A plasma proteomic signature for atherosclerotic cardiovascular disease risk prediction in the UK Biobank cohort

Trisha P. Gupte[1], Zahra Azizi[1], Pik Fang Kho[1], Jiayan Zhou[1], Ming-Li Chen[1], Daniel J. Panyard[2], Rodrigo Guarischi-Sousa[1,3], Austin T. Hilliard[1,3], Disha Sharma[1], Kathleen Watson[4], Fahim Abbasi[1,5], Philip S. Tsao[1,5], Shoa L. Clarke[1,5], Themistocles L. Assimes[1,5,6]

1 Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA, USA

2 Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

3 Palo Alto Veterans Institute for Research (PAVIR), Stanford, CA, USA

4 Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, USA

5 Stanford Cardiovascular Institute, Stanford University School of Medicine, Stanford, CA, USA

6 Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA, USA

Corresponding author:

Themistocles L. Assimes, MD PhD

Palo Alto VA Hospital, 3801 Miranda Avenue, Palo Alto, CA, 94304

Email: tassimes@stanford.edu

Abstract

Background: While risk stratification for atherosclerotic cardiovascular disease (ASCVD) is essential for primary prevention, current clinical risk algorithms demonstrate variability and leave room for further improvement. The plasma proteome holds promise as a future diagnostic and prognostic tool that can accurately reflect complex human traits and disease processes. We assessed the ability of plasma proteins to predict ASCVD.

Method: Clinical, genetic, and high-throughput plasma proteomic data were analyzed for association with ASCVD in a cohort of 41,650 UK Biobank participants. Selected features for analysis included clinical variables such as a UK-based cardiovascular clinical risk score (QRISK3) and lipid levels, 36 polygenic risk scores (PRSs), and Olink protein expression data of 2,920 proteins. We used least absolute shrinkage and selection operator (LASSO) regression to select features and compared area under the curve (AUC) statistics between data types. Randomized LASSO regression with a stability selection algorithm identified a smaller set of more robustly associated proteins. The benefit of plasma proteins over standard clinical variables, the QRISK3 score, and PRSs was evaluated through the derivation of Δ AUC values. We also assessed the incremental gain in model performance using proteomic datasets with varying numbers of proteins. To identify potential causal proteins for ASCVD, we conducted a two-sample Mendelian randomization (MR) analysis.

Result: The mean age of our cohort was 56.0 years, 60.3% were female, and 9.8% developed incident ASCVD over a median follow-up of 6.9 years. A protein-only LASSO model selected 294 proteins and returned an AUC of 0.723 (95% CI 0.708-0.737). A clinical variable and PRS-only LASSO model selected 4 clinical variables and 20 PRSs and achieved an AUC of 0.726 (95% CI 0.712-0.741). The addition of the full proteomic dataset to clinical variables and PRSs resulted in a Δ AUC of 0.010 (95% CI 0.003-0.018). Fifteen proteins selected by a stability selection algorithm offered improvement in ASCVD prediction over the QRISK3 risk score [Δ AUC: 0.013 (95% CI 0.005-0.021)]. Filtered and clustered versions of the full proteomic dataset (consisting of 600-1,500 proteins) performed comparably to the full dataset for ASCVD prediction. Using MR, we identified 11 proteins as potentially causal for ASCVD.

Conclusion: A plasma proteomic signature performs well for incident ASCVD prediction but only modestly improves prediction over clinical and genetic factors. Further studies are warranted to better elucidate the clinical utility of this signature in predicting the risk of ASCVD over the standard practice of using the QRISK3 score.

Introduction

Atherosclerotic cardiovascular disease (ASCVD) is one of the leading causes of morbidity and mortality worldwide. A substantial contributor to the high burden of disease includes suboptimal risk stratification coupled with the inefficient application of established primary prevention strategies[1-3]. While ASCVD outcomes have improved in recent decades, clinical risk algorithms have historically demonstrated variability and leave room for further improvement[4]. Among ASCVD clinical risk prediction models currently in use, the UK-based QRISK3 cardiovascular clinical risk score has notably demonstrated excellent performance in its validation cohort and in the UK Biobank (UKB)[5,6]. Unlike other models, the QRISK3 score incorporates significant past medical history and current medication use in addition to standard clinical variables.

In the last decade, high-throughput profiling of circulating plasma proteins has emerged as a powerful tool for both predicting and understanding the underlying biology of complex human traits[7]. By capturing dynamic changes in protein expression, proteomic profiling is well-advantaged to reflect interactions between individuals' genetics, environment, lifestyle, and more. Further, the incorporation of proteomic profiling data to traditional cardiovascular risk factors has also been shown to enhance prediction of cardiometabolic disease[8-10]. However, the ability of proteins to augment existing clinical risk prediction models as robust as the QRISK3 score has yet to be tested.

We aimed to evaluate the predictive value of clinical, genetic, and proteomic factors for ASCVD in the UKB. We hypothesized that the incorporation of proteomic data would provide modest improvement in prediction accuracy beyond that offered by the QRISK3 score, other clinical variables, and polygenic risk scores (PRSs).

