Peripheral artery disease (PAD) is a leading cause of cardiovascular morbidity and mortality; however, the extent to which genetic factors increase risk for PAD is largely unknown. Using electronic health record data, we performed a genome-wide association study in the Million Veteran Program testing ~32 million DNA sequence variants with PAD (31,307 cases and 211,753 controls) across veterans of European, African and Hispanic ancestry. The results were replicated in an independent sample of 5,117 PAD cases and 389,291 controls from the UK Biobank. We identified 19 PAD loci, 18 of which have not been previously reported. Eleven of the 19 loci were associated with disease in three vascular beds (coronary, cerebral, peripheral), including LDLR, LPL and LPA, suggesting that therapeutic modulation of low-density lipoprotein cholesterol, the lipoprotein lipase pathway or circulating lipoprotein(a) may be efficacious for multiple atherosclerotic disease phenotypes. Conversely, four of the variants appeared to be specific for PAD, including F5 p.R506Q, highlighting the pathogenic role of thrombosis in the peripheral vascular bed and providing genetic support for Factor Xa inhibition as a therapeutic strategy for PAD. Our results highlight mechanistic similarities and differences among coronary, cerebral and peripheral atherosclerosis and provide therapeutic insights.

PAD is a complex disease impacted by both lifestyle and inheritance. Despite its high prevalence, only a few studies have evaluated PAD genetics, with published genome-wide association studies (GWAS) having revealed only three loci reaching genome-wide significance. Furthermore, it is uncertain if the genetic mechanisms underlying atherosclerotic disease of the peripheral arteries and the coronary and cerebral arteries are shared or distinct. Large-scale biobanks combining genetic data with electronic health record (EHR)-derived phenotypes are under development throughout the world. The Million Veteran Program (MVP) was established in 2011 to study how genes affect health in the Veterans
Affairs (VA) Healthcare System. Approximately 10% of individuals greater than the age of 55 seeking care in the VA Healthcare System have PAD, making MVP an ideal cohort for performing a large-scale PAD genetic analysis. Leveraging the MVP resource, we sought to: (1) perform a genetic discovery analysis of PAD; (2) explore the spectrum of phenotypic consequences associated with PAD risk variants; and (3) identify genetic signals that differentiate PAD from vascular disease in other arterial beds.

We designed a two-phased GWAS (Fig. 1). Initial discovery was performed in MVP, testing for association separately among individuals of European (white), African (black) and Hispanic ancestry. The results were then meta-analyzed across ancestral groups. For variants with suggestive associations ($P < 10^{-4}$) with PAD, we sought replication in the UK Biobank. We then combined statistical evidence across MVP and UK Biobank and set a significance threshold of $P < 5 \times 10^{-8}$ (genome-wide significance).

The MVP discovery analysis was comprised of 31,307 individuals (24,009 white, 5,373 black, 1,925 Hispanic) with PAD and 211,753 PAD-free controls; their baseline characteristics are presented in Supplementary Table 1. Participants with PAD were more likely to be older, male, prescribed statin therapy, have a history of smoking and affected with type 2 diabetes (T2D). To validate our PAD phenotype, the minimum ankle–brachial index (mABI) was extracted for 17,861 individuals with ABI measurements available in MVP. As expected, we observed a median mAABI of less than 0.9 for PAD cases and approximately 1 for PAD controls across all three ethnic groups (Supplementary Table 2 and Extended Data Fig. 1). We further validated our MVP PAD phenotype with manual chart reviews and observed a specificity of 88% (95% confidence interval (CI) = 75.7–94.5%) and sensitivity of 100% (95% CI = 89.8–100%), commensurate with that published in the literature.

Through genotype imputation, we obtained 20.3 million, 32.4 million and 31.2 million DNA sequence variants for analysis in white, black and Hispanic participants, respectively (Supplementary Table 1). Following trans-ethnic meta-analysis in the discovery phase, a total of 554 variants at 25 loci met a genome-wide significance threshold (Extended Data Fig. 2). We replicated all three previously described genome-wide PAD loci with at least nominal ($P < 0.05$) significance (Supplementary Table 3). A total of 1,276 variants demonstrated association $P < 10^{-8}$ in the MVP discovery analysis. Of those, 552 were also available for independent testing in the UK Biobank (5,117 PAD cases, 389,291 controls) and were taken forward for replication. Following replication, 19 loci exceeded genome-wide significance ($P < 5 \times 10^{-8}$, Table 1 and Supplementary Table 4). Of the 19 PAD loci, 15 were directionally consistent across whites, blacks and Hispanics in MVP, 8 demonstrated at least nominal significance in blacks and 3 in Hispanics (Supplementary Table 5); 18 of the loci have not been previously reported (Extended Data Fig. 3).

The $LPA$ variant rs118039278 was the top association result (6.4% frequency for the A allele; odds ratio (OR) = 1.25; 95% CI: 1.22–1.30; $P = 1.57 \times 10^{-45}$). Of the six signals from MVP that did not replicate in the UK Biobank, two were rare variants that were not available in UK Biobank following quality control (European minor allele frequency (MAF) < 0.005), and the remaining four did not meet the pre-specified $P < 0.05$ for independent replication (Supplementary Table 6). All three previously reported suggestive ($5.0 \times 10^{-4} < P < 0.05$) PAD associations at the $SH2B3$/$PTPN11$ (ref. 8), $HDAC9$ (ref. 9) and $CHRNA3$ (ref. 9) loci were observed at genome-wide significance.

We next studied whether DNA sequence variants were associated with PAD severity as determined by mAABI. We performed a GWAS of mAABI as a continuous trait for 13,382 European, 3,284 African and 998 Hispanic ancestry individuals in MVP, restricted to those with an ABI < 1.4 as previously described. Baseline characteristics for these individuals are depicted in Supplementary Table 7. Following trans-ethnic meta-analysis, only the known 9p21-ABI association passed the genome-wide significance threshold (rs1333045, 46.8% frequency for the T allele; $P = 0.064$; 95% CI: 0.042–0.086; $P = 8.3 \times 10^{-4}$). However, we observed that 6 of the 19 PAD risk variants identified in our PAD case/control analysis were associated with reduced mAABI at nominal significance ($P < 0.05$, Supplementary Table 8). Notably, the mAABI GWAS lead 9p21 variant (rs1333045) was different than the lead variant identified in the PAD case-control analysis at this locus (rs1537372).

