

1        **Large-scale plasma proteomics in the UK Biobank modestly**  
2                    **improves prediction of major cardiovascular events in a**  
3                    **population without previous cardiovascular disease**

4        **Patrick Royer<sup>1,2,3</sup>, Elias Björnson<sup>1</sup>, Martin Adiels<sup>1,4</sup>, Rebecca Josefson<sup>1</sup>, Eva Hagberg<sup>1,2</sup>, Anders**  
5        **Gummesson<sup>1,5</sup>, Göran Bergström<sup>\*1,2</sup>**

6        1 Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy,  
7        Gothenburg University, Gothenburg, Sweden.

8        2 Region Västra Götaland, Sahlgrenska University Hospital, Department of Clinical Physiology,  
9        Gothenburg, Sweden.

10       3 Department of Critical Care, University Hospital of Martinique, Fort-de-France, Martinique,  
11       French West Indies, France.

12       4 School of Public Health and Community Medicine, Institute of Medicine, University of  
13       Gothenburg, Gothenburg, Sweden.

14       5 Region Västra Götaland, Sahlgrenska University Hospital, Department of Clinical Genetics and  
15       Genomics, Gothenburg, Sweden.

16  
17       \*To whom correspondence should be addressed: Prof Göran Bergström, Department of  
18       Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, Gothenburg  
19       University, Gothenburg, Sweden.

20       Phone: +46 70 509 4405. Email: [goran.bergstrom@hjl.gu.se](mailto:goran.bergstrom@hjl.gu.se)

21

1 **Abstract**

2 **Background and Aims:** Improved identification of individuals at high risk of developing  
3 cardiovascular disease would enable targeted interventions and potentially lead to reductions  
4 in mortality and morbidity. Our aim was to determine whether use of large-scale proteomics  
5 improves prediction of cardiovascular events beyond traditional risk factors (TRFs).

6 **Methods:** Using proximity extension assays, 2919 plasma proteins were measured in 38 380  
7 participants of the UK Biobank. Both data- and hypothesis-driven feature selection and trained  
8 models using extreme gradient boosting machine learning were used to predict risk of major  
9 cardiovascular events (MACE: fatal and non-fatal myocardial infarction, stroke and coronary  
10 artery revascularisation) during a 10-year follow-up. Area under the curve (AUC) and net  
11 reclassification index (NRI) were used to evaluate the additive value of selected protein panels  
12 to MACE prediction by Systematic COronary Risk Evaluation 2 (SCORE2) or the 10 TRFs used in  
13 SCORE2.

14 **Results:** SCORE2 and SCORE2 refitted to UK Biobank data predicted MACE with AUCs of 0.740  
15 and 0.749, respectively. Data-driven selection identified 114 proteins of greatest relevance for  
16 prediction. Prediction of MACE was not improved by using these proteins alone (AUC of 0.758)  
17 but was significantly improved by combining these proteins with SCORE2 or the 10 TRFs  
18 (AUC=0.771,  $p<0.001$ , NRI=0.140, and AUC=0.767,  $p=0.03$ , NRI 0.053, respectively). Hypothesis-  
19 driven protein selection (113 proteins from five previous studies) also improved risk prediction  
20 beyond TRFs while a random selection of 114 proteins did not.

21 **Conclusions:** Large-scale plasma proteomics with data- and hypothesis-driven protein selection  
22 modestly improves prediction of future MACE beyond TRFs.

1 **Keywords**

2 Plasma proteomics, myocardial infarction, stroke, prediction, clinical risk factors, machine

3 learning, UK Biobank.

4

5

6

7

8

9

10

11

12

## 1 Introduction

2 Cardiovascular disease (CVD) is the main global cause of death.<sup>1</sup> CVD-related morbidity and  
3 mortality would likely be reduced if intense primary prevention efforts were focused on the  
4 group of people at highest risk. Currently, high-risk individuals are identified by estimating their  
5 10-year risk of major cardiovascular events (MACE) using traditional risk factors (TRFs)  
6 ensembled into one of many risk scores.<sup>2,3</sup> However, current risk scores lack precision for both  
7 the individual and timing of the event,<sup>4,5</sup> and there is an intense search for novel biomarkers  
8 that can help to improve the scores. The recent development of large-scale, targeted  
9 proteomics offers an unprecedented opportunity to test whether novel biomarkers for MACE  
10 can be found in the plasma proteome.<sup>6</sup>

11 Protein sets derived from large-scale proteomics have shown superior prediction when added  
12 to clinical risk scores for secondary prevention of MACE in several studies.<sup>7-10</sup> The results are  
13 less clear for primary prevention of MACE. In studies using plasma protein panels derived from  
14 aptamer-based affinity reagents, prediction of incident cardiovascular events was equivalent<sup>11</sup>  
15 or modestly superior<sup>12</sup> to a traditional risk score. However, in a study that used a protein panel  
16 derived from paired, nucleotide-labelled antibody probes, prediction of events was clearly  
17 superior compared with a model built on TRFs.<sup>13</sup>

18 In the current study, we tested whether sets of plasma proteins measured using paired  
19 antibody probes could predict incident MACE beyond TRFs in a population (38 000 participants  
20 from the UK Biobank) without a history of previous cardiovascular disease. Sets of plasma  
21 proteins were selected from 2919 measured proteins using both data- and hypothesis-driven  
22 techniques.