Methods

Study population

We analyzed a cohort of UKB participants with normalized protein expression (NPX) data and excluded individuals with prevalent ASCVD, history of statin use, or in whom the QRISK3 score could not be computed. The study design of the UKB has been previously described extensively[11]. At participants' baseline visits, trained healthcare providers conducted verbal interviews and administered questionnaires to obtain information on past medical history, family history, lifestyle, and sociodemographic and psychosocial factors. Physical measures along with the collection of blood, urine, and saliva samples were also obtained. By integrating participants' electronic health record (EHR) data, health outcomes data including outpatient and inpatient International Classification of Disease, Tenth Revision (ICD-10) codes are available within the database. The UKB received ethical approval from the Northwest Multicenter Research Ethics Committee and obtained informed consent from all participants at the time of recruitment.

Measurement of protein biomarkers

The UKB conducted proteomic profiling in a random sampling of UKB participants with plasma samples collected at baseline visits. Using the antibody-based Proximity Extension Assay (PEA) by Olink, the UKB measured NPX data for a total of 2,923 proteins. The sample handling, processing, and quality control protocols implemented by the UKB have been previously described in a summary document (biobank.ndph.ox.ac.uk/ukb/ukb/docs/PPP_Phase_1_QC_dataset_companion_doc.pdf) and in

two publications[12,13]. We identified GLIPR1, NPM1, and PCOLCE as proteins with a high degree of missingness (> 50%) and excluded these from analysis. The remaining missing NPX values were imputed with their mean values. All NPX values were provided as log-transformed by the UKB and standardized by us prior to analysis.

Measurement of the outcome

In accordance with the QRISK3 score, we defined ASCVD as the composite outcome of either transient ischemic attack (TIA), ischemic stroke, or coronary heart disease. The incidence of ASCVD was recorded using UKB's first occurrences data (UKB categories 2401-2417), which is organized by ICD-10 codes and was generated by mapping primary care data (UKB category 3000), hospital inpatient data (UKB category 2000), death register records (UKB fields 40001 and 40002), and self-reported medical condition codes (UKB field 20002). The ICD-10 codes used to record ASCVD can be found in the **Supplementary File**.

Measurement of clinical variables

Our primary clinical variable was the QRISK3 score, which incorporates several established clinical predictors to provide an individual's 10-year predicted risk of developing ASCVD[5]. We computed the score using the QRISK3 function in the QRISK3 R package. We further considered additional clinical predictors measured at baseline including hemoglobin A1c (HbA1c), lipoprotein (a) (Lp(a)), LDL cholesterol, triglyceride levels, prior alcohol use, and physical activity status. Lastly, we considered 36 standard polygenic risk scores (PRSs) calculated with genome wide data on DNA sequence variation, summarizing the genetic predisposition or liability to 36 health traits or disease conditions[14,15]. All clinical variables and PRSs were standardized prior to running analyses. A full list of variables included in our analyses can be found in the **Supplementary File**.

Statistical analyses

A flowchart of the study design and analysis plan is shown in **Figure 1**. We used least absolute shrinkage and selection operator (LASSO) regression to evaluate the relative ability of clinical variables, PRSs, and proteins to associate with incident ASCVD. We randomly divided the cohort into a training set (70%) and test set (30%) before building five LASSO models using 10-fold cross validation. These five LASSO models included a model with clinical variables alone, a model with PRSs alone, a model with proteins alone, a combined model with clinical variables and PRSs, and a combined model in which proteins were added to clinical variables and PRSs. We compared area under the curve (AUC) statistics to assess the relative ability of each model to predict ASCVD. To ascertain the incremental value offered by proteins beyond clinical variables and PRSs, we calculated Δ AUC values and generated a corresponding 95% confidence interval by bootstrapping 1,000 samples.

We next utilized the R package stabs to run the randomized LASSO stability selection (RLSS) algorithm, which was initially presented by Meinhausen and Bühlmann and later refined by Shah and Samsworth[16,17]. This algorithm was applied in the training set to develop a more robust proteomic signature (PS) for ASCVD prediction. We used default parameters when applying this algorithm, which included a weakness value of 0.8, a cutoff value of 0.8, and a per-family error rate of two. To assess the predictive value of this PS, we calculated the AUC of a LASSO model incorporating this smaller set of proteins in the test set. We subsequently calculated a Δ AUC

value with a 95% confidence interval to determine the added predictive benefit of this PS in ASCVD prediction beyond that offered by QRISK3.

To further explore correlation structure and the incremental improvement in model performance with varying sizes of proteomic datasets, we used two methods to reduce the number of proteomic predictors. First, we implemented a filtering-based approach and computed a correlation matrix of all 2,920 proteins to identify pairs of proteins with a correlation value > 0.3 and > 0.5. We randomly removed one protein from each pair to form two smaller datasets of about 600 and 1,500 proteins, respectively. Second, we used a clustering-based approach and applied principal component analysis (PCA) and K-means clustering to form 600 and 1,500 clusters of proteins. From each cluster, we randomly selected a protein to form two additional smaller datasets of 600 and 1,500 proteins, respectively. Through standard LASSO regression, we evaluated ASCVD prediction performance of all four proteomic datasets by generating AUCs.