Understanding the full spectrum of phenotypic consequences of a given DNA sequence variant can help to identify the mechanism by which a variant or gene leads to disease. Termed a phenome-wide association study (PheWAS), this approach examines the association of a risk variant across a range of phenotypes (Extended Data Fig. 4). Using a median of 65 distinct EHR-derived diagnosis codes (International Classification of Disease 9th and 10th revisions (ICD-9/10)) per participant, we tested each of the 19 PAD lead risk variants across 1,101 disease phenotypes. We found that several of the newly identified DNA sequence variants correlated with a range of known risk factors for PAD (Fig. 2 and Supplementary Table 9). For example, rs7903146 within $TCF7L2$ is one of the strongest known genetic predictors of T2D12 and associated with T2D in our PheWAS. The PAD association for rs7903146 was significantly reduced when controlling for T2D in the regression model, suggesting this variant confers PAD risk through its effect on T2D (Extended Data Fig. 4). The Factor V Leiden variant (F5 p.R506Q) is one of the strongest known genetic predictors of T2D and associated with T2D in our PheWAS. The PAD association for rs7903146 was significantly reduced when controlling for T2D in the regression model, suggesting this variant confers PAD risk through its effect on T2D. The association of DNA sequence variants with PAD was tested separately in three mutually exclusive ancestry groups and the results combined using an inverse-variance weighted fixed-effects meta-analysis in the discovery phase. Variants with suggestive association (two-sided logistic regression $P < 10^{-4}$) were then brought forward for independent replication in the UK Biobank.
near the HLA-B gene was associated with a number of autoimmune diseases including celiac disease, Graves’ disease, systemic lupus erythematosus and type 1 diabetes. In total, we identified 158 statistically significant ($P < 5.0 \times 10^{-8}$) PhEwas associations across the 19 genetic variants implicating many known PAD risk factors based on the traits they relate to, including lipids, T2D, smoking, thrombosis and hypertension. We supplemented our MVP PhEwas with data from PhenoScanner V2 (ref. 19), an online resource of association statistics from previously conducted GWAS and UK Biobank. In total, we identified 443 additional PhEwas associations from the PhenoScanner database at $P < 5.0 \times 10^{-8}$ (Supplementary Table 10). We subsequently prioritized likely candidate causal PAD risk genes by aggregating evidence from (1) prior genetic, clinical, or functional studies, (2) our PhEwas results, (3) cis-expression quantitative trait locus (eQTL) from the Genotype-Tissue Expression Project (GTEx) V7 data set, (4) recently published protein quantitative trait locus (pQTL) data derived from the human plasma of 3,301 participants of the INTERVAL study, and (5) results from a transcriptome-wide association study (TWAS) using RNA-seq data from post-mortem tibial artery tissue (388 individuals) and MVP European PAD summary statistics. This analysis revealed several candidate causal genes including $F5$, $LPA$, $SORT1$, $LPL$ and $LDLR$ (Supplementary Tables 11 and 12).

We next sought to understand how DNA sequence variants might differ in their contribution to vascular disease risk in the peripheral, coronary and cerebral arterial territories. Analysis of shared heritability provides a mechanism that describes the relationship of common variant risk across phenotypes. Using linkage disequilibrium score regression, we examined the genetic correlation ($r_g$) between PAD and both coronary artery disease (CAD) and large artery stroke (LAS). We used summary statistics from the European MVP PAD analysis, along with summary data of 60,801 coronary disease cases and 123,504 disease-free controls from the CARDIoGRAMplusC4D consortia, and 6,688 LAS cases and 454,450 controls from the 2018 MEGASTROKE analysis. We noted a stronger positive correlation between PAD and both coronary artery disease ($r_g = 0.88, P = 5.5 \times 10^{-10}$) than for PAD and CAD ($r_g = 0.62, P = 1.57 \times 10^{-4}$). Based on these findings, we explored the differential effects of individual genetic variants on PAD, LAS and CAD.

### Table 1 | PAD risk loci discovered in the MVP biobank and replicated in the UK Biobank

| Chr:Pos | rsid   | EA | NEA | EAF | Overall OR$^a$ | Overall 95% CI$^b$ | Overall P$^c$ | Annotation                  | Gene/Locus$^d$ |
|---------|--------|----|-----|-----|---------------|-------------------|---------------|-----------------------------|----------------|
| 1:10981792 | rs7528419 | A  | G   | 0.772 | 1.07 | 1.05–1.09 | 2.54 $\times 10^{-6}$ | 3’ UTR variant | CELSR2/SORT1            |
| 1:169519049 | rs6025 | T  | C   | 0.026 | 1.2 | 1.14–1.26 | 1.63 $\times 10^{-12}$ | Missense variant (Factor V Leiden) | $FS$           |
| 6:160985526 | rs11039278 | A  | G   | 0.068 | 1.26 | 1.22–1.30 | 1.57 $\times 10^{-4}$ | Intron variant | $LPA$         |
| 6:31065071 | rs3130968 | T  | C   | 0.144 | 1.07 | 1.05–1.10 | 3.16 $\times 10^{-10}$ | Regulatory region variant | (HLA-B)    |
| 7:19049388 | rs2107595 | G  | A   | 0.187 | 1.08 | 1.05–1.10 | 2.49 $\times 10^{-13}$ | Regulatory region variant | (HDAC9)     |
| 7:22786532 | rs47721727 | G  | A   | 0.202 | 1.08 | 1.05–1.10 | 3.65 $\times 10^{-10}$ | Intergenic variant | (IL6)      |
| 8:19819217 | rs332 | A  | C   | 0.706 | 1.06 | 1.04–1.07 | 2.53 $\times 10^{-9}$ | Introgen variant | $LPL$        |
| 9:136149229 | rs505922 | C  | T   | 0.334 | 1.06 | 1.04–1.07 | 7.10 $\times 10^{-8}$ | Intron variant | $ABO$         |
| 9:22103183 | rs1537372 | T  | G   | 0.421 | 1.12 | 1.10–1.14 | 4.32 $\times 10^{-9}$ | Intron variant | CDKN2B-AS1/9p21 |
| 10:114758349 | rs7903146 | C  | T   | 0.293 | 1.06 | 1.04–1.08 | 3.76 $\times 10^{-6}$ | Intron variant | TCF7L2          |
| 11:102710471 | rs566125 | T  | C   | 0.127 | 1.08 | 1.05–1.11 | 4.37 $\times 10^{-9}$ | Intron variant | MMP3          |
| 11:46342834 | rs7476 | C  | A   | 0.364 | 1.06 | 1.04–1.08 | 8.33 $\times 10^{-3}$ | 3’ UTR variant | CREB3L1       |
| 12:112871372 | rs11066301 | G  | A   | 0.413 | 1.06 | 1.04–1.08 | 2.96 $\times 10^{-6}$ | Intron variant | PTPN11        |
| 12:79951566 | rs48442266 | G  | A   | 0.388 | 1.06 | 1.04–1.07 | 1.01 $\times 10^{-9}$ | Upstream gene variant | RP11-359M6.3 |
| 13:110828891 | rs1975514 | C  | T   | 0.357 | 1.05 | 1.04–1.07 | 8.32 $\times 10^{-10}$ | Intron variant | COL4A1         |
| 14:70501364 | rs55784307 | A  | C   | 0.183 | 1.06 | 1.04–1.09 | 2.93 $\times 10^{-8}$ | Downstream gene variant | SMO2C         |
| 15:78915864 | rs10851907 | A  | G   | 0.41 | 1.06 | 1.05–1.08 | 1.49 $\times 10^{-10}$ | Upstream gene variant | CHRNA3       |
| 17:66089393 | rs62084752 | C  | G   | 0.216 | 1.07 | 1.05–1.09 | 1.58 $\times 10^{-8}$ | Upstream gene variant | LOC732538 |
| 19:11191729 | rs138294113 | C  | T   | 0.879 | 1.09 | 1.06–1.11 | 1.20 $\times 10^{-10}$ | Intergenic variant | (LDLR)          |

$^a$Overall OR, 95% CI and P (two-sided) represent logistic regression statistics following meta-analysis of MVP and UK Biobank (N = 36,424 PAD cases and 601,044 controls). $^b$Genes for variants that are outside the transcript boundary of a protein-coding gene are shown with nearest candidate gene in parentheses (for example, (HLA-B)). $^c$Overall OR, 95% CI and P (two-sided) represent logistic regression statistics following meta-analysis of MVP and UK Biobank (total N = 36,424 PAD cases and 601,044 controls). $^d$Genes for variants that are outside the transcript boundary of a protein-coding gene are shown with nearest candidate gene in parentheses (for example, (HLA-B)).
The 19 lead PAD risk variants identified in our GWAS analysis were tested for their effects on CAD and LAS in white MVP participants. The results were then summarized with combined statistics from the CARDIoGRAMplusC4D or MEGASTROKE studies, respectively. We observed that 14 PAD risk variants demonstrated at least nominal association (p < 0.05) with CAD, and 12 with LAS (Supplementary Tables 13–16). In a sensitivity analysis, the PAD effect estimates at the SORT1, LPA, 9p21 and LDLR loci were attenuated, suggesting that some of the PAD risk may be driven by co-morbidity or shared causal pathways when accounting for the concomitant CAD and LAS diagnoses (Supplementary Tables 17–19). The COL4A1 locus, previously associated with CAD[4], and small vessel disease of the brain[3], was found to be associated with PAD and CAD but not with LAS in our analysis. Data from the MEGASTROKE study demonstrate evidence of association with small artery stroke (P = 1.4 × 10^−8) for this variant, suggesting it may be acting differently in the cerebral bed.