## 1 **Methods**

### 2 *Study population*

3 The study population comprised participants from the UK Biobank who were randomly selected  
4 for plasma proteomic analysis at the baseline visit.<sup>14</sup> Individuals with earlier MACE and with  
5 more than 20% missing protein measurements were excluded. We used ICD-9 and ICD-10  
6 (International Classification of Diseases 10th revision) diagnostic codes and OPCS-4 (OPCS  
7 Classification of Interventions and Procedures version 4) when identifying prevalent disease  
8 before the base-line examination (Table S1). UK Biobank has approval from the North-West  
9 Multi-centre Research Ethics Committee (MREC) as a Research Tissue Bank (RTB). The current  
10 study was also approved by the Swedish ethical review authority (2021-04030).

11

### 12 *Definition of outcome data*

13 MACE was defined as fatal or non-fatal cardiovascular events (myocardial infarction,  
14 revascularization procedure, ischaemic stroke, and intracerebral haemorrhage) during the 10-  
15 year follow-up after inclusion in the study. We used ICD-10 diagnostic codes and OPCS-4 for  
16 outcome data (Table S1).

17

### 18 *Traditional risk factors*

19 Baseline characteristics were collected at the baseline examination as previously described.<sup>15</sup>  
20 The 10 TRFs used in the models are: age, sex, systolic blood pressure, current smoking status,  
21 diabetes status and age at diagnosis, glycated haemoglobin, estimated glomerular filtration  
22 rate, HDL, and total cholesterol.

1

2 *Proteomic analyses*

3 Detailed information on the proteomics technology as well as the normalization and quality  
4 control steps has already been published.<sup>14, 16</sup> In brief, 2941 plasma protein analytes  
5 corresponding to 2923 unique proteins were measured using the antibody-based Olink Explore  
6 3072 proximity extension assay (PEA) technology. Protein measurements were expressed as  
7 normalized protein expression (NPX), a Log<sub>2</sub> scale arbitrary unit. In the present study, proteins  
8 with more than 20% missing NPX values across samples were excluded and missing values for  
9 protein and clinical variables were imputed using the K-nearest neighbours (KNN) algorithm.<sup>17</sup>

10

11 *Data analyses and statistics*

12 Extreme gradient boosting machine learning models<sup>18</sup> were trained with grid search and 5-fold  
13 cross-validation to predict the 10-year risk of MACE using different subsets and combinations of  
14 clinical and protein data. The data set was randomly divided into a training set (80%) and a test  
15 set (20%). Two clinical risk prediction models were chosen as reference: (1) SCORE2 (Systematic  
16 COronary Risk Evaluation 2;<sup>19</sup> SCORE2-Diabetes for individuals with diabetes<sup>20</sup>) was used to  
17 calculate the 10-year fatal and non-fatal cardiovascular disease risk for each participant; and (2)  
18 a refitted risk score was trained on UK Biobank data using the same TRFs as those included in  
19 SCORE2 (termed refitted SCORE2).

20

21 Four protein models were tested: a “complete” protein model including all proteins; a  
22 “hypothesis-driven” protein model based on proteins found to be predictive of MACE in

1 previous studies;<sup>11-13, 21, 22</sup> a “data-driven” protein model obtained after a protein feature  
2 selection procedure on the training set using the Boruta algorithm;<sup>23</sup> and a corresponding  
3 “random” protein model consisting of random proteins whose number was equivalent to that  
4 of the data-driven model. Eight combined models were formed by combining the four protein  
5 datasets with either the calculated SCORE2 (one variable) or the TRFs included in SCORE2 (10  
6 variables). The workflow is described in Figure S1.

7  
8 The performance of the models was assessed using the area under the receiver operating curve  
9 (AUC) and the categorical net reclassification index (NRI) using a 5 and 10% risk threshold.<sup>24</sup>  
10 AUC for the SCORE2 and refitted SCORE2 were compared with AUC for the protein and  
11 combined models using DeLong’s test.<sup>25</sup> The calibration of the models was evaluated by  
12 plotting reliability curves.

13  
14 Statistical and machine learning analyses were performed using R version 4.0.4 (R Foundation  
15 for Statistical Computing, Vienna, Austria). Baseline characteristics between participants who  
16 experienced or did not experience a MACE during the 10-year follow-up were compared using  
17 mean, standard deviation and t-test for continuous variables, and proportion and Chi-square  
18 test for categorical variables. A two-tailed P value of <0.05 was considered statistically  
19 significant.

20

## 1 **Results**

### 2 *Characteristics of study population*

3 Flowchart of inclusion is presented in Figure S1. In total, 46 799 randomly selected participants  
4 from the UK Biobank (total n=502 414) had available data on proteomics and 45 666 of these  
5 had no history of MACE at the baseline visit. After excluding individuals and proteins with more  
6 than 20% missing protein measurements, the final cohort comprised 38 380 individuals with  
7 2919 unique proteins measured. Baseline characteristics of the participants divided by MACE  
8 are presented in Table 1 without imputations. The pattern of missing data is presented in Table  
9 S2. Individuals who experienced a MACE during the 10-year follow-up (n=1661, 4.3%) had a  
10 much more severe cardiovascular risk factor profile at baseline compared to those who did not  
11 experience events (Table 1).

12

### 13 *Selection of proteins for prediction models*

14 Missing NPX values represented 2.4% of the total 112 031 220 protein data points and were  
15 imputed with the KNN algorithm. Workflow for protein selection is described in Figure S1. In  
16 the data-driven protein set, 114 proteins were selected by the Boruta algorithm as being of  
17 relevance for MACE prediction after 276 iterations on the training set (n=30 704). In the  
18 hypothesis-driven protein set, 113 proteins were compiled from 115 candidate biomarker  
19 proteins identified in five previous publications;<sup>11-13, 21, 22</sup> the remaining two candidate  
20 biomarkers identified in these publications (AGP1 and TREM1) had not been analysed in the  
21 current UK Biobank sample. The random protein set was created with the same number of  
22 proteins (114) as in the data-driven protein set. The complete protein set included all 2919



1 proteins. The data-driven and hypothesis-driven protein sets had 20 proteins in common while  
2 the random protein set had two and four proteins in common with the data-driven and  
3 hypothesis-driven protein sets respectively. A list of all proteins measured, and the different  
4 protein sets is found in Table S3.