Finally, we used two-sample Mendelian randomization (MR) analysis to identify potential causal proteins for ASCVD. In a cohort of 15,016 UKB participants, we performed genome-wide association studies (GWASs) for all 2,920 proteins. Effect estimates for ASCVD were obtained by performing a GWAS analysis of ASCVD (defined as either a TIA, ischemic stroke, or coronary heart disease) in a cohort of UKB participants of European ancestry who did not have proteomics data available (n_{cases} = 64,386 & $n_{controls}$ = 299,887). A full description of the methods used for both GWAS analyses can be found in the **Supplementary File.**

The inverse variance weighted method (IVW) served as our primary approach for the MR analysis. In any MR analysis, there are three assumptions which should be satisfied: 1) the genetic variants used as instrumental variables should be associated with the outcome of interest, 2) the genetic instruments should not be associated with other confounder variables, and 3) the genetic instruments should affect the outcome only via the exposure of interest rather than through alternative pathways. While MR analyses utilizing *cis*-genetic instruments have typically been found to satisfy these assumptions, we conducted additional analyses to address these assumptions[18]. To address the first assumption, we used established methods to calculate the proportion of variance explained and *F* statistic (with equations provided in the **Supplementary File**). Various sensitivity analyses were also conducted, including the MR-Egger method, which we used to calculate the MR-Egger intercept and assess for pleiotropy. All analyses were conducted using the TwoSampleMR package in R.

Results

Cohort characteristics

We analyzed NPX data of 2,920 proteins in a total of 41,650 participants. Baseline characteristics for the cohort are shown in **Table 1**. The mean age at recruitment was 56.0 years (SD, 8.2 years), 60.3% were female, and 93.1% were of self-reported white ethnicity. Over a median follow-up of 6.9 years, 9.8% developed incident ASCVD.

Standard LASSO regression

Consistent with prior reports, the QRISK3 score performed very well in predicting ASCVD with an AUC of 0.720 (95% CI 0.706-0.734) (**Fig. 2**). When additional clinical variables were added to QRISK3, the AUC marginally increased at 0.724 (95% CI 0.710-0.738). A LASSO model built on PRSs alone did not perform well with an AUC of 0.575 (95% CI 0.558-0.591) while a LASSO model incorporating the full proteomic dataset on its own performed comparably to the clinical variable-only model with an AUC of 0.723 (95% CI 0.717-0.745). In our combined models, we observed that the addition of PRSs to clinical variables resulted in a slightly higher AUC point estimate of 0.731 (95% CI 0.717-0.745). Incorporating the full proteomic dataset in addition to clinical variables and PRSs modestly improved the AUC to 0.741 (95% CI 0.727-0.755) and resulted in a Δ AUC of 0.010 (95% CI 0.003-0.018). A full list of the clinical variables, PRSs, and proteins selected by these LASSO models can be found in **Supp. Table 1**.

Stability selection with randomized LASSO regression

A randomized LASSO stability selection (RLSS) analysis selected 15 proteins, which spanned a wide range of known functions (**Fig. 3**). These stability selection proteins offered most of the improvement in ASCVD prediction explained by the full proteomic dataset with an AUC of 0.711 (95% CI 0.696-0.726). This robust proteomic signature also offered modest improvement in ASCVD prediction over that provided by the QRISK3 score [Δ AUC: 0.013 (95% CI 0.005-0.021)].

Clustering and filtering analyses

We assessed and compared the prediction performance of smaller proteomic datasets formed through filtering-based or clustering-based approaches to that of the full proteomic dataset provided by the UKB [AUC of 0.723 (95% CI 0.708-0.737)] (**Supp. Fig. 1**). A smaller dataset (consisting of 600 proteins) formed through a filtering-based approach with a correlation threshold of > 0.3 performed the worst with an AUC of 0.691 (95% CI 0.675-0.706). Despite consisting of approximately the same number of proteins (645 vs. 600), a smaller dataset formed through a clustering-based approach with the creation of 600 clusters performed better with an AUC of 0.705 (95% CI 0.690-0.720). Finally, datasets formed through a filtering-based approach with a correlation threshold of > 0.5 and through a clustering-based approach with the creation of 1,500 clusters performed comparably to the full proteomic dataset with AUCs of 0.719 (95% CI 0.704-0.734) and 0.717 (95% CI 0.702-0.732), respectively.

Two-sample Mendelian randomization analysis

To identify potentially causal proteins for ASCVD, we conducted a two-sample Mendelian randomization (MR) analysis. Of 2,920 proteins, we identified genome-wide significant *cis*-protein quantitative loci (*cis*-pQTLs) for 1,745 based on a significance threshold of 5×10^{-8}. The minimum *F* statistic of our genetic instruments was 27.7 (**Supp. Table 2**). Effect estimates for ASCVD were obtained by running a GWAS in 364,273 UKB participants of European ancestry using REGENIE (**Supp. Fig. 2a-b**). We identified 11 proteins as having a potentially causal effect on ASCVD based on a Bonferroni-corrected threshold (p-value = 4.31×10^{-5}) (**Fig. 4**). For several proteins in which the initial number of associated SNPs was low (nSNPs < 3), we were not able to obtain results for sensitivity analyses. For proteins with a higher number of initial associated SNPs, however, we found that results from the IVW method generally aligned with results from other sensitivity analyses (**Supp. Fig. 4)**. Full results for the two-sample MR

analysis including annotations of whether each protein tested was selected by a standard LASSO model or the RLSS algorithm can be found in **Supp. Table 2**.

Discussion

We evaluated the relative ability of clinical variables, PRSs, and proteins to predict ASCVD by utilizing high-throughput proteomic profiling data provided by the UKB. In this study, we investigated whether the incorporation of proteomic data enhanced prediction offered by an existing clinical risk prediction model, as well as other clinical and genetic factors. We highlight three primary sets of findings from our results.