Common mechanisms emerged for the 11 PAD risk variants demonstrating significant association in all three (coronary, cerebral and peripheral) vascular beds including lipids (LDLR, LPA, LPL and SORT1), hypertension (PTPN11), and diabetes (TCF7L2). Conversely, variants in the RP11-359M6.3, HLA-B, CHRNA3, and F5 loci were uniquely associated with PAD, implying that smoking and thrombosis may play a greater role in PAD than disease in other arterial territories (Extended Data Figs. 7 and 8).

The novel PAD risk variant Factor V Leiden (F5 p.R506Q) is the most common cause of inherited thrombophilia[23], as the protein-altering consequence of the variant results in a resistance to proteolysis by activated protein C[24]. In a combined analysis of 111,216 coronary disease cases and 248,081 controls from MVP (9,388 Factor V Leiden carriers) and CARDIoGRAMplusC4D, we observed no evidence of an association between F5 p.R506Q and CAD (OR = 1.01; 95% CI: 0.97–1.05; P = 0.72, Fig. 3a). Similarly, for 7,393 LAS cases and 628,737 controls from MVP and MEGASTROKE, we observed no evidence of an association between F5 p.R506Q and LAS (OR = 1.03; 95% CI: 0.89–1.20; P = 0.65, Fig. 3b). In contrast, F5 p.R506Q was associated with a 20% increased risk of PAD in individuals of European ancestry in MVP (OR = 1.20; 95% CI: 1.14–1.27; P = 8.81 × 10^−11, Fig. 3c).

To better understand Factor V Leiden’s relationship with PAD, we tested its association with increasingly severe disease manifestations, including claudication, rest pain, tissue loss and major amputation. In total, we identified 5,797 individuals with intermittent claudication, 1,000 with rest pain, 1,773 with evidence of tissue loss and 438 who had undergone a major amputation among white MVP participants (Supplementary Table 2). We observed significant associations for the Factor V Leiden mutation with each subtype of PAD (Fig. 3c). Interestingly, the variant’s effect estimate increased as PAD severity increased, with carriers having a 62% increased risk of undergoing a PAD-related major amputation (OR = 1.62; 95% CI: 1.16–2.26; P = 0.005).

Recent evidence has linked tobacco use to an increased risk for thrombotic sequelae[29,30]. We hypothesized that there may be an interaction between smoking and F5 p.R506Q carrier status on PAD risk. We observed that the presence of F5 p.R506Q had greater effect on PAD among current smokers (OR = 1.40; 95% CI: 1.25–1.58; P = 1.20; 95% CI: 1.14–1.27; P = 8.81 × 10^−11, Fig. 3c).

The association of the thrombophilic Factor V Leiden variant, F5 p.R506Q, with different types of vascular disease were analyzed, as depicted in forest plots. Associations with all PAD cases, as well as PAD cases of increasing severity (c), and PAD cases stratified by smoking status (d) among European-ancestry MVP participants are shown. Two-sided logistic regression P values are displayed. Gray boxes reflect the inverse-variance weight for each study or subgroup.
vasculature, including the Factor V Leiden variant. These results are demonstrative of how large biobanks that couple genetic variation with dense EHR data can be leveraged for biological insights that can inform clinical care.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41591-019-0492-5.

Received: 26 September 2018; Accepted: 17 May 2019; Published online: 8 July 2019

References
1. GBD 2016 Causes of Death Collaborators. Global, regional, and national age–sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 390, 1151–1210 (2017).
2. Wahlström, C. M. & Magnusson, P. K. Genetic influences on peripheral arterial disease in a twin population. Arterioscler Thromb Vasc Biol. 31, 678–682 (2011).
3. Murabito, J. M. et al. Association between chromosome 9p21 variants and the ankle–brachial index identified by a meta-analysis of 21 genome-wide association studies. Circ Cardiovasc Genet. 5, 100–112 (2012).
4. Matsukura, M. et al. Genome-wide association study of peripheral arterial disease in a Japanese population. PLoS ONE 10, e0139262 (2015).
5. Collins, R. What makes UK Biobank special? Lancet 379, 1173–1174 (2012).
6. Gaziano, J. M. et al. Million veteran program: A mega-biobank to study genetic influences on health and disease. J. Clin. Epidemiol. 70, 214–223 (2016).
7. Fan, J. et al. Billing code algorithms to identify cases of peripheral artery disease from administrative data. J. Am. Med. Inform. Assoc. 20, e349–e354 (2013).
8. Kullo, I. J. et al. The ATXN2-SH2B3 locus is associated with peripheral arterial disease: an electronic medical record-based genome-wide association study. Front Genet. 5, 57 (2014).
9. Thorgerisson, T. E. et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 452, 638–642 (2008).
10. Denny, J. C. et al. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics 26, 1205–1210 (2010).
11. Denny, J. C. et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat. Biotechnol. 31, 1102–1110 (2013).
12. Voight, B. F. et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat. Genet. 42, 579–589 (2010).
13. Klarin, D., Emdin, C. A., Natarajan, P., Conrad, M. F. & Kathiresan, S. Genetic analysis of venous thromboembolism in UK biobank identifies the ZFPM2 locus and implicates obesity as a causal risk factor. Circ. Cardiovasc. Genet. 10, e001643 (2017).
14. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707–713 (2010).
15. de Bakker, P. P. I. et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat. Genet. 38, 1166–1172 (2006).
16. Conte, M. S. et al. Society for Vascular Surgery practice guidelines for atherosclerotic occlusive disease of the lower extremities: management of asymptomatic disease and claudication. J. Vasc. Surg. 61, 2s–41s (2015).
17. Staley, J. R. et al. PhenoScanner: a database of human genotype-phenotype associations. Bioinformatics 32, 3207–3209 (2016).
18. GTEx Consortium. Genetic effects on gene expression across human tissues. Nat. Genet. 52, 525–531 (2020).
19. Zhang, Y. et al. Genetic analysis of venous thromboembolism in UK biobank identifies the ZFPM2 locus and implicates obesity as a causal risk factor. Circ. Cardiovasc. Genet. 10, e001643 (2017).
20. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707–713 (2010).
21. de Bakker, P. P. I. et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat. Genet. 38, 1166–1172 (2006).
22. Conte, M. S. et al. Society for Vascular Surgery practice guidelines for atherosclerotic occlusive disease of the lower extremities: management of asymptomatic disease and claudication. J. Vasc. Surg. 61, 2s–41s (2015).
23. Staley, J. R. et al. PhenoScanner: a database of human genotype-phenotype associations. Bioinformatics 32, 3207–3209 (2016).
24. GTEx Consortium. Genetic effects on gene expression across human tissues. Nat. Genet. 52, 525–531 (2020).
25. Zhang, Y. et al. Genetic analysis of venous thromboembolism in UK biobank identifies the ZFPM2 locus and implicates obesity as a causal risk factor. Circ. Cardiovasc. Genet. 10, e001643 (2017).
26. Sun, B. B. et al. Genomic atlas of the human plasma proteome. Nature 558, 73–79 (2018).
27. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. 48, 245–252 (2016).
28. Anttila, V. et al. Analysis of shared heritability in common disorders of the brain. Science 360, eaap8757 (2018).
29. Anttila, V. et al. Analysis of shared heritability in common disorders of the brain. Science 360, eaap8757 (2018).
30. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. Nat. Genet. 47, 1236–1241 (2015).
31. CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat. Genet. 47, 1121–1130 (2015).
25. Malik, R. et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat. Genet. 50, 524–537 (2018).