5

### 6 *Performance of prediction models*

7 The predictive performance of the clinical, protein and combined models was evaluated on the  
8 test set (n=7676) and AUC results are presented in Table 2. SCORE2 and refitted SCORE2  
9 predicted MACE with an AUC of 0.740 and 0.749 respectively. The complete, data-driven, and  
10 hypothesis-driven protein models all had numerically higher AUC (0.773, 0.758, 0.759,  
11 respectively) but the difference was only significant for the complete protein model (p=0.003  
12 and 0.014 compared to SCORE2 and refitted SCORE2, respectively). The random protein model  
13 performed significantly worse than the two clinical models (AUC=0.712).

14

15 When the proteins selected in the data-driven model were combined with the 10 TRFs used in  
16 SCORE2, they significantly outperformed predictions compared to both SCORE2 ( $\Delta$ AUC =  
17 +0.029, p<0.001) and refitted SCORE2 ( $\Delta$ AUC = +0.016, p=0.031). When the proteins selected in  
18 the hypothesis-driven model were combined with TRFs, they numerically increased the AUC  
19 compared to refitted SCORE2 ( $\Delta$ AUC = +0.018, p=0.057). The hypothesis-driven protein set  
20 significantly increased AUC when combined with SCORE2 compared to SCORE2 alone ( $\Delta$ AUC =  
21 +0.029, p<0.001). The complete protein set did not outperform the data-driven protein set  
22 when combined with either TRFs (p=0.198) or SCORE2 (p=0.310). The random protein set

1 combined with TRFs or SCORE2 did not significantly change the discrimination compared to the  
2 clinical models alone.

3

4 Reclassification tables and the calculated NRI for all combined models are presented in Figure 1  
5 and Figure S2. At the 5% risk threshold, NRI ranged from 0.026 (random protein set) to 0.044  
6 (complete protein set), and no protein set significantly improved the reclassification when  
7 combined with TRFs and compared with the refitted SCORE2 (Table S2). However, at the 10%  
8 risk threshold, the NRI was significantly increased when the complete protein set (NRI=0.046,  
9  $p=0.039$ ) or the data-driven protein set (NRI=0.053,  $p=0.020$ ) was combined with TRFs and  
10 compared with the refitted SCORE2 (Figure 1B and D). A significantly increased NRI was also  
11 seen when the data-driven (0.048,  $p=0.032$ ) or the hypothesis-driven (0.049,  $p=0.046$ ) protein  
12 sets were combined with SCORE2 at the 5% risk threshold. Further, all four protein sets  
13 significantly increased NRI when combined with SCORE2 at the 10% risk threshold (0.112,  
14 0.140, 1.148 and 0.061,  $p<0.01$ , for the complete, data-driven, hypothesis-driven, and random  
15 protein sets, respectively).

16

17 Model performance at the 5% and 10% cut-off values for 10-year risk of MACE can be analysed  
18 in detail using the reclassification tables (Figure 1 and Figure S2). Using the 5% cut-off, the  
19 models tend to improve classification mainly in the non-MACE group. On the contrary, when  
20 using the 10% cut-off, classification is mainly improved in the MACE group. For example, the  
21 largest reclassification in the non-MACE group was seen when the complete protein set was  
22 combined with SCORE2 and compared to the SCORE2 model using a 5% cut-off: 1290

1 individuals were correctly reclassified while only 187 were incorrectly reclassified (15% correct  
2 reclassification). The largest reclassification in the MACE group was seen when the hypothesis-  
3 driven protein set was combined with SCORE2 and compared to the SCORE2 model using the  
4 10% cut-off for risk. A total of 73 individuals were correctly reclassified while only 10 were  
5 incorrectly reclassified (18% correct reclassification).

6  
7 As shown by reliability curves in Table S4, the clinical, protein and combined models were  
8 correctly calibrated.

9  
10 The 113 candidate protein biomarkers presented in five previous studies in primary  
11 prevention<sup>11-13, 21, 22</sup> and measured in the UK Biobank are shown in Table S4. Three proteins  
12 (GDF15, MMP12 and NTproBNP) were found in 3 studies, ten proteins were found in 2 studies  
13 while the remaining 111 proteins were found only once.

14

## 1 Discussion

2 In this study, we used data from individuals without previous CVD from the UK Biobank to test  
3 whether prediction models for MACE could be improved by addition of subsets derived from  
4 2919 measured proteins. Using a data-driven feature selection, we identified 114 proteins as  
5 being of relevance for prediction of MACE. Prediction by this panel of proteins was equal to that  
6 of SCORE2 and a refitted model based on the TRFs used in SCORE2 but trained on UK Biobank  
7 data. More importantly, the discriminative capacity was significantly increased when the 114  
8 proteins were used in combination with the TRFs used in SCORE2 (10 variables) or the SCORE2-  
9 calculated risk (one variable). We also showed that a hypothesis-driven dataset of 113 proteins  
10 previously suggested as biomarkers of CVD<sup>11-13, 21, 22</sup> added discriminative capacity to SCORE2. A  
11 model using 114 randomly selected proteins did not improve discrimination, supporting the  
12 specificity of the selected proteins. We consider our protein selection successful since the AUC  
13 of the combined complete protein model was not different from the AUC of the combined  
14 model using data-driven protein selection.