First, we found that a protein-only model derived by standard LASSO regression performed comparably to the QRISK3 score, which performed exceptionally well on its own. When additional clinical variables and PRSs were combined with QRISK3, ASCVD prediction was comparable to performance offered by QRISK3 alone. The further addition of proteins to clinical variables and PRSs resulted in a modest improvement in prediction. Unlike most ASCVD risk prediction models, the UK-based QRISK3 cardiovascular clinical risk score captures a significant portion of an individual's past medical history and current medication use, which explains its strong performance in UK-based populations such as its validation cohort and in the UKB[5,6]. In the setting of our baseline model already exhibiting strong performance, our observation of only modest improvements in ASCVD prediction with the addition of other clinical variables and multi-omic datasets is not surprising[19]. However, when practitioners may not have access to all the clinical information needed to compute a score generated by QRISK3, the comparable performance of our protein-only model suggests that proteins could potentially serve as an alternative tool for risk stratification purposes in the future.

Second, we show that a substantially smaller set of proteins selected by a stability selection algorithm accounted for most of the prediction performance offered by the full dataset. In addition to aiding in feature reduction, this algorithm also identifies more "stable" features that may be more transportable to other populations. Recently, others created a custom quantitative PEA panel measuring up to 21 proteins called the CVD-21 tool[20]. With this tool in mind, a future where a robust proteomic signature such as the one we describe is generated for risk prediction through absolute quantification is now foreseeable. Third, smaller proteomic datasets, created by filtering and clustering methods to reduce high degrees of correlation in the full dataset, do not meaningfully change the performance of ASCVD prediction. When looking to future implementations of plasma proteomic profiling, our findings suggest that the additional costs associated with measuring more than ~1,500 proteins may be avoided without drastically affecting prediction performance.

Several proteins were repeatedly selected by standard LASSO models incorporating proteins alone, as well as proteins in addition to clinical variables and PRSs. In particular, GDF15, or growth/differentiation factor 15, has previously been associated with a host of diseases within the cardiometabolic spectrum[21,22]. With known functions in the suppression of food intake and inflammation, GDF15 is now an appealing drug target in the management of obesity, T2DM, and CVD[23,24]. LTBP2 also carried a large beta coefficient in our protein-only LASSO model. While this protein has not previously been associated with ASCVD, others have demonstrated its

potential as a novel biomarker of heart failure and other diseases[25,26]. Interestingly, we also identified dementia-related proteins such as BCAN and NEFL as predictive of ASCVD as well[7,27].

In addition to risk prediction, high-throughput proteomic profiling can further our understanding of the underlying biology of disease and aid in identifying novel drug targets. From our two-sample MR analysis, we corroborated several previously known causal targets of ASCVD including PCSK9, LPA, and IL6R[28-31]. Additionally, we identified FN1 as potentially causal, which has been demonstrated to enhance endothelial inflammation in atherosclerotic plaques[32]. Lastly, we identified CELSR2 as potentially causal for ASCVD. With known functions in serum lipid and cholesterol metabolism, variants in the *CELSR2* gene and its neighboring genes, *PSCR1* and *SORT1* were first identified in GWASs of cardiovascular disease over fifteen years ago[33].

The well-documented reproducibility and stability of the Olink platform is a key strength of our study[34]. By selecting QRISK3, a robust cardiovascular clinical risk score, as our baseline model, we also contribute valuable insights on whether proteins can augment existing prediction models for ASCVD. Our study has some notable limitations including the lack of genetic diversity within the UKB. Our two-sample MR analysis was restricted to participants of European ancestry only. To avoid the further perpetuation of health disparities in biomedical research, future studies in more diverse populations are needed. Additionally, while previous studies have also highlighted the potential of plasma proteins to enhance cardiometabolic health prediction, we acknowledge potential limitations in relying on protein measurements in plasma rather than using protein measurements from human tissue[9] .

In summary, our findings suggest that plasma proteomic profiling modestly enhances prediction of ASCVD beyond an already well-performing cardiovascular clinical risk score, QRISK3, as well as other routinely available clinical variables and PRSs. Further investigations in more diverse study populations are needed to better understand the potential benefits multi-omics data could provide for ASCVD prediction. We also show that utilizing more robust proteomic datasets does not appreciably affect prediction performance when compared to the full proteomic dataset provided by the UKB. Finally, we contribute a list of potentially causal proteins for ASCVD using expanded plasma proteomic from the UKB.

Acknowledgement:
The UKB received ethical approval from the Northwest Multicenter Research Ethics Committee and obtained informed consent from all participants at the time of recruitment. This study was conducted under UK Biobank application number 52374.

Funding and Assistance
This study was supported by a grant from the National Institutes of Health 1R01DK114183. TPG was supported by the Sarnoff Cardiovascular Research Foundation Fellowship.

Conflict of Interest
None of the authors have conflicts of interest to report

Author Contributions

TPG and TLA conceived and designed the study. TPG, ZA, PFK, JZ, and carried out the analyses. TPG and TLA drafted the manuscript. TPG, ZA, PFK, JZ, MLC, DJP, RGS, ATH, DS, KW, FA, PST, SLC, and TLA verified the underlying data. TLA is responsible for the integrity of the work as a whole. All authors acquired and interpreted the data, critically revised the paper and had final responsibility for the decision to submit for publication.