26. Sibon, I. et al. COL4A1 mutation in Axenfeld–Rieger anomaly with leukoencephalopathy and stroke. Ann. Neurol. 62, 177–184 (2007).

27. Greengard, J. S., Eichinger, S., Griffin, J. H. & Bauer, K. A. Brief report: variability of thrombosis among homozygous siblings with resistance to activated protein C due to an Arg->Gln mutation in the gene for factor V. N. Engl. J. Med. 331, 1559–1562 (1994).

28. Bertina, R. M. et al. Mutation in blood coagulation factor V associated with resistance to activated protein C. Nature 369, 64–67 (1994).

29. Holst, A. G., Jensen, G. & Prescott, E. Risk factors for venous thromboembolism: results from the Copenhagen City Heart Study. Circulation 121, 1896–1903 (2010).

30. Cheng, Y. J. et al. Current and former smoking and risk for venous thromboembolism: a systematic review and meta-analysis. PLoS Med. 10, e1001515 (2013).

31. Willey, J. et al. Epidemiology of lower extremity peripheral artery disease in veterans. J. Vasc. Surg. 68, 527–535.e5 (2018).

32. Fowkes, F. G. et al. Comparison of global estimates of prevalence and risk factors for peripheral artery disease in 2000 and 2010: a systematic review and analysis. Lancet 382, 1329–1340 (2013).

33. Myocardial Infarction Genetics and CARDIoGRAM Exome Consortium Investigators. Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease. N. Engl. J. Med. 374, 1134–1144 (2016).

34. Musunuru, K. et al. Exome sequencing, ANGPTL3 mutations, and familial combined hyperlipidemia. N. Engl. J. Med. 363, 2220–2227 (2010).

35. Dewey, F. E. et al. Genetic and pharmacologic inactivation of ANGPTL3 and cardiovascular disease. N. Engl. J. Med. 377, 211–221 (2017).

36. Emdin, C. A. et al. Phenotypic characterization of genetically lowered human lipoprotein(a) levels. J. Am. Coll. Cardiol. 68, 2761–2772 (2016).

37. Swanberg, M. et al. MHC2TA is associated with differential MHC molecule expression and susceptibility to rheumatoid arthritis, multiple sclerosis and myocardial infarction. Nat. Genet. 37, 486–494 (2005).

38. Anand, S. S. et al. Rivaroxaban with or without aspirin in patients with stable peripheral or carotid artery disease: an international, randomised, double-blind, placebo-controlled trial. Lancet 391, 219–229 (2017).

Acknowledgements
This research is based on data from the MVP, Office of Research and Development, Veterans Health Administration and was supported by award no. MVP000. This publication does not represent the views of the Department of Veterans Affairs, the US Food and Drug Administration, or the US Government. This research was also supported by funding from: the Department of Veterans Affairs awards nos. B01-BX03340 (K.C. and P.W.F.W.), B01-BX003362 (P.S.T. and K.M.C) and IK2-CX001780 (S.M.D) and the VA Informatics and Computing Infrastructure (VINCI) VA HSR RES 1300457 (P.A., O.V.P. and S.L.D.); the National Institutes of Health grants nos. R01-HL131977 (J.A.B.), R03-AG05090 (S.A.), R01-HL138306 (J.Chen), R01-HL127564 (S.K.) and K08-HL140203 (P.N.); and the American Heart Association grant no. 18SFRN33960373 (J.A.B.) and 15MCPRP2558005 (S.M.D.). Data on CAD have been contributed by the CARDIoGRAMplusCAD investigators. Data on large artery stroke have been contributed by the MEGASTROKE investigators. The MEGASTROKE project received funding from sources specified at http://www.megastroke.org/acknowledgements.html.

Author contributions
D.K., J.L., T.L.A., K.A., M.C., M.S.F., J.A.B., J.M.G., J. Concato, D.J.R., K.C., K.-M.C., P.W.F.W., C.J.O., S.K., P.S.T. and S.M.D were responsible for the concept and design. The acquisition, analysis or interpretation of data were performed by D.K., J.L., K.A., M.C., T.L.A., J.H., K.M.L., Q.S., J.E.H., P.N., S.A., A.S., Y.V.S., M.V., M.S.F., L.W., J.Chen, D.S., J.S.L., D.R.M., P.R., P.R.A., O.V.P., S.L.D., W.E.B., J.A.B., J. Concato, J.M.G., D.J.R., K.C., K.-M.C., P.W.F.W., C.J.O., S.K., P.S.T. and S.M.D. The authors D.K., P.S.T. and S.M.D. drafted the manuscript. The critical revision of the manuscript for important intellectual content was carried out by D.K., T.L.A., Y.V.S., K.C., P.N., D.J.R., K.C., K.-M.C., C.J.O., J.A.B., M.S.F., O.V.P., P.W.W., S.K., P.S.T. and S.M.D. The authors D.K., P.N., Y.V.S., M.S.F., J.Chen, J.A.B., J. Concato, J.M.G., D.J.R., K.C., K.-M.C., C.J.O., P.W.F.W., S.K., P.S.T. and S.M.D provided administrative, technical, or material support.

Competing interests
J.A.B. reports consulting with AstraZeneca, Bristol Myers Squibb, Amgen, Merck, Sanofi, Amgen, Pfizer, and Sanofi; Antidote Pharmaceutical and Boehringer Ingelheim. He serves on the DSMC for Bayer and Novartis. S.L.D. has received research grant support from the following for-profit companies through the University of Utah or the Western Institute for Biomedical Research (VA Salt Lake City’s affiliated non-profit): AbbVie Inc., Anolix LLC, Astellas Pharma Inc., AstraZeneca Pharmaceuticals LP, Boehringer Ingelheim International GmbH, Celgene Corporation, Eli Lilly and Company, Genentech Inc., Genomic Health, Inc., Gilead Sciences Inc., GlaxoSmithKline PLC, Innocrin Pharmaceuticals Inc., Janssen Pharmaceuticals Inc., Kantar Health, Myriad Genetic Laboratories, Inc., Novartis International AG and PAREXEL International Corporation. S.M.D. receives research support to his institution from CytoVAS and RenalayxAI. S.K. is a founder of Maze Therapeutics, Verve Therapeutics and San Therapeutics. He holds equity in Catabasis and San Therapeutics. He is a member of the scientific advisory boards for Regeneron Genetics Center and Corvidia Therapeutics; served as a consultant for Acceleron, Eli Lilly, Novartis, Merck, Novo Nordisk, Novo Ventures, Ionis, Alnylam, Aegerion, Huag Partners, Noble Insights, Leerink Partners, Bayer Healthcare, Illumina, Color Genomics, MedGenome, Quest and Medscape; and reports patents related to a method of identifying and treating a person having a predisposition to or afflicted with cardiometabolic disease (20180010185) and a genetics risk predictor (20190017119). O.V.P. has received research grants from the following for-profit organizations through the University of Utah or Western Institute for Biomedical Research: Anolix LLC, AstraZeneca Pharmaceuticals LP, Genentech Inc., Genomic Health, Inc., Gilead Sciences Inc., Janssen Pharmaceuticals Inc., Novartis International AG and PAREXEL International Corporation.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0492-5. Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0492-5.

Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.M.D.