15  
16 There are a few large studies in the literature using targeted proteomics to predict CVD events  
17 in populations without previous CVD<sup>11-13, 21, 22</sup>. In a study that used aptamer technology and a  
18 case-control design, a panel of 13 proteins was shown to be equally effective to a traditional  
19 risk model in predicting CVD.<sup>11</sup> In another study using paired antibody probes in a case-control  
20 design, a model based on 50 proteins was superior to a model using refitted TRFs.<sup>13</sup> Our  
21 findings support and extend this result by showing that a data-driven selection of proteins,  
22 from a large set of proteins measured using paired antibody probes, can be used to improve

1 classification of MACE in a large unselected population sample. A recent study using aptamer  
2 technology showed that 70 proteins in combination with TRFs significantly improved the AUC  
3 slightly but only increased NRI significantly when a high risk threshold was used.<sup>12</sup> There are  
4 also two studies showing associations of protein subsets with CVD independent of TRFs.<sup>21, 22</sup> In  
5 the above five studies, a total of 115 unique proteins has been suggested as candidate  
6 biomarkers. Our study supports the selection of these candidate proteins since our hypothesis-  
7 driven panel of proteins also added discriminatory capacity for MACE beyond that of SCORE2.

8  
9 A few of the individual protein candidate biomarkers from our own study and the five  
10 previously published papers<sup>11-13, 21, 22</sup> are common to more than one study (e.g., GDF15,  
11 MMP12 and NTproBNP are in three of the previous studies and in the current study), but most  
12 candidate proteins are unique to one study. This lack of reproducibility could indicate that the  
13 biological signal of CVD risk is not strong enough to overcome variation in data design, cohort  
14 definitions and choice of outcome. It is also possible that the field of large-scale proteomics is  
15 not mature enough to provide stable measurements of large series of biomarkers under  
16 different conditions. Continued work on analytic validity, repeatability, replication, and external  
17 validation is required to improve candidate biomarker generalizability.<sup>26</sup>

18  
19 An important question to address is whether the significant shifts in discriminatory capacity we  
20 found are of clinical benefit. The shifts in AUC were numerically small (up to 0.035) which was  
21 also true for the improvements in NRI (0.04-0.15). Depending on the risk threshold used and  
22 the model tested, up to 18% of participants in the group that will later suffer MACE could be

1 correctly reclassified by combining the data-driven selection of protein biomarkers with  
2 SCORE2. This number are close to the ones presented recently when evaluating large-scale  
3 proteomics for prediction of CVD events in a large Icelandic population<sup>12</sup> and appear to  
4 represent a modest improvement. It also appears from our and this recent analysis<sup>12</sup> that the  
5 benefits of adding proteins to a risk prediction model is best seen at a higher risk threshold. If  
6 true, proteins could be better used as a diagnostic test in a population with a high-pretest  
7 likelihood of diseases than as a screening tool in the general population at low risk. This notion  
8 is also supported by the success of protein biomarkers in secondary prevention<sup>7-10</sup> relative to  
9 primary prevention.<sup>11-13, 21, 22</sup>

10

11 Our study has several limitations. First, our study lacks an external validation set, which limits  
12 the generalisability of our findings. Second, proteins were measured using dual binding affinity  
13 proteomics; although this method is known to have high protein target specificity and a high  
14 number of phenotypic associations,<sup>27</sup> we cannot be sure that our findings can be replicated  
15 using other proteomic platforms. Third, this study was not designed to identify individual  
16 biomarkers, but to create a reliable panel of proteins that could predict MACE; we did not test  
17 for a potential causal role of individual proteins in the development of MACE since this was not  
18 the focus of this paper. Fourth, this report describes some of the potential benefits of using  
19 protein scores for risk estimation. A detailed health-economic assessment of both costs and  
20 benefits of using large-scale proteomics needs to be performed to assess the net clinical  
21 benefit<sup>28</sup> of these improvements, which is beyond the scope of this report. Further,

1 intervention trials are needed to test whether the incremental improvement in classification of  
2 risk, potentially introduced by proteomics, can be transferred into fewer CVD events.

3

#### 4 **Conclusion**

5 Using machine learning and a large set of proteins measured using dual antibody probes, we  
6 could improve identification of MACE with a panel of 114 proteins selected using a data-driven  
7 technique. Similar, although not as convincing, improvements were achieved using a  
8 hypothesis-driven panel of 113 proteins. The improvements are, however, relatively small and  
9 the clinical utility of adding these biomarkers in primary prevention will have to be established.

10

#### 11 **Acknowledgements**

12 This research has been conducted using the UK Biobank Resource under Application Number  
13 82018. The authors are very grateful for the excellent editorial assistance of Rosie Perkins.

#### 14 **Funding**

15 This study was supported by the Swedish Heart Lung Foundation (20210383), the Swedish  
16 Research Council (2019-01140) and grants from the Swedish state under the agreement  
17 between the Swedish government and the county councils, the ALF-agreement (ALFGBG-  
18 718851, ALFGBG-991828).