Protein and Gene Abbreviations

- GDF15: Growth/differentiation factor 15
- LTBP2: Latent-transforming growth factor beta-binding protein 2
- BCAN: Brevican core protein
- NEFL: Neurofilament light polypeptide
- PCSK9: Proprotein convertase subtilisin/kexin type 9
- LPA: Apolipoprotein(a)
- IL6R: Interleukin-6 receptor subunit alpha
- FN1: Fibronectin
- CELSR2: Cadherin EGF LAG seven-pass G-type receptor 2
- *PSCR1*: Proline/serine-rich coiled-coil 1
- *SORT1*: Sortilin

References

1. Johnson NB, Hayes LD, Brown K, Hoo EC, Ethier KA. CDC National Health Report: leading causes of morbidity and mortality and associated behavioral risk and protective factors--United States, 2005-2013. *MMWR Suppl*. 2014;63:3-27.
2. Weir HK, Anderson RN, Coleman King SM, Soman A, Thompson TD, Hong Y, Moller B, Leadbetter S. Heart Disease and Cancer Deaths - Trends and Projections in the United States, 1969-2020. *Prev Chronic Dis*. 2016;13:E157. doi: 10.5888/pcd13.160211
3. Mensah GA, Fuster V, Murray CJL, Roth GA, Mensah GA, Abate YH, Abbasian M, Abd-Allah F, Abdollahi A, Abdollahi M, et al. Global Burden of Cardiovascular Diseases and Risks, 1990-2022. *Journal of the American College of Cardiology*. 2023;82:2350-2473. doi: doi:10.1016/j.jacc.2023.11.007
4. Kaasenbrood L, Boekholdt SM, Van Der Graaf Y, Ray KK, Peters RJ, Kastelein JJ, Amarenco P, LaRosa JC, Cramer MJ, Westerink J. Distribution of estimated 10-year risk of recurrent vascular events and residual risk in a secondary prevention population. *Circulation*. 2016;134:1419-1429.
5. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017;357:j2099. doi: 10.1136/bmj.j2099
6. Tamlander M, Mars N, Pirinen M, Palotie A, Daly M, Riley-Gills B, Jacob H, Paul D, Runz H, John S, et al. Integration of questionnaire-based risk factors improves polygenic risk scores for human coronary heart disease and type 2 diabetes. *Communications Biology*. 2022;5:158. doi: 10.1038/s42003-021-02996-0
7. You J, Guo Y, Zhang Y, Kang J-J, Wang L-B, Feng J-F, Cheng W, Yu J-T. Plasma proteomic profiles predict individual future health risk. *Nature Communications*. 2023;14:7817. doi: 10.1038/s41467-023-43575-7
8. Nowak C, Carlsson AC, Östgren CJ, Nyström FH, Alam M, Feldreich T, Sundström J, Carrero J-J, Leppert J, Hedberg P, et al. Multiplex proteomics for prediction of major

cardiovascular events in type 2 diabetes. *Diabetologia*. 2018;61:1748-1757. doi: 10.1007/s00125-018-4641-z

9. Hoogeveen RM, Pereira JPB, Nurmohamed NS, Zampoleri V, Bom MJ, Baragetti A, Boekholdt SM, Knaapen P, Khaw K-T, Wareham NJ, et al. Improved cardiovascular risk prediction using targeted plasma proteomics in primary prevention. *European Heart Journal*. 2020;41:3998-4007. doi: 10.1093/eurheartj/ehaa648
10. Zanini JC, Pietzner M, Langenberg C. Integrating genetics and the plasma proteome to predict the risk of type 2 diabetes. *Current Diabetes Reports*. 2020;20:1-11.
11. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12:e1001779. doi: 10.1371/journal.pmed.1001779
12. Elliott P, Peakman TC, Biobank oboU. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *International Journal of Epidemiology*. 2008;37:234-244. doi: 10.1093/ije/dym276
13. Sun BB, Chiou J, Traylor M, Benner C, Hsu Y-H, Richardson TG, Surendran P, Mahajan A, Robins C, Vasquez-Grinnell SG, et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature*. 2023;622:329-338. doi: 10.1038/s41586-023-06592-6
14. Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars N, Benner C, Aguirre M, Venkataraman GR, Wainberg M, Ollila HM, Kiiskinen T, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nature Genetics*. 2021;53:185-194. doi: 10.1038/s41588-020-00757-z
15. Thompson D, Wells D, Selzam S, Peneva I, Moore R, Sharp K, Tarran W, Beard E, Riveros-Mckay F, Giner-Delgado C, et al. UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. In: medRxiv; 2022.
16. Meinshausen N, Bühlmann P. Stability Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 2010;72:417-473. doi: 10.1111/j.1467-9868.2010.00740.x
17. Shah RD, Samworth RJ. Variable Selection with Error Control: Another Look at Stability Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 2012;75:55-80. doi: 10.1111/j.1467-9868.2011.01034.x
18. Zheng J, Haberland V, Baird D, Walker V, Haycock PC, Hurle MR, Gutteridge A, Erola P, Liu Y, Luo S, et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat Genet*. 2020;52:1122-1131. doi: 10.1038/s41588-020-0682-6
19. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem*. 2008;54:17-23. doi: 10.1373/clinchem.2007.096529
20. Siegbahn A, Eriksson N, Assarsson E, Lundberg M, Ballagi A, Held C, Stewart RA, White HD, Åberg M, Wallentin L. Development and validation of a quantitative Proximity Extension Assay instrument with 21 proteins associated with cardiovascular risk (CVD-21). *Plos one*. 2023;18:e0293465.
21. Echouffo-Tcheugui JB, Daya N, Matsushita K, Wang D, Ndumele CE, Al Rifai M, Hoogeveen RC, Ballantyne CM, Selvin E. Growth Differentiation Factor (GDF)-15 and Cardiometabolic Outcomes among Older Adults: The Atherosclerosis Risk in

