/oxfordjournals.aje.a115184.
18. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al; eMERGE Network. The Electronic Medical Records and Genomics

(eMERGE) Network: past, present, and future. *Genet Med*. 2013;15:761–771. doi: 10.1038/gim.2013.72.
19. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al; eMERGE Team. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics*. 2011;4:13. doi: 10.1186/1755-8794-4-13.
20. Bowton E, Field JR, Wang S, Schildcrout JS, van Driest SL, Delaney JT, et al. Biobanks and electronic medical records: enabling cost-effective research. *Sci Transl Med*. 2014;6:234cm3. doi: 10.1126/scitranslmed.3008604.
21. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther*. 2008;84:362–369. doi: 10.1038/clpt.2008.89.
22. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000;155:945–959.
23. Zuvich RL, Armstrong LL, Bielinski SJ, Bradford Y, Carlson CS, Crawford DC, et al. Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. *Genet Epidemiol*. 2011;35:887–898. doi: 10.1002/gepi.20639.
24. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods*. 2013;10:5–6. doi: 10.1038/nmeth.2307.
25. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*. 2012;44:955–959. doi: 10.1038/ng.2354.
26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559–575. doi: 10.1086/519795.
27. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. 2010;26:1205–1210. doi: 10.1093/bioinformatics/btq126.
28. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31:1102–1110. doi: 10.1038/nbt.2749.
29. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88:76–82. doi: 10.1016/j.ajhg.2010.11.011.
30. Aschard H, Vilhjálmsson BJ, Joshi AD, Price AL, Kraft P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am J Hum Genet*. 2015;96:329–339. doi: 10.1016/j.ajhg.2014.12.021.
31. Majumdar A, Haldar T, Witte JS. Determining which phenotypes underlie a pleiotropic signal. *Genet Epidemiol*. 2016;40:366–381. doi: 10.1002/gepi.21973.
32. Crow RS, Prineas RJ, Rautaharju P, Hannan P, Liebson PR. Relation between electrocardiography and echocardiography for left ventricular mass in mild systemic hypertension (results from Treatment of Mild Hypertension Study). *Am J Cardiol*. 1995;75:1233–1238. doi: 10.1016/S0002-9149(99)80769-3.
33. Weber C, Noels H. Atherosclerosis: current pathogenesis and therapeutic options. *Nat Med*. 2011;17:1410–1422. doi: 10.1038/nm.2538.
34. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al. Heart disease and stroke statistics-2016 update: a report from the American Heart Association. *Circulation*. 2016;133:e38–e360. doi: 10.1161/CIR.0000000000000350.
35. Inzucchi SE. Clinical practice. Diagnosis of diabetes. *N Engl J Med*. 2012;367:542–550. doi: 10.1056/NEJMcp1103643.
36. Raynor LA, Pankow JS, Duncan BB, Schmidt MI, Hoogeveen RC, Pereira MA, et al. Novel risk factors and the prediction of type 2 diabetes in the Atherosclerosis Risk in Communities (ARIC) study. *Diabetes Care*. 2013;36:70–76. doi: 10.2337/dc12-0609.
37. Long MT, Fox CS. The Framingham Heart Study–67 years of discovery in metabolic disease. *Nat Rev Endocrinol*. 2016;12:177–183. doi: 10.1038/nrendo.2015.226.
38. Eckel RH, Grundy SM, Zimmet PZ. The metabolic syndrome. *Lancet*. 2005;365:1415–1428. doi: 10.1016/S0140-6736(05)66378-7.
39. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al; ReproGen Consortium; Psychiatric Genomics Consortium; Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3. An atlas of genetic correlations across human diseases and traits. *Nat Genet*. 2015;47:1236–1241. doi: 10.1038/ng.3406.
40. Visscher PM, Hemani G, Vinkhuyzen AA, Chen GB, Lee SH, Wray NR, et al. Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet*. 2014;10:e1004269. doi: 10.1371/journal.pgen.1004269.
41. Hu FB, Willett WC. Optimal diets for prevention of coronary heart disease. *JAMA*. 2002;288:2569–2578. doi: 10.1001/jama.288.20.2569.
42. Smith SC, Milani RV, Arnett DK, Crouse JR, McDermott MM, Ridker PM, et al. Atherosclerotic vascular disease conference: Writing Group II: risk factors. *Circulation*. 2004;109:2613–2616. doi: 10.1161/01.CIR.0000128519.60762.84.
43. Sabater-Lleal M, Huang J, Chasman D, Naitza S, Dehghan A, Johnson AD, et al; VTE Consortium; STROKE Consortium; Wellcome Trust Case Control Consortium 2 (WTCCC2); C4D Consortium; CARDIoGRAM Consortium. Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. *Circulation*. 2013;128:1310–1324. doi: 10.1161/CIRCULATIONAHA.113.002251.
44. Zacho J, Tybjaerg-Hansen A, Jensen JS, Grande P, Sillesen H, Nordestgaard BG. Genetically elevated C-reactive protein and ischemic vascular disease. *N Engl J Med*. 2008;359:1897–1908. doi: 10.1056/NEJMoa0707402.
45. Ebrahim S, Davey Smith G. Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Hum Genet*. 2008;123:15–33. doi: 10.1007/s00439-007-0448-6.
46. Gage SH, Davey Smith G, Ware JJ, Flint J, Munafò MR. G = E: what GWAS can tell us about the environment. *PLoS Genet*. 2016;12:e1005765. doi: 10.1371/journal.pgen.1005765.
47. Amos CI, Spitz MR, Cinciripini P. Chipping away at the genetics of smoking behavior. *Nat Genet*. 2010;42:366–368. doi: 10.1038/ng0510-366.
48. Mosley JD, Witte JS, Larkin EK, Bastarache L, Shaffer CM, Karnes JH, et al. Identifying genetically driven clinical phenotypes using linear mixed models. *Nat Commun*. 2016;7:11433. doi: 10.1038/ncomms11433.
49. Gianola D, de los Campos G, Toro MA, Naya H, Schön CC, Sorensen D. Do molecular markers inform about pleiotropy? *Genetics*. 2015;201:23–29. doi: 10.1534/genetics.115.179978.

CLINICAL PERSPECTIVE

Although the gold standard for defining epidemiological risk profiles is longitudinal studies, this approach is hampered by a need for long observation times. Here, we tested the idea that electronic health records systems coupled to dense genomic data can be leveraged to circumvent this time constraint. We describe an approach that uses baseline risk factors from epidemiological studies (here, we studied the ARIC study [Atherosclerosis Risk in Communities] cohort) and measures their association with electronic health record phenotypes. This approach is enabled by genetics methods that measure the extent to which phenotypes are modulated by a common set of genetic variants. We evaluate electronic health record phenotypes for type 2 diabetes mellitus and ischemic heart disease using data collected through the Electronic Medical Records and Genomics network. We show that associations identified using our genetic approach are similar to those identified using longitudinal association approaches. Our analyses validate systolic blood pressure, triglyceride, and total-to-HDL cholesterol levels as important measures of ischemic heart disease risk. This analytic paradigm should prove useful not only to evaluate the contribution of known ischemic heart disease risk factors to contemporary populations, but also to rapidly evaluate new putative risk factors and biomarkers of disease.