19

#### 20 **Disclosure of interest**

21 No author reports conflicts of interest.

22

1 **Data availability statement**

2 The data that support the findings of this study are available from UK Biobank.

3



## 1   **References**

- 2       1.    Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, Barengo NC,  
3            Beaton AZ, Benjamin EJ, Benziger CP, Bonny A, Brauer M, Brodmann M, Cahill TJ, Carapetis J,  
4            Catapano AL, Chugh SS, Cooper LT, Coresh J, Criqui M, DeCleene N, Eagle KA, Emmons-Bell S,  
5            Feigin VL, Fernandez-Sola J, Fowkes G, Gakidou E, Grundy SM, He FJ, Howard G, Hu F, Inker L,  
6            Karthikeyan G, Kassebaum N, Koroshetz W, Lavie C, Lloyd-Jones D, Lu HS, Mirijello A,  
7            Temesgen AM, Mokdad A, Moran AE, Muntner P, Narula J, Neal B, Ntsekhe M, Moraes de  
8            Oliveira G, Otto C, Owolabi M, Pratt M, Rajagopalan S, Reitsma M, Ribeiro ALP, Rigotti N,  
9            Rodgers A, Sable C, Shakil S, Sliwa-Hahnle K, Stark B, Sundstrom J, Timpel P, Tleyjeh IM,  
10           Valgimigli M, Vos T, Whelton PK, Yacoub M, Zuhlke L, Murray C, Fuster V, Group G-N-  
11            JGBoCDW. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update  
12            From the GBD 2019 Study. *J Am Coll Cardiol* 2020;**76**(25):2982-3021.
- 13       2.    Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, Himmelfarb CD,  
14            Khera A, Lloyd-Jones D, McEvoy JW, Michos ED, Miedema MD, Munoz D, Smith SC, Jr., Virani  
15            SS, Williams KA, Sr., Yeboah J, Ziaeian B. 2019 ACC/AHA Guideline on the Primary Prevention of  
16            Cardiovascular Disease: A Report of the American College of Cardiology/American Heart  
17            Association Task Force on Clinical Practice Guidelines. *Circulation* 2019;**140**(11):e596-e646.
- 18       3.    Visseren FLJ, Mach F, Smulders YM, Carballo D, Koskinas KC, Back M, Benetos A, Biffi A,  
19            Boavida JM, Capodanno D, Cosyns B, Crawford C, Davos CH, Desormais I, Di Angelantonio E,  
20            Franco OH, Halvorsen S, Hobbs FDR, Hollander M, Jankowska EA, Michal M, Sacco S, Sattar N,  
21            Tokgozoglul, Tonstad S, Tsioufis KP, van Dis I, van Gelder IC, Wanner C, Williams B, Societies  
22            ESCNC, Group ESCSD. 2021 ESC Guidelines on cardiovascular disease prevention in clinical  
23            practice. *Eur Heart J* 2021;**42**(34):3227-3337.

- 1 4. Dekkers OM, Mulder JM. When will individuals meet their personalized probabilities? A  
2 philosophical note on risk prediction. *Eur J Epidemiol* 2020;**35**(12):1115-1121.
- 3 5. Emberson J, Whincup P, Morris R, Walker M, Ebrahim S. Evaluating the impact of population  
4 and high-risk strategies for the primary prevention of cardiovascular disease. *Eur Heart J*  
5 2004;**25**(6):484-91.
- 6 6. Cui M, Cheng C, Zhang L. High-throughput proteomics: a methodological mini-review. *Lab*  
7 *Invest* 2022;**102**(11):1170-1181.
- 8 7. Ganz P, Heidecker B, Hveem K, Jonasson C, Kato S, Segal MR, Sterling DG, Williams SA.  
9 Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes  
10 Among Patients With Stable Coronary Heart Disease. *JAMA* 2016;**315**(23):2532-41.
- 11 8. Nurmohamed NS, Belo Pereira JP, Hoogeveen RM, Kroon J, Kraaijenhof JM, Waissi F,  
12 Timmerman N, Bom MJ, Hoefler IE, Knaapen P, Catapano AL, Koenig W, de Kleijn D, Visseren  
13 FLJ, Levin E, Stroes ESG. Targeted proteomics improves cardiovascular risk prediction in  
14 secondary prevention. *Eur Heart J* 2022;**43**(16):1569-1577.
- 15 9. Wallentin L, Eriksson N, Olszowka M, Grammer TB, Hagström E, Held C, Kleber ME, Koenig W,  
16 März W, Stewart RAH, White HD, Åberg M, Siegbahn A. Plasma proteins associated with  
17 cardiovascular death in patients with chronic coronary heart disease: A retrospective study.  
18 *PLoS Medicine* 2021;**18**(1).
- 19 10. Williams SA, Ostroff R, Hinterberg MA, Coresh J, Ballantyne CM, Matsushita K, Mueller CE,  
20 Walter J, Jonasson C, Holman RR, Shah SH, Sattar N, Taylor R, Lean ME, Kato S, Shimokawa H,  
21 Sakata Y, Nochioka K, Parikh CR, Coca SG, Omland T, Chadwick J, Astling D, Hagar Y, Kureshi N,  
22 Loupy K, Paterson C, Primus J, Simpson M, Trujillo NP, Ganz P. A proteomic surrogate for  
23 cardiovascular outcomes that is sensitive to multiple mechanisms of change in risk. *Sci Transl*  
24 *Med* 2022;**14**(639):eabj9625.