Peer review information: Michael Basson was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

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

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2019
Methods

Study populations. We conducted genetic association analyses using DNA samples and phenotypic data from two cohorts: MVP and UK Biobank. In MVP, individuals aged 19 to over 100 years were recruited from VA Medical Centers across the United States. In our initial MVP analysis, we evaluated 31,307 individuals (24,009 white, 5,373 black, 1,925 Hispanic) with PAD, and 211,753 controls free of clinical evidence of disease. For variants with suggestive associations ($P < 10^{-6}$), we sought replication of our findings in the UK Biobank (Fig. 1 and Extended Data Fig. 9). In the UK Biobank, individuals aged 45 to 69 years old were recruited from across the United Kingdom for participation. In this study, we identified 5,117 PAD cases and 389,291 controls of European ancestry. MVP received ethical/study protocol approval by the VA Central Institutional Review Board, the analysis in UK Biobank was approved by a local Institutional Review Board at Partners Healthcare (protocol 2013P001840) and informed consent was obtained for all participants. Additional information regarding experimental design and participants are provided in the Nature Research Reporting Summary.

Genetic data and quality control. DNA extracted from whole blood was genotyped in MVP using a customized Affymetrix Axiom biobank array, the MVP 1.0 Genotyping Array. Veterans (former US military personnel) of three mutually exclusive ethnic groups were identified for analysis: (1) non-Hispanic whites (European ancestry), (2) non-Hispanic blacks (African ancestry) and (3) self-identified Hispanics. Before imputation, variants that were poorly called or that deviated from their expected allele frequency based on reference data from the 1000 Genomes Project$^c$ were excluded. After pre-phasing using EAGLE.v2 (ref. 38), 1,000 Genomes and 1000 Genomes Project$^c$ phase 3, v.5 reference panel were imputed into MVP participants via Minimac3 software$^c$. Ethnicity-specific principal component analysis was performed using EIGENSOFT v6.2 software$^c$. Participants were then divided into three mutually exclusive ethnic groups based on self-identified race/ethnicity and admixture analysis using the ADMIXTURE v1.3 software$^{15}$. (1) non-Hispanic whites (self-identified as ‘non-Hispanic’, white, and >80% genetic European ancestry), (2) non-Hispanic blacks (self-identified as ‘non-Hispanic’, black, and >50% genetic African ancestry) and (3) Hispanics (self-identified only). In total, 312,571 white, black and Hispanic MVP participants passed our sample-level quality control.

In MVP, sample and variant quality control was performed as previously described$^3$. In brief, duplicate samples, samples with more heterozygosity than expected, an excess (>2.5%) of missing genotypes, or discordance between genetically inferred sex and phenotypic gender were excluded. In addition, one individual from each pair of related individuals (kinship $>0.0884$ as measured by KING$^c$v2.0 software) were removed.

Following imputation, variant level quality control was performed using the EasyQC R package$^c$ (www.R-project.org), and exclusion metrics included: ancestry-specific Hardy–Weinberg equilibrium$^c$, $P < 1 \times 10^{-6}$, posterior call probability <0.9, imputation quality <0.5 (MAF $<0.0003$), call rate <95% for common variants (MAF $>1\%$) and call rate <99% for rare variants (MAF $<1\%$). Variants within the same region also were excluded if they deviated >10% from their expected allele frequency based on reference data from the 1000 Genomes Project$^c$. Following variant level quality control, we obtained 20.3 million, 32.4 million and 31.2 million DNA sequence variants for analysis in white, black and Hispanic participants, respectively.

In the UK Biobank, analysis was performed separately in white individuals after genotyping using either the UK BILEVE or UK Biobank Axiom Arrays. Approximately 500,000 individuals were genotyped and subsequently imputed to the Haplotype Reference Consortium (HRC). Details of these procedures are described elsewhere$^c$. We performed genome-wide association testing for PAD in the UK Biobank using all variants in the HRC reference with MAF $>0.5\%$ and imputation quality INFO $>0.3$. To avoid potential population stratification, only European-ancestry samples were included in the analysis. This subset was selected based on self-reported white ethnicity that was subsequently confirmed using genetic principal components analysis. Outliers within the self-reported white samples in the first six principal components of ancestry were detected and subsequently removed using the R package aberration$^c$. In addition, individuals with sex chromosome aneuploidy (neither XX or XY), discordant self-reported sex and genetic sex, or excessive heterozygosity or missingness, as defined centrally by the UK Biobank were removed. Finally, one individual from each pair of second-degree or closer relatives (kinship $>0.0884$) was removed, selectively retaining PAD cases when possible.

PAD definitions. From the 312,571 multi-ethnic participants passing quality control in MVP, individuals were defined as having PAD based on possessing at least two of the ICD-9/10 codes/Current Procedural Terminology (CPT) codes outlined in Supplementary Table 20 in their EHR, or having one code and at least two visits to a vascular surgeon within a 14 month period$^c$. Individuals were defined as controls if they had zero diagnosis/procedure codes suggesting a diagnosis of PAD. Those listed in Supplementary Table 21 and Supplementary Table 22 in their EHR reflected two or more separate encounters in the VA Healthcare System in each of the 2 years before enrollment in MVP. Manual chart review was performed by two trained nurse chart abstractors with a vascular surgeon reviewing discordant case results; the cases of chart review for 50 cases and 50 controls otherwise representative of the overall cohort were used for determining the sensitivity and specificity of the phenotyping algorithm. In the UK Biobank, individuals were defined as having PAD based on possession of at least one of the self-reported illness codes, OPCS procedure codes or ICD codes in Supplementary Table 22 in their EHR. All other individuals were defined as controls. In both cohorts, individuals were not excluded from the PAD control group if they possessed diagnosis codes for either CAD or LAS.

Assignment of smoking status in MVP. Smoking status was derived from an algorithm that utilized diagnosis codes, medications, clinic identifier codes and smoking-related health factors from the VA EHR to classify individuals as never, former or current smokers from American Heart Association abstract A18809$^c$.

Ankle–brachial index extraction and GWAS quality control. A natural language processing algorithm was used to extract ABI data from the EHR in MVP. Resultant values were manually inspected for accuracy. In total, 261,835 ABI measurements across 17,861 individuals were available for analysis. We selected each individual’s minimum ABI (mABI) for association analysis to minimize confounding from treatment or revascularization.

We performed a GWAS of mABI in 13,382 European, 3,284 African and 998 Hispanic ancestry MVP participants after restricting to those with value $<1.4$ as previously described$^3$. Sample and variant quality controls were performed in identical manner to the MVP PAD case/control analysis, with the exception of excluding variants with MAF $<0.01$ given the smaller sample size. In total, we obtained 9.2 million, 15.6 million and 10.8 million DNA sequence variants for analysis in white, black and Hispanic participants, respectively.

PhesVar of PAD risk variants. Understanding the full spectrum of phenotypic consequences of a given DNA sequence variant may shed light on the mechanism by which a variant/gene leads to disease. For lead PAD risk variants identified in our primary analysis, we performed a PhesVar of 1,101 distinct diseases in MVP leveraging the full catalog of EHR ICD-9 diagnosis codes in 176,913 white veterans passing PhesVar quality control using the R package PhesVar$^c$ and its associated disease definitions with the exception of CAD defined as previously described$^3$. Diseases were required to have a prevalence of $>0.2\%$ (300 cases) to be included in the PhesVar analysis. Lead PAD risk DNA sequence variants were tested using logistic regression adjusting for age, sex and five principal components under the assumption of additive effects.