Communities Study. *Clinical Chemistry*. 2021;67:653-661. doi: 10.1093/clinchem/hvaa332

22. May BM, Pimentel M, Zimerman LI, Rohde LE. GDF-15 as a Biomarker in Cardiovascular Disease. *Arquivos Brasileiros de Cardiologia*. 2021;116:494-500.
23. Adela R, Banerjee SK. GDF-15 as a target and biomarker for diabetes and cardiovascular diseases: a translational prospective. *Journal of diabetes research*. 2015;2015:490842.
24. Wang D, Day EA, Townsend LK, Djordjevic D, Jørgensen SB, Steinberg GR. GDF15: emerging biology and therapeutic applications for obesity and cardiometabolic disease. *Nature Reviews Endocrinology*. 2021;17:592-607. doi: 10.1038/s41574-021-00529-7
25. Bai Y, Zhang P, Zhang X, Huang J, Hu S, Wei Y. LTBP-2 acts as a novel marker in human heart failure – a preliminary study. *Biomarkers*. 2012;17:407-415. doi: 10.3109/1354750X.2012.677860
26. Boucherat O, Yokokawa T, Krishna V, Kalyana-Sundaram S, Martineau S, Breuils-Bonnet S, Azhar N, Bonilla F, Gutstein D, Potus F. Identification of LTBP-2 as a plasma biomarker for right ventricular dysfunction in human pulmonary arterial hypertension. *Nature cardiovascular research*. 2022;1:748-760.
27. Guo Y, You J, Zhang Y, Liu W-S, Huang Y-Y, Zhang Y-R, Zhang W, Dong Q, Feng J-F, Cheng W, et al. Plasma proteomic profiles predict future dementia in healthy adults. *Nature Aging*. 2024;4:247-260. doi: 10.1038/s43587-023-00565-0
28. Saleheen D, Haycock PC, Zhao W, Rasheed A, Taleb A, Imran A, Abbas S, Majeed F, Akhtar S, Qamar N, et al. Apolipoprotein(a) isoform size, lipoprotein(a) concentration, and coronary artery disease: a mendelian randomisation analysis. *Lancet Diabetes Endocrinol*. 2017;5:524-533. doi: 10.1016/s2213-8587(17)30088-8
29. Rosa M, Chignon A, Li Z, Boulanger M-C, Arsenault BJ, Bossé Y, Thériault S, Mathieu P. A Mendelian randomization study of IL6 signaling in cardiovascular diseases, immune-related disorders and longevity. *NPJ genomic medicine*. 2019;4:23.
30. Arsenault BJ. From the garden to the clinic: how Mendelian randomization is shaping up atherosclerotic cardiovascular disease prevention strategies *European Heart Journal*. 2022;43:4447-4449. doi: 10.1093/eurheartj/ehac394
31. Reyes-Soffer G, Ginsberg HN, Berglund L, Duell PB, Heffron SP, Kamstrup PR, Lloyd-Jones DM, Marcovina SM, Yeang C, Koschinsky ML. Lipoprotein (a): a genetically determined, causal, and prevalent risk factor for atherosclerotic cardiovascular disease: a scientific statement from the American Heart Association. *Arteriosclerosis, Thrombosis, and vascular biology*. 2022;42:e48-e60.
32. Al-Yafeai Z, Yurdagul Jr A, Peretik JM, Alfaidi M, Murphy PA, Orr AW. Endothelial FN (Fibronectin) deposition by α5β1 integrins drives atherogenic inflammation. *Arteriosclerosis, thrombosis, and vascular biology*. 2018;38:2601-2614.
33. Castillo-Avila RG, González-Castro TB, Tovilla-Zárate CA, Martínez-Magaña JJ, López-Narváez ML, Juárez-Rojop IE, Arias-Vázquez PI, Borgonio-Cuadra VM, Pérez-Hernández N, Rodríguez-Pérez JM. Association between Genetic Variants of CELSR2-PSRC1-SORT1 and Cardiovascular Diseases: A Systematic Review and Meta-Analysis. *Journal of Cardiovascular Development and Disease*. 2023;10:91.
34. Haslam DE, Li J, Dillon ST, Gu X, Cao Y, Zeleznik OA, Sasamoto N, Zhang X, Eliassen AH, Liang L, et al. Stability and reproducibility of proteomic profiles in epidemiological studies: comparing the Olink and SOMAscan platforms. *Proteomics*. 2022;22:e2100170. doi: 10.1002/pmic.202100170