- 1 11. Williams SA, Kivimaki M, Langenberg C, Hingorani AD, Casas JP, Bouchard C, Jonasson C,  
2 Sarzynski MA, Shipley MJ, Alexander L, Ash J, Bauer T, Chadwick J, Datta G, DeLisle RK, Hagar Y,  
3 Hinterberg M, Ostroff R, Weiss S, Ganz P, Wareham NJ. Plasma protein patterns as  
4 comprehensive indicators of health. *Nat Med* 2019;**25**(12):1851-1857.
- 5 12. Helgason H, Eiriksdottir T, Ulfarsson MO, Choudhary A, Lund SH, Ivarsdottir EV, Hjorleifsson  
6 Eldjarn G, Einarsson G, Ferkingstad E, Moore KHS, Honarpour N, Liu T, Wang H, Hucko T,  
7 Sabatine MS, Morrow DA, Giugliano RP, Ostrowski SR, Pedersen OB, Bundgaard H, Erikstrup C,  
8 Arnar DO, Thorgeirsson G, Masson G, Magnusson OT, Saemundsdottir J, Gretarsdottir S,  
9 Steinthorsdottir V, Thorleifsson G, Helgadottir A, Sulem P, Thorsteinsdottir U, Holm H,  
10 Gudbjartsson D, Stefansson K. Evaluation of Large-Scale Proteomics for Prediction of  
11 Cardiovascular Events. *JAMA* 2023;**330**(8):725-735.
- 12 13. Hoogeveen RM, Pereira JPB, Nurmohamed NS, Zampoleri V, Bom MJ, Baragetti A, Boekholdt  
13 SM, Knaapen P, Khaw KT, Wareham NJ, Groen AK, Catapano AL, Koenig W, Levin E, Stroes ESG.  
14 Improved cardiovascular risk prediction using targeted plasma proteomics in primary  
15 prevention. *European Heart Journal* 2020;**41**(41):3998-4007.
- 16 14. Sun BB, Chiou J, Traylor M, Benner C, Hsu YH, Richardson TG, Surendran P, Mahajan A, Robins  
17 C, Vasquez-Grinnell SG, Hou L, Kvikstad EM, Burren OS, Davitte J, Ferber KL, Gillies CE, Hedman  
18 AK, Hu S, Lin T, Mikkilineni R, Pendergrass RK, Pickering C, Prins B, Baird D, Chen CY, Ward LD,  
19 Deaton AM, Welsh S, Willis CM, Lehner N, Arnold M, Worheide MA, Suhre K, Kastenmuller G,  
20 Sethi A, Cule M, Raj A, Alnylam Human G, AstraZeneca Genomics I, Biogen Biobank T, Bristol  
21 Myers S, Genentech Human G, GlaxoSmithKline Genomic S, Pfizer Integrative B, Population  
22 Analytics of Janssen Data S, Regeneron Genetics C, Burkitt-Gray L, Melamud E, Black MH,  
23 Fauman EB, Howson JMM, Kang HM, McCarthy MI, Nioi P, Petrovski S, Scott RA, Smith EN,  
24 Szalma S, Waterworth DM, Mitnau LJ, Szustakowski JD, Gibson BW, Miller MR, Whelan CD.

- 1 Plasma proteomic associations with genetics and health in the UK Biobank. *Nature*  
2 2023;**622**(7982):329-338.
- 3 15. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J,  
4 Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins  
5 R. UK biobank: an open access resource for identifying the causes of a wide range of complex  
6 diseases of middle and old age. *PLoS Med* 2015;**12**(3):e1001779.
- 7 16. Wik L, Nordberg N, Broberg J, Bjorkesten J, Assarsson E, Henriksson S, Grundberg I, Pettersson  
8 E, Westerberg C, Liljeroth E, Falck A, Lundberg M. Proximity Extension Assay in Combination  
9 with Next-Generation Sequencing for High-throughput Proteome-wide Analysis. *Mol Cell*  
10 *Proteomics* 2021;**20**:100168.
- 11 17. Fix E, Hodges JL. Discriminatory Analysis. Nonparametric Discrimination: Consistency  
12 Properties. *International Statistical Review / Revue Internationale de Statistique*  
13 1989;**57**(3):238-247.
- 14 18. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In. *Proceedings of the 22nd*  
15 *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San  
16 Francisco, California, USA: Association for Computing Machinery; 2016, 785–794.
- 17 19. group Sw, collaboration ESCCr. SCORE2 risk prediction algorithms: new models to estimate 10-  
18 year risk of cardiovascular disease in Europe. *Eur Heart J* 2021;**42**(25):2439-2454.
- 19 20. Group SC-DW, the ESCCRC. SCORE2-Diabetes: 10-year cardiovascular risk estimation in type 2  
20 diabetes in Europe. *Eur Heart J* 2023;**44**(28):2544-2556.
- 21 21. Molvin J, Jujic A, Melander O, Pareek M, Rastam L, Lindblad U, Daka B, Leosdottir M, Nilsson  
22 PM, Olsen MH, Magnusson M. Proteomic exploration of common pathophysiological pathways  
23 in diabetes and cardiovascular disease. *ESC Heart Fail* 2020;**7**(6):4151-4158.

- 1        22.    Ho JE, Lyass A, Courchesne P, Chen G, Liu C, Yin X, Hwang SJ, Massaro JM, Larson MG, Levy D.  
2            Protein biomarkers of cardiovascular disease and mortality in the community. *Journal of the*  
3            *American Heart Association* 2018;**7**(14).
- 4        23.    Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. *Journal of Statistical*  
5            *Software* 2010;**36**(11):1 - 13.
- 6        24.    Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive  
7            ability of a new marker: from area under the ROC curve to reclassification and beyond.  
8            *Statistics in medicine* 2008;**27**(2):157-72; discussion 207-12.
- 9        25.    DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more  
10           correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*  
11           1988;**44**(3):837-45.
- 12       26.    Ioannidis JP, Khoury MJ. Improving validation practices in "omics" research. *Science*  
13           2011;**334**(6060):1230-2.
- 14       27.    Katz DH, Robbins JM, Deng S, Tahir UA, Bick AG, Pampana A, Yu Z, Ngo D, Benson MD, Chen ZZ,  
15           Cruz DE, Shen D, Gao Y, Bouchard C, Sarzynski MA, Correa A, Natarajan P, Wilson JG, Gerszten  
16           RE. Proteomic profiling platforms head to head: Leveraging genetics and clinical traits to  
17           compare aptamer- and antibody-based methods. *Sci Adv* 2022;**8**(33):eabm5164.
- 18       28.    Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for  
19           evaluating risk prediction instruments: a critical review. *Epidemiology* 2014;**25**(1):114-21.