We supplemented our MVP PhesVar with data from PhenoScanner V2 (ref. 39), an online resource of association statistics from previously conducted GWAS and UK Biobank and used a genome-wide significant $P$ value threshold (two-sided $P < 5 \times 10^{-8}$). PhenoScanner data sources are outlined in Supplementary Table 23.

eQTl/pQTL associations and PAD transcriptome-wide association study. To identify loci that might influence gene expression, we used previously published eQTL mapping data from the Genotype-Tissue Expression (GTex) Consortium Project across 44 tissues$^18$. We queried the 19 lead PAD risk variants identified in our study for overlap with genome-wide significant variant-gene pairs from the GTEx portal. Similarly, to identify loci that might influence protein concentrations in plasma, we used published pQTL data generated from an aptamer-based multiplex protein assay to quantify 3,622 plasma proteins in 3,301 healthy participants from the INTERVALL study$^{20,21}$. We queried the 19 lead PAD risk variants identified in our study for overlap with genome-wide significant variant-protein pairs.

We then performed a TWAS using summary statistics from the European MVP PAD analysis and gene-expression reference panels of tibial artery from GTEx v7 in 388 independent samples, as previously described$^3$. In brief, for a given gene, variant–expression weights in the 1-mb cI region were first computed with the BSLMM$^c$, which “models effects on expression as a mixture of normal distributions to account for the sparse expression architecture. Given weights $w$, lipid $Z$ scores $Z$, and variant–correlation (LD) matrix $D$, the association between predicted expression and lipids (i.e., the TWAS statistic) was estimated as $Z_w=w^T_{w}Dw^T_{w}$” (ref. 3). We computed $w$ using either the variants genotyped in each expression reference panel or imputed HapMap3 variants. To account for multiple hypotheses, we applied a Bonferroni corrected two-sided $P < 6.2 \times 10^{-10}$ ($0.05/8089$ genes).

Shared heritability within PAD, CAD and LAS. To better understand how common genetic variation influences risk for atherosclerosis in multiple vascular beds, we used linkage disequilibrium score regression$^c$ to calculate the genetic correlation between PAD–CAD and PAD–LAS. Summary statistics from the European MVP PAD GWAS, the trans-ethnic LAS MEGASTROKE GWAS meta-analysis (greater than two-thirds European)$^{25}$ and the transancestral LAS MEGASTROKE GWAS meta-analysis were used for this analysis. Of note, we used the transancestral meta-analysis statistics from MEGASTROKE because the sample size of the European-ancestry only analysis lacked sufficient power for estimation of genetic correlation.

PAD associations independent of CAD and LAS. DNA sequence variants might differ in contribution to risk for atherosclerosis in the peripheral, coronary and
cerebrovascular beds. For 19 lead PAD risk variants identified in our primary analysis, we first tested their effect on CAD and LAS in white MVP participants and combined results with summary statistics from the CARDIoGRAMplusC4D and MEGASTROKE consortia. We performed a sensitivity analysis for variants demonstrating at least a nominally significant association with either CAD or LAS, by re-testing their association with PAD after including CAD or LAS status as a covariate in the association model. We also re-tested their association including both CAD and LAS as covariates in a single model. Variants associated with both CAD and LAS individually, and that remained associated with PAD after adjustment for CAD/LAS suggest the presence of a common mechanism or pathway leading to the development of atherosclerosis in multiple arterial beds. Conversely, associations that are present uniquely with PAD suggest a mechanism specific to the peripheral vascular bed.

MVP CAD definition and analysis. CAD was defined based on ICD-9/10 and CPT codes using the method described by Dewey and colleagues. CAD cases were defined as individuals who, based on ICD-9, ICD-10 and CPT codes had an inpatient admission with a primary diagnosis of CAD, a combination of CAD associated inpatient or outpatient encounters on two or more distinct days noted in the longitudinal VA EHR or as for-service data, or a coronary revascularization at the time of analysis (Supplementary Table 24). We identified 50,415 CAD cases and 124,577 controls available for analysis among 174,992 white MVP participants. Genotyped and imputed DNA sequence variants were tested for association with CAD through logistic regression adjusting for age, sex and five principal components of ancestry. Results were then combined with publicly available summary data of CAD case patients identified in the CARDIoGRAMplusC4D consortium study6 using an inverse-variance weighted fixed-effects method. For variants with a high amount of heterogeneity across the two ethnic groups (I² statistic >75%, for example, rs4842266), results were combined using a random effects method.

MVP LAS adjusted analysis and definition. LAS was defined based on the groupings proposed by Denny et al. and the phaco 433.11, occlusion of the cerebral arteries with cerebral infarction, which is defined as the occurrence of the any of following ICD-9-CM codes on two or more distinct dates: 433.01, 433.11, 433.21, 433.31, 433.81, 433.91. We identified 705 LAS cases and 174,287 controls available for analysis among 174,992 white MVP participants. Genotyped and imputed DNA sequence variants were tested for association with LAS through logistic regression adjusting for age, sex and five principal components of ancestry. Results were then combined with publicly available summary data of the transancestral LAS MEGASTROKE meta-analysis6 of 6,688 LAS cases and 454,450 controls using an inverse-variance weighted fixed-effects method.

T2D definition for TCF7L2 adjusted analysis. To better understand how the TCF7L2 locus affects PAD risk, rs7903146 was re-tested for association with PAD after adjusting for T2D status in the 174,992 white MVP participants. T2D was defined based on the groupings proposed by Denny et al.7, which identified 78,431 MVP participants with T2D (58,603 males and 5,273 black females). We then tested the association of rs7903146 in MVP with T2D through logistic regression adjusting for age, sex and five principal components of ancestry separately in whites and blacks. We then re-tested for association with PAD through logistic regression adjusting for age, sex, T2D and five principal components of ancestry.

We report logistic regression two-sided P values.

Factor V Leiden genotypes and risk of vascular disease. One of the variants most strongly associated with PAD in the discovery analysis was the Factor V Leiden mutation, the most common cause of inherited thrombophilia. The variant’s protein-altering consequence (F5 p.R506Q) results in a resistance to protein C by activated protein C and a hypercoaguable state. We sought to better understand Factor V Leiden’s relationship with atherosclerosis by testing its association with CAD, LAS and increasingly severe presentations of PAD. Individuals were defined as having claudication, rest pain, tissue loss, or major amputation if they met our EHR-based definition for PAD and possessed at least one diagnosis code depicted in Supplementary Table 25. If an individual possessed diagnosis codes for more than one severe PAD presentation (for example, claudication and rest pain), they were placed in the most severe PAD classification was selected. We then evaluated for evidence of an interaction between smoking and F5 p.R506Q carrier status.

Statistical analysis. In our primary analysis, genotyped and imputed DNA sequence variants were tested for association with PAD using logistic regression adjusting for age, sex and five principal components of ancestry using an additive model using the SNPTEST v2.5.4 (mathgen.stats.ox.ac.uk/genetics/software/snptest/snptest.html) statistical software program. In our MVP discovery analysis, we performed association analyses separately for each ethnic group, and then meta-analyzed using an inverse-variance weighted fixed-effects method. A P < 5.0 × 10⁻⁸ was used to declare genome-wide significance for the continuous mABI trait.

For variants with suggestive PAD associations (P < 10⁻⁶) we sought replication of our findings in UK Biobank. Association testing was performed in 5,117 PAD cases and 389,291 controls using a logistic regression model adjusted for age at baseline, sex, genotyping array and the first 10 principal components of ancestry. All testing was performed in PLINK v2 (https://www.cog-genomics.org/plink/2.0/). We defined significant novel PAD associations as those that were at least nominally significant in replication (P < 0.05), were directionally consistent in both cohorts, and had an overall P < 5 × 10⁻⁸ (genome-wide significance) in the discovery and replication cohorts combined. Novel loci were defined as being greater than 500,000 base pairs away from a known PAD genome-wide associated lead variant. Additionally, linkage disequilibrium information from the 1000 Genomes Project was used to determine independent variants where a locus extended beyond 500,000 base pairs. All logistic regression values of P were two-sided.