Tables and Figures

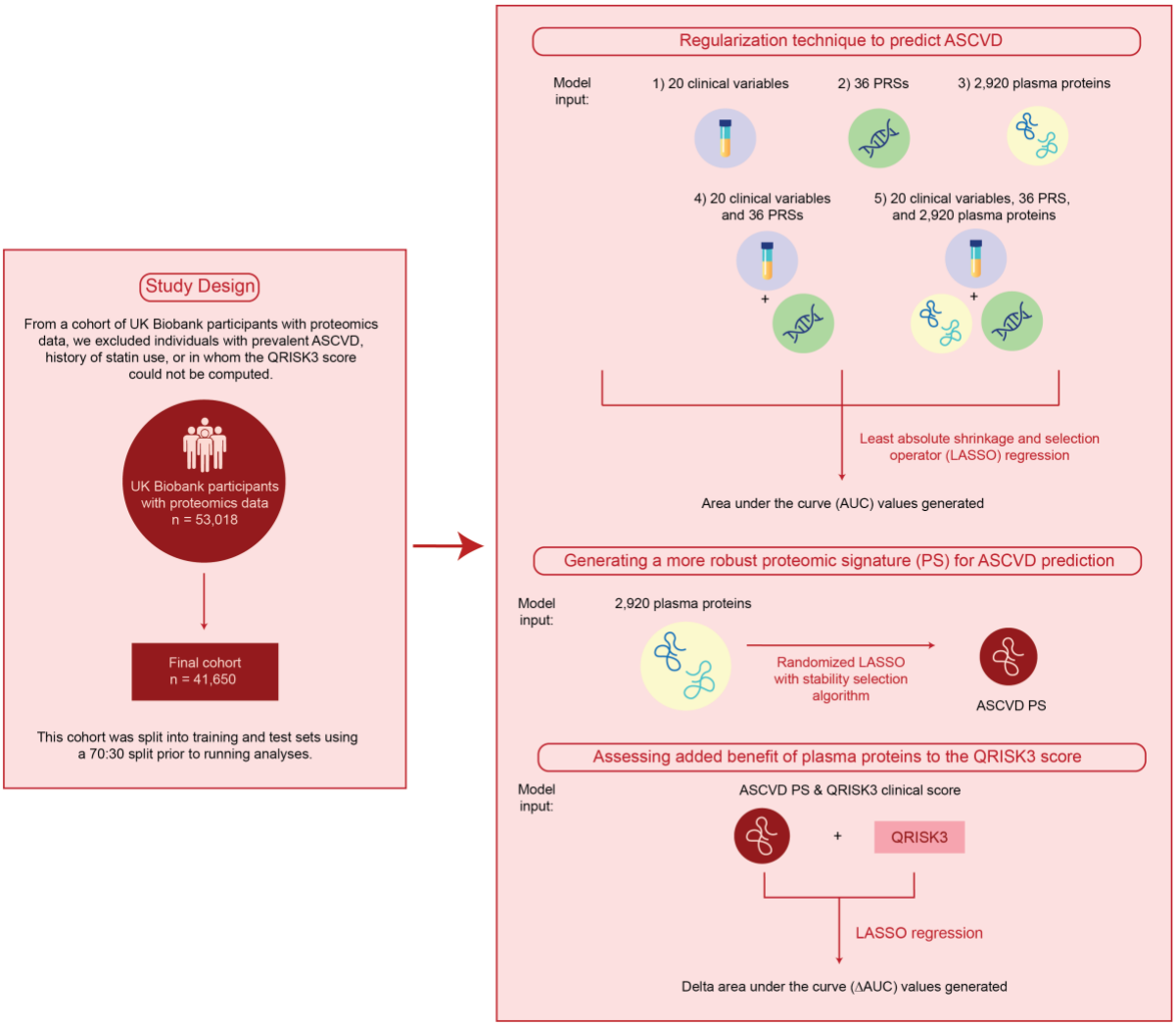
Table 1. Demographics and clinical characteristics of the study population

	n (%)
Female (%)	25,106 (60.3)
Age at baseline visit	56.0 (8.2)
Self-reported ethnicity group (%)	
White	38,758 (93.1)
African	599 (1.4)
South Asian	592 (1.4)
Chinese	123 (0.2)
Mixed	304 (0.7)
Other	1,221 (2.9)
BMI (kg/m2)	27.1 (4.7)
SBP (mmHg)	137.1 (18.7)
DBP (mmHg)	82.1 (10.1)
LDL (mmol/mol)	3.7 (0.8)
HDL (mmol/mol)	1.5 (0.4)
TG (mmol/mol, median [IQR])	1.4 [0.4, 2.5]
Cholesterol (mmol/mol)	5.8 (1.1)
HbA1c (mmol/mol)	35.6 (5.9)
Blood pressure lowering medication (%)	2,100 (5.0)
Past medical history of HTN (%)	9,056 (21.7)
Family history of heart disease (%)	19,442 (46.7)
Ever smoked status (%)	24,207 (58.1)
Self-reported alcohol intake (%)	
Never or missing	3,583 (8.6)
One to three times a month or special occasions	9,553 (22.9)
Once or twice a week	10,857 (26.1)
Three or four times a week	9,414 (22.6)
Daily or almost daily	8,243 (19.8)
Physical activity category (%)	
Missing	9,564 (23.0)
Low	5,861 (14.1)

Moderate	16,200 (38.9)
High	10,012 (24.0)
Incident ASCVD (%)	
Any ASCVD	4,086 (9.8)
TIA	456 (1.1)
Ischemic stroke	666 (1.6)
Coronary heart disease	3,234 (7.8)
Follow-up time in years	6.9 (5.2)