20  
21  
22  
23  
24

Table 1. Baseline characteristics of participants

	Total	MACE within ten years		
	n=38380	No n=36719 (95.7)	Yes n=1661 (4.3)	p value
Male sex, n (%)	17241 (44.9)	16091 (43.8)	1150 (69.2)	<0.001
Age (years)	56.51 (8.12)	56.31 (8.12)	60.94 (6.70)	<0.001
BMI (kg/m <sup>2</sup> )	27.38 (4.77)	27.33 (4.76)	28.44 (4.88)	<0.001
Waist (cm)	90.05 (13.35)	89.79 (13.29)	95.81 (13.44)	<0.001
Diabetes, n (%)	563 (1.5)	469 (1.3)	94 (5.7)	<0.001
Age at diabetes diagnosis (years)	56.60 (7.67)	56.57 (7.68)	56.74 (7.68)	0.836
Systolic blood pressure (mmHg)	82.28 (10.69)	82.19 (10.65)	84.15 (11.24)	<0.001
Diastolic blood pressure (mmHg)	139.62 (19.55)	139.29 (19.47)	146.91 (19.86)	<0.001
HDL cholesterol (mmol/l)	1.46 (0.38)	1.46 (0.38)	1.33 (0.39)	<0.001
LDL cholesterol (mmol/l)	3.57 (0.87)	3.57 (0.86)	3.51 (0.98)	0.007
Total cholesterol (mmol/l)	5.72 (1.14)	5.72 (1.13)	5.56 (1.29)	<0.001
Triglyceride (mmol/l)	1.74 (1.03)	1.73 (1.03)	1.96 (1.07)	<0.001
eGFR (mL/min/1.73 m <sup>2</sup> )	94.82 (12.89)	94.99 (12.77)	90.88 (14.80)	<0.001
C-reactive protein (mg/l)	2.59 (4.33)	2.55 (4.27)	3.46 (5.50)	<0.001
Glucose (mmol/l)	5.12 (1.20)	5.11 (1.16)	5.48 (1.90)	<0.001
Haemoglobin A1c (mmol/mol)	36.07 (6.59)	35.93 (6.32)	39.28 (10.31)	<0.001
Smoking - current, n (%)	4117 (10.7)	3807 (10.4)	310 (18.7)	<0.001
Smoking - previous, n (%)	13129 (34.2)	12486 (34.0)	643 (38.7)	<0.001
Smoking - never, n (%)	20968 (54.7)	20274 (55.3)	694 (41.8)	<0.001
Hyperlipidemia medication, n (%)	6085 (15.9)	5486 (14.9)	599 (36.1)	<0.001

Hypertension medication, n (%)	7528 (19.6)	6872 (18.7)	656 (39.5)	<0.001
--------------------------------	-------------	-------------	------------	--------

MACE, major adverse cardiovascular event; BMI, body mass index; HDL, high-density lipoprotein; LDL, low-density lipoprotein; eGFR, estimated glomerular filtration rate.

Values are expressed as mean and standard deviation for continuous variables, and number and proportion for categorical variables.

t-test and Chi-square test were used for continuous and categorical variables.

1

2

Table 2. Performance of the clinical, protein and combined models for predicting the 10-year risk of MACE.

Model	AUC [95% CI]	P value	P value
<b>Clinical model</b>			
Refitted Score2 (10)	0.749 [0.724 - 0.773]		
Score2 (1)	0.740 [0.715 - 0.764]		
		<b>vs refitted Score2</b>	<b>vs Score2</b>
<b>Protein model</b>			
Complete (2919)	0.773 [0.749 - 0.796]	0,014	0,003
Data-driven (114)	0.758 [0.733 - 0.783]	0,379	0,123
Hypothesis-driven (113)	0.759 [0.734 - 0.784]	0,348	0.110
Random (114)	0.712 [0.685 - 0.738]	0,002	0.040
<b>Combined protein model</b>			
<b>With traditional risk factors included in Score2</b>		<b>vs refitted Score2</b>	
Complete (2929)	0.771 [0.748 - 0.795]	0,008	
Data-driven (124)	0.767 [0.742 - 0.791]	0,031	
Hypothesis-driven (123)	0.765 [0.741 - 0.789]	0,057	
Random (124)	0.749 [0.724 - 0.774]	0,953	
<b>With Score2</b>		<b>vs Score2</b>	
Complete (2920)	0.775 [0.751 - 0.799]	<0.001	
Data-driven (115)	0.771 [0.748 - 0.795]	<0.001	
Hypothesis-driven (114)	0.769 [0.744 - 0.793]	<0.001	
Random (115)	0.750 [0.726 - 0.775]	0,061	

MACE, major adverse cardiovascular event; AUC, area under the receiver operating characteristic curve; CI, confidence interval.

AUC for the protein and combined models were respectively compared with AUC for the clinical models (refitted Score2 and Score2) using DeLong's test. The number of variables included in each model is reported in brackets.

- 1
- 2
- 3



1 **Figure legends**

2

3 **Figure 1.** *Reclassification tables.*

4 The figure shows reclassification results in the test set (7676 participants and 344 events) when

5 **(A,B)** complete (2919) and **(C,D)** data-driven (114) protein sets are combined with the 10 TRFs

6 included in SCORE2 for predicting the 10-year risk of first major cardiovascular event (MACE).