In our Phase analysis, DNA sequence variants were tested using logistic regression using an untransformed mABI value adjusting for age, sex and five principal components of ancestry. We performed association analyses separately for each ethnic group and then meta-analyzed the results using an inverse-variance weighted fixed-effects method. A P < 5.0 × 10⁻⁸ was used to declare genome-wide significance for the continuous mABI trait.

In our PheWAS analysis, DNA sequence variants were tested using logistic regression adjusting for sex, and five principal components of ancestry against disease-free controls and declared to be significantly associated with the disease if they met a P < 5.0 × 10⁻⁸. In our CAD/LAS sensitivity analysis, risk variants identified in the primary analysis were tested for association with CAD/LAS in white MVP participants and combined with either (1) publicly available summary data of 60,801 CAD case patients and 123,504 disease-free controls in the CARDIoGRAMplusC4D consortium study9 or (2) with the transancestral LAS MEGASTROKE meta-analysis of 6,688 LAS cases and 454,450 controls using an inverse-variance weighted fixed-effects method. Variants demonstrating at least nominally significant (P < 0.05) significance with CAD/LAS were then re-tested with PAD after adjusting for CAD or LAS status in MVP. Variants were declared to be still associated with PAD if they demonstrated a reduction in association signal when adjusting for CAD/LAS status but Pₐₚ₂ remained <0.05. All logistic regression values of P were two-sided.

In our Factor V Leiden analysis, the Leiden variant was tested for association with each subtype of PAD (intermittent claudication, rest pain, tissue loss, major amputation), as compared to PAD free controls, through logistic regression adjusting for sex, and five principal components of ancestry. Lastly, we evaluated for evidence of a Factor V Leiden-smoking interaction by stratifying MVP participants into current smokers and former/never smokers and performed a Cochran’s Q test for interaction. In our Factor V Leiden analysis, we set a Bonferroni-adjusted level of significance of P = 0.05 / 7 tests = 0.007. All values of P in the Factor V Leiden analysis were two-sided.

To determine the specificity and sensitivity values of our PAD phenotype, we performed a manual chart review and calculated the resultant values using the R v.3.2.0 software (Supplementary Table 26). Sensitivity refers to the ratio of (true positives)/(true positives + false negatives) and specificity refers to the ratio of (true negatives)/(true negatives + false positives).

Natural language processing algorithm for ABI. ABI values measured for participants in the VA Healthcare System are not recorded in a structured format in the EHR. Instead, the values can be found in clinical reports in narrative or semi-structured format (Extended Data Fig. 10). In order to make these ABI values available for the PAD phenotype definition, we developed a natural language processing system to identify instances of ABI values recorded within clinical notes. The system was developed in several stages and the results of an initial iteration of the system development were reported previously. To develop a rule-based natural language processing system that could scale to process the 6.5 million documents associated with the 31,307 patients in the discovery analysis cohort, we utilized the Leo framework, which builds on the Unstructured Information Management Architecture – Asynchronous Scaleout. The system achieved 4% precision as valued on 1,000 manually labeled clinical notes. A sensitivity analysis showed 89.8% recall on an instance level across 200 documents selected from the same day as a PAD diagnosis code.

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

Data availability. The full summary level association data from the MVP transancestry PAD meta-analysis from this report are available through dbGAP, accession code phs001672.v2.p1. Additional data that support the findings of this study are available on request from the corresponding author (S.M.D.); these data are not publicly available due to US Government and Department of Veteran’s Affairs restrictions relating to participant privacy and consent. Data contributed by CARDIoGRAMplusC4D investigators are available online (http://cardiogramplusc4d.org/). Data on LAD have been contributed by the MEGASTROKE investigators and are available online (http://www.megastroke.org/).
The genetic and phenotypic UK Biobank data are available upon application to the UK Biobank. Source data has been provided for Figs. 2 and 3 and Extended Data Figs. 4, 5 and 7. Additional data that were used to generate the figures in this study are available on request from the corresponding author (S.M.D.) or through dbGAP as listed above.

Code availability
Code to perform analyses in this manuscript are available from the authors upon request (D.K. and S.M.D.), or from the URLs associated with each software in the Methods section.