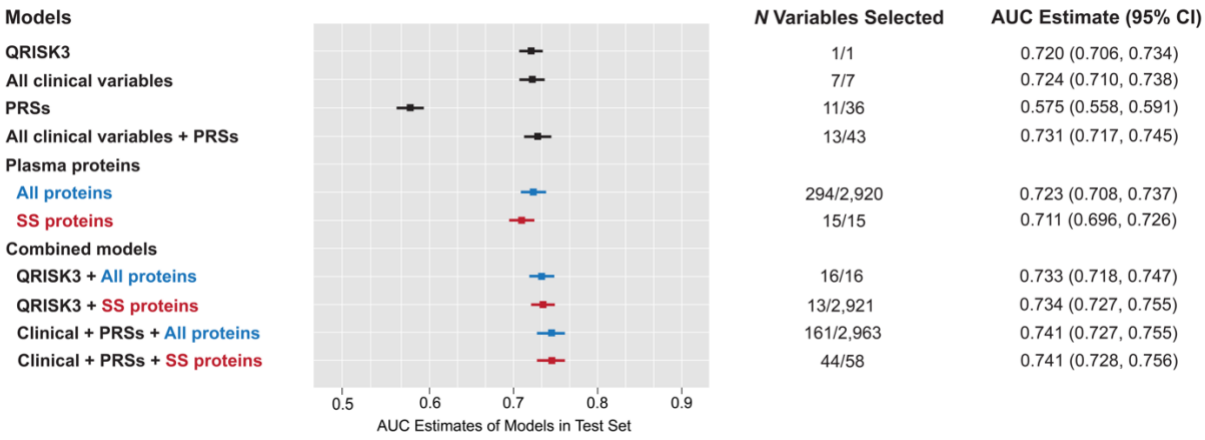
All continuous measurements were documented in mean (SD) unless otherwise specified.
Abbreviations: IQR: interquartile range, BMI: body mass index, SBP: systolic blood pressure, DBP: diastolic blood pressure, LDL: low-density lipoprotein, HDL: high-density lipoprotein, TG: triglyceride, HbA1c: hemoglobin A1c, HTN: hypertension, ASCVD: atherosclerotic cardiovascular disease, TIA: transient ischemic attack

Figure 1. Study design and analysis workflow



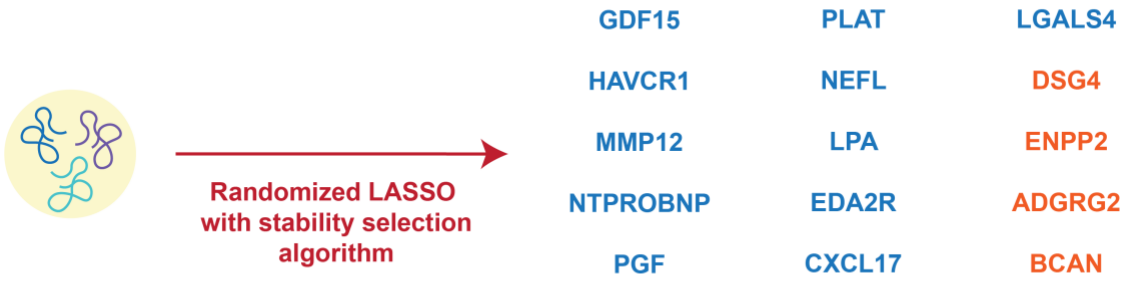
Abbreviations: ASCVD: atherosclerotic cardiovascular disease, PRSs: polygenic risk scores

Figure 2. Area under the curves (AUCs) of clinical variables, polygenic risk scores, and plasma proteins



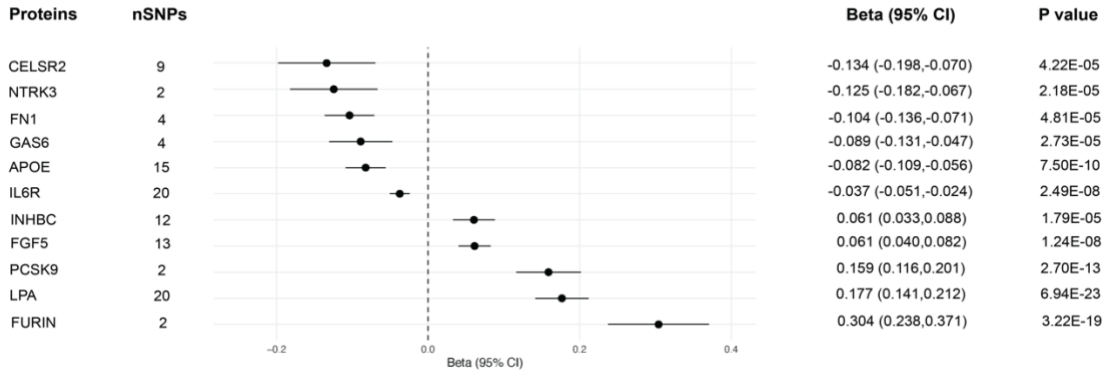
Footnote: Models performed using training and test sets in the study cohort. "All proteins" in blue represent the full proteomic dataset. "SS proteins" in red refer to proteins selected by a randomized LASSO regression model with stability selection algorithm. **Abbreviations:** LASSO: least absolute shrinkage and selection operator, AUC: area under the curve, PRSs: polygenic risk scores, SS: stability selection

Figure 3. Plasma proteins identified by a stability selection algorithm with randomized LASSO regression



Footnote: Proteins listed in blue were positively associated with atherosclerotic cardiovascular disease and those listed in orange were negatively associated. **Abbreviations:** LASSO: least absolute shrinkage and selection operator

Figure 4. Potentially causal proteins for atherosclerotic cardiovascular disease identified in a two-sample Mendelian randomization analysis



Footnote: Forest plot of potentially causal proteins for atherosclerotic cardiovascular disease based on a Bonferroni-corrected threshold (p-value = 4.31×10^{-5}). **Abbreviations:** SNP: single nucleotide polymorphism, 95% CI: 95% confidence interval