7 Panels A and C show net reclassification index (NRI) for the 5% risk threshold, and panels B and

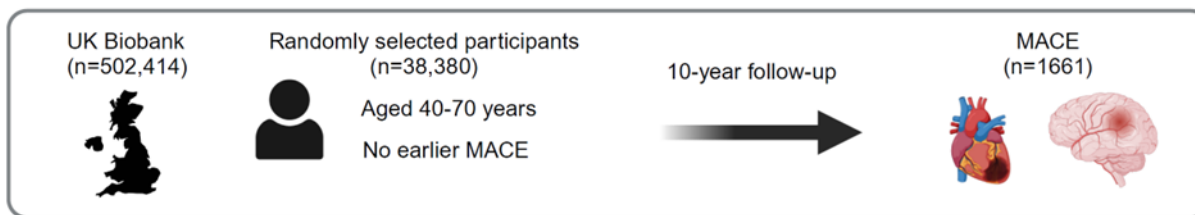
8 D show NRI for the 10% risk threshold.

9 CI, confidence interval.

10

## 1 Structured Graphical Abstract legend.

### Large-scale plasma proteomics for predicting major adverse cardiovascular events (MACE)



Extreme Gradient Boosting algorithm  
For predicting 10-year risk of MACE

Model	AUC [95% CI]	p value
<b>Clinical</b>		
Refitted Score2	0.75 [0.72 - 0.77]	
Score2	0.74 [0.72 - 0.76]	
<b>Combined</b>		
<b>- Traditional risk factors included in Score2</b>		
		<b>vs refitted Score2</b>
Complete proteins (n=2919)	0.77 [0.75 - 0.80]	0.008
Data-driven proteins (n=114)	0.77 [0.74 - 0.79]	0.031
<b>- Score2</b>		
		<b>vs Score2</b>
Complete proteins (n=2919)	0.78 [0.75 - 0.80]	<0.001
Data-driven proteins (n=114)	0.77 [0.75 - 0.80]	<0.001

Large-scale plasma proteomics modestly improves prediction of future MACE beyond traditional risk factors

2

## 3 Key Question



4 Can plasma proteins predict the 10-year risk of first major adverse cardiovascular events (MACE)? Can  
5 they outperform or add a prognostic value to clinical models?

## 6 Key Finding

7 An extreme gradient boosting model trained and tested on 30,704 and 7,676 adults from the UK  
8 Biobank with a data-driven set of 114 plasma proteins improved the prediction of first MACE when  
9 combined to clinical models Score2 and a refitted version of Score2. A protein model alone including a  
10 large set of 2919 plasma proteins outperformed the predictions of the clinical models.

## 11 Take-home Message

12 Plasma proteomics may improve the clinical predictions of first MACE. Further research should precise  
13 the cost benefits and the optimal size of the predictive protein panel to use.

 Correctly reclassified  
 Incorrectly reclassified

**A**

		TRF included in Score2 + Complete proteins		
		Refitted Score2	<5%	≥5%
MACE	<5%	101	34	135 (39)
	≥5%	31	178	209 (61)
	n (%)	132 (38)	212 (62)	344 (100)
No MACE	<5%	4931	419	5350 (73)
	≥5%	676	1306	1982 (27)
	n (%)	5607 (76)	1725 (24)	7332 (100)

NRI [95% CI]: 0.044 [-0.003 - 0.091]  
P value: 0.067

**B**

		TRF included in Score2 + Complete proteins		
		Refitted Score2	<10%	≥10%
MACE	<10%	209	39	248 (72)
	≥10%	20	76	96 (28)
	n (%)	229 (67)	115 (33)	344 (100)
No MACE	<10%	6464	314	6778 (92)
	≥10%	248	306	554 (8)
	n (%)	6712 (92)	620 (8)	7332 (100)

NRI [95% CI]: 0.046 [0.002 - 0.090]  
P value: 0.039

**C**

		TRF included in Score2 + Data-driven proteins		
		Refitted Score2	<5%	≥5%
MACE	<5%	106	29	135 (39)
	≥5%	25	184	209 (61)
	n (%)	131 (38)	213 (62)	344 (100)
No MACE	<5%	4912	438	5350 (73)
	≥5%	648	1334	1982 (27)
	n (%)	5560 (76)	1772 (24)	7332 (100)

NRI [95% CI]: 0.040 [-0.003 - 0.083]  
P value: 0.065

**D**

		TRF included in Score2 + Data-driven proteins		
		Refitted Score2	<10%	≥10%
MACE	<10%	207	41	248 (72)
	≥10%	20	76	96 (28)
	n (%)	227 (66)	117 (34)	344 (100)
No MACE	<10%	6474	304	6778 (92)
	≥10%	242	312	554 (8)
	n (%)	6716 (92)	616 (8)	7332 (100)

NRI [95% CI]: 0.053 [0.008 - 0.097]  
P value: 0.020

medRxiv preprint doi: <https://doi.org/10.1101/2024.03.13.24304196>; this version posted March 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](#) .

medRxiv preprint doi: <https://doi.org/10.1101/2024.03.13.24304196>; this version posted March 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](#) .

medRxiv preprint doi: <https://doi.org/10.1101/2024.03.13.24304196>; this version posted March 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](#) .

medRxiv preprint doi: <https://doi.org/10.1101/2024.03.13.24304196>; this version posted March 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](#) .

medRxiv preprint doi: <https://doi.org/10.1101/2024.03.13.24304196>; this version posted March 15, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](#) .