References
39. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 526, 68–74 (2015).
40. Loh, P. R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. 48, 811–816 (2016).
41. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44, 955–959 (2012).
42. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, 904–909 (2006).
43. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664 (2009).
44. Klarin, D. et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat. Genet. 50, 1514–1523 (2018).
45. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. Bioinformatics 26, 2867–2873 (2010).
46. Winkler, T. W. et al. Quality control and conduct of genome-wide association meta-analyses. Nat. Protoc. 9, 1192–1212 (2014).
47. Hyde, C. L. et al. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. Nat. Genet. 48, 1031–1036 (2016).
48. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209 (2018).
49. Bellenguez, C., Strange, A., Freeman, C., Donnelly, P. & Spencer, C. C. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. Bioinformatics 28, 134–135 (2012).
50. Arya, S. et al. Race and socioeconomic status independently affect risk of major amputation in peripheral artery disease. J. Am. Heart Assoc. 7, e007425 (2018).
51. Song, R. J. et al. Abstract 18809: Development of an electronic health record-based algorithm for smoking status using the Million Veteran Program (MVP) cohort survey response. Circulation 134 A18809 (2016).
52. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics 30, 2375–2376 (2014).
53. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 9, e1003264 (2013).
54. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190–2191 (2010).
55. Alba, P. R. et al. Ankle brachial index extraction system. In AMIA Annu. Symp. Proc. https://symposium2018.zerista.com/event/member?item_id=8296486 2018.
56. Cornia, R., Patterson, O. V., Ginter, T. & DuVall, S. L. Rapid NLP development with Leo. In AMIA Annual Symposium Proceedings 1356 (2014).
57. Ferrucci, D. & Lally, A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat. Lang. Eng. 10, 327–348 (2004).
Extended Data Fig. 1 | Distribution of minimum ankle-brachial index values in the MVP. Histogram of minimum mABI values extracted from the electronic health record for 17,861 participants of the MVP. These values, restricted to those with an minimum ABI of <1.4, were used for the subsequent mABI GWAS.
Extended Data Fig. 2 | Quantile-quantile plot for the discovery trans-ethnic PAD GWAS in MVP. The expected logistic regression association $P$ values versus the observed distribution of $P$ values for PAD association ($N$ = 31,307 PAD cases and 211,753 controls) are displayed. Quantile-quantile plots were inspected for ancestry-specific analyses and genomic control values were <1.20 for each racial group (data not shown). No systemic inflation was observed ($\lambda_{GC} = 1.05$). All $P$ values were two-sided.
Extended Data Fig. 3 | Manhattan plot for the PAD GWAS. Plot of \(-\log_{10}(P)\) for association of imputed variants by chromosomal position for all autosomal polymorphisms analyzed in the PAD GWAS (N = 36,424 PAD cases and 601,044 controls). The genes nearest to the top associated variants are displayed. Genes highlighted in red represent novel PAD loci (18). Genes for variants that are outside the transcript boundary of a protein-coding gene are shown with nearest candidate gene in parentheses (for example, (LDLR)). Logistic regression two-sided \(P\) values are displayed.
Extended Data Fig. 4 | TCF7L2 mediates its effect on PAD via T2D. a, Forest plot depicting the replication of the known TCF7L2/rs7903146-T2D association signal in MVP for both white and black participants. b, The same variant is also associated with PAD risk in whites and blacks in MVP. However, when controlling for T2D status in the regression model, the association signal is dramatically reduced, suggesting that TCF7L2 PAD risk is mediated through its effect on T2D. Logistic regression two-sided values of $P$ are displayed. Gray boxes reflect the inverse-variance weight for each ancestry.
Extended Data Fig. 5 | Forest plot for association of the CHRNA3 locus and peripheral artery disease risk stratified by smoking status. When stratifying European MVP participants by smoking status (ever smokers versus never smokers), nearly all the association signal resides within the ever smoker group. Previous reports of variation at the CHRNA3 locus demonstrate that carriers of the PAD risk allele have a reduced likelihood of cigarette smoking cessation. This suggests that the PAD-CHRNA3 association is driven by a greater burden of tobacco exposure in those who carry the nicotine dependence/PAD risk allele. Logistic regression two-sided values of $P$ are displayed. Gray boxes reflect the inverse-variance weight for each subgroup.
Extended Data Fig. 6 | Peripheral artery disease risk loci and known causal risk factors. Peripheral artery disease risk loci identified in this GWAS analysis are depicted along with the plausible relationship to the underlying causal risk factor. Loci names are based on the nearest genes; however, the causal gene(s) remains unclear for some associated loci and, as such, the resultant annotation may prove incorrect in some cases.
Extended Data Fig. 7 | Peripheral artery disease risk variants and association with LAS and CAD. For the 19 PAD risk variants identified in our study, logistic regression $Z$-scores of association (aligned to the PAD risk allele) were obtained from MVP (PAD, $N = 31,307$ PAD cases and $211,753$ controls) and publicly available summary statistics for large artery stroke (MVP+MEGASTROKE consortium$^{35}$, $N = 7,393$ LAS cases and $628,737$ controls) and coronary artery disease (MVP+CARDioGRAMplusC4D consortium$^{24}$, $N = 111,216$ CAD cases and $248,081$ controls). A positive $Z$-score (red) indicates a positive association between the PAD risk allele and the disease, while a negative $Z$-score (blue) indicates an inverse association. Boxes are outlined in cyan if the variant is uniquely associated with PAD (two-sided logistic regression $P_{\text{PAD}} < 5 \times 10^{-8}$, $P_{\text{CAD}}$ and $P_{\text{LAS}} > 0.05$).
Extended Data Fig. 8 | Peripheral artery disease risk variants and mechanistic overlap with LAS and CAD. Venn diagram of each of the 19 PAD risk loci in a based on their association with PAD (N = 31,307 PAD cases and 211,753 controls; two-sided $P_{\text{FDR}} < 5 \times 10^{-8}$), CAD (N = 111,216 CAD cases and 248,081 controls; two-sided $P < 0.05$) and LAS (N = 7,393 LAS cases and 628,737 controls; two-sided $P < 0.05$) using logistic regression. Each locus is depicted along with the plausible relationship to the underlying causal risk factor separately by color. Loci names are based on the nearest genes; however, the causal gene(s) remains unclear for some associated loci and as such, the resultant annotation may prove incorrect in some cases.
Extended Data Fig. 9 | Overall study design. The primary analysis consisted of a genome-wide association study to identify novel PAD risk variants. Secondary analyses involved a genome-wide association study of minimum ABI, a closer examination the 19 PAD risk variants through PheWAS, a candidate causal gene analysis using eQTL/pQTL/TWAS data, a PAD analysis accounting for CAD/LAS status and a focused Factor V Leiden analysis.
Extended Data Fig. 10 | Natural language processing for index extraction. Examples of semi-structured text that contains targeted indices for extraction using natural language processing (NLP). TBI, toe–brachial index; PT, posterior tibial artery; AT, anterior tibial artery.

| Right  | Left  |
|--------|-------|
| Brachial 160 180 | Toe Brachial TBI |
| PT 80 74 | Right 86 114 .72 |
| AT 100 46 | Left 100 118 .84 |
| Toe 52 20 | |
| ABI 0.63 0.46 | |
| TBI 0.33 0.13 | |
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted. Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated
- Clearly defined error bars
- State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

Phenotypic data was collected from the electronic health record and genetic data using the Million Veteran Program (MVP) Axiom array.

Data analysis

Imputation was performed using MiniMac3/EAGLE v2, and data was collected and cleaned using the EasyQC package, SNPTESTv2.5.4, EIGENSOFT v6, METAL (released 2011), and KING 2.0 software programs as outlined in the online methods. Clear code for analysis is available at their associated website (see text). Additional analyses were performed in R-3.2.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The full summary level association data from the MVP trans-ancestry PAD meta-analysis from this report are available through dbGAP, accession code
Additional data that support the findings of this study are available on request from the corresponding author SMD; these data are not publicly available due to U.S. Government and Department of Veteran’s Affairs restrictions relating to participant privacy and consent. Data contributed by CARDioGRAMplusC4D investigators are available online (http://www.CARDIOGRAMPLUSC4D.org/). Data on large artery stroke have been contributed by the MEGASTROKE investigators and are available online (http://www.megastroke.org/). The genetic and phenotypic UK Biobank data are available upon application to the UK Biobank.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

All samples available of three ancestries (European, African, Hispanic) were used for analysis (after quality control, see Supplementary Table 1 for full details). Sample size was determined based on using all genetic data available from MVP/UK Biobank. Participants were excluded if they failed to meet case or control definitions.

Data exclusions

Data were excluded if they did not pass our pre-established quality control metrics, or if they did not fall within the three main ancestries used for analysis.

Replication

Replication was performed using data from UK Biobank of 25 genome-wide significant loci identified in MVP. Of the 6 signals that did not replicate, 2 were rare variants that were not available in UK Biobank following quality control (European MAF < 0.005), and the remaining four did not meet the pre-specified P < 0.05 for independent replication (see Supplementary Table 6).

Randomization

Randomization is not applicable, as this is a population based case-control analysis of prevalent data.

Blinding

Blinding is not applicable, as this is a population based case-control analysis of prevalent data.

Reporting for specific materials, systems and methods

Materials & experimental systems

- n/a Involved in the study
- Unique biological materials
- Antibodies
- Eukaryotic cell lines
- Palaeontology
- Animals and other organisms
- Human research participants

Methods

- n/a Involved in the study
- ChIP-seq
- Flow cytometry
- MRI-based neuroimaging

Human research participants

Demographics and participant counts for the European, African, and Hispanic MVP participants [European - Mean Age PAD cases = 74.4 years, 97.5% male, Mean Age PAD controls = 66.9 years, 91.9% male; African - Mean Age PAD cases = 69.6 years, 96.1% male, Mean Age PAD controls = 60.4 years, 85.2% male; Hispanic - Mean Age PAD cases = 71.6 years, 97.9% male, Mean Age PAD controls = 59.0 years, 89.7% male] that passed our quality control and were included in the analysis are depicted in Supplementary Table 1 and Supplementary Table 7.

Recruitment

Individuals aged 19 to 104 years have been recruited voluntarily from more than 50 VA Medical Centers nationwide for participation in the Million Veteran Program biobank study. Recruitment is currently occurring in person at selected sites in the VHA health care system. Every Veteran is assigned a study ID number, which is used to track them throughout the entire process of recruitment, enrollment, sample collection and use; this approach also provides a level of protection for personal identifiers from the outset. Given that study enrollment is voluntary, biases of this study are similar to those of any mega-biobank with voluntary enrollment, including survivorship bias. A complete description of the entire MVP Biobank study including recruitment can be found at PMID: 26441289.
In UK Biobank, individuals aged 45 to 69 years old were recruited from across the United Kingdom for participation. Given that study enrollment is voluntary, biases of this study are similar to those of any mega-biobank with voluntary enrollment, including survivorship bias. A complete description of the entire UK Biobank study including recruitment can be found at PMID: 30305743.