**Supplementary Methods – Details regarding data preparation and the statistical methods**

1. Preparation and imputation of data for statistical analysis

Protein measurements for 723 proteins were available initially. Outlying values (protein measurements above 80) were set to be missing values. Further, proteins with a call rate below 50% were excluded from the dataset. For the remaining 273 proteins, missing values were imputed based on observations from other proteins. Due to complexity, imputation was organized in 10 clusters, defined by hierarchical clustering where the similarity of two proteins was measured by the absolute value of their Pearson correlation. Within each cluster a multiple imputation assuming a multivariate normal distribution for the protein measurements and missingness at random was used.

To assure numerical stability, the imputation algorithm included a 1 percent (relative to the number of observations) ridge prior. The number of imputations was set to five. To judge the impact of a change in imputation method, a rerun of the subsequently described main analyses with a single imputation obtained by the k-nearest-neighbour method (1) was performed. Finally, for each protein, measurements were standardized using the observed standard deviation and mean (average over the different imputations).

Similarly, multiple imputation relying on a multivariate normal distribution was used to produce five imputations for missing clinical data. The list of variables included in this imputation process is contained in Supplementary Table 1, see below. Although the subsequent statistical analysis required only information on few clinical variables, the imputation procedure was based on a larger set of variables, most of these having sporadic missing values.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Overall** | **Missing** |
| N | 497 |  |
| Males (%) | 354 (71.2) | 0 |
| Age (mean (SD)) | 64.44 (13.96) | 0 |
| Body mass index (mean (SD)) | 26.79 (4.63) | 14.3 |
| Heparin treatment before hospitalization (%) | 361 (72.6) | 0 |
| Aspirin treatment before hospitalization (%) | 434 (87.3) | 0 |
| Adenosine-diphosphate receptor blocker treatment before hospitalization (%) | 347 (70.4) | 0.8 |
| Cardiogenic shock (%) | 35 ( 7.0) | 0 |
| STEMI diagnosis (%) | 381 (76.7) | 0 |
| Anterior STEMI diagnosis (%) | 131 (26.4) | 0 |
| Normal coronary angiography (%) | 390 (78.5) | 0 |
| Multi-vessel disease at coronary angiography (%) | 155 (31.2) | 0 |
| Flow after removal of thrombus (mean (SD)) | 2.91 (0.35) | 32.8 |
| Left ventricular ejection fraction (mean (SD)) | 46.58 (12.14) | 9.9 |
| Systolic blood pressure at hospitalization (mean (SD)) | 126.27 (27.44) | 2.4 |
| Diastolic blood pressure at hospitalization (mean (SD)) | 71.57 (15.08) | 2.6 |
| Heart rate at hospitalization (mean (SD)) | 80.12 (20.21) | 2 |
| Cardiac arrest before hospitalization (%) | 92 (18.7) | 1 |
| Cardiac arrest with coma before hospitalization (%) | 60 (12.1) | 0 |
| Known diabetes (%) | 75 (15.9) | 4.8 |
| Known hypertension (%) | 196 (42.0) | 6 |
| Known kidney disease (%) | 12 ( 2.4) | 0.4 |
| Known hypercholesterolemia (%) | 130 (27.8) | 5.8 |
| Known peripheral artery disease (%) | 18 ( 3.8) | 4.2 |
| Known previous stroke (%) | 24 ( 5.0) | 3.8 |
| Known ischemic heart disease (%) | 99 (20.0) | 0.4 |
| Known heart failure (%) | 18 ( 3.6) | 0.2 |
| Familiar disposition for ischemic heart disease (%) | 142 (34.5) | 17.1 |
| Former or current smoker (%) | 282 (64.4) | 11.9 |

**Supplementary Table 1: Summary of variables used for imputation of clinical data**

Note: Multiple imputation was performed relying on a multivariate normal distribution in order to produce five imputations for the missing clinical data. SD: standard deviation, STEMI: ST-elevation myocardial infarction.

2. Statistical analysis

In total three outcomes were considered: presence of definitive STEMI[[1]](#footnote-1), presence of CS[[2]](#footnote-2), and time to death (as defined by survival for 365 days from admission). For each outcome, two types of analyses were performed to investigate relations to proteins: a univariate association analysis and predictive modelling (see below). Both analyses relied on regression modelling, especially logistic regression (with odds ratios (OR)) for analysis of the presence of definitive STEMI or the presence of CS and Cox regression (with hazard ratios (HR)) for analysis of time to death.

Before including proteins in the analyses, for each outcome, a standard model containing only sex, age, heparin, systolic blood pressure and heart rate was estimated and model fit was checked using Hosmer-Lemeshow's goodness of fit test for CS/definitive STEMI. If necessary, the standard model was extended to include quadratic effects. Model fit for the mortality outcome was investigated by considering an a Cox proportional hazard model including quadratic effects in the numeric variables. If these effects were significant at a 5% level, they were afterwards included in the Cox proportional hazard model.

2.1. Univariate association analysis

Using each of the five sets of imputed proteins, the associations of single proteins by multivariable regression were investigated. For this, the standard model (only adjusting for the demographic/clinical variables) was extended to also include a single protein. Results from the five analyses were then combined according to Rubin 1987 (2) yielding overall effect measures. As this was done for each protein, a Benjamini-Hochberg’s correction for multiple testing (3) was considered, which controls the false discovery rate (FDR)[[3]](#footnote-3) of findings. A cut-off of 0.05 was used in order to declare a protein to be a finding.

2.2 Predictive modelling

For each of the three outcomes, a multivariable prediction model was built by extending the standard model using all imputed proteins as additional potential predictors. Since the number of potential predictors is relatively large in relation to the available observations, one can expect that most predictors do not contribute to a stable prediction. To derive a stable prediction model, the variable selection procedure introduced by Meinshausen et al. 2010 (4) was applied. This procedure allows the control of the expected number of selected variables with low selection probability (ENSP[[4]](#footnote-4)) that represent uninformative predictors, i.e. one controls the number of false discoveries. This control is achieved under quite general conditions on the distribution of the predictors via the complementary pair subsampling procedure of Shah and Samworth 2013 (5). The central building block of this procedure is a regression modelling approach that allows the selection of a predetermined number, say q, of predictors in an optimal manner. In the present study the penalized regression ’lasso’[[5]](#footnote-5) (6) adapted for the outcome being either binary or a survival time was employed. Afterwards a number n of random subsamples of the actual data was generated and a regression model to each such subsample was fitted, hence obtaining n sets of q predictor variables. Based on these sets one estimates the selection probability of the predictors via their relative frequencies of having been chosen. Finally, one retains only those predictors as being stable, which selection probabilities are larger than a pre-chosen threshold probability θ. The influence of the tuning parameters q and θ is relatively uncritical (Meinshausen et al., 2010, page 423). The chosen q and θ can be shown to determine an upper limit for the ENSP. n=50 random subsamples together with q=60 and θ = 0.6, corresponding to an ENSP of about 34, was used. The selection probabilities were obtained running a stability selection for all protein imputations. The overall final prediction model was constituted by the clinical covariates and all those proteins that were present at least 60% of all times. The parameters of the final prediction model were obtained by firstly fitting a regression model for each protein imputation and secondly combining the estimates into overall estimates according to Rubin 1987 (2).

Prediction performance was evaluated as follows: for the final prediction model, fitted values were calculated as average over the fitted values from each of the five protein imputations. Prediction performance was measured in terms of area under the curve (AUC)[[6]](#footnote-6).

This prediction performance is referred to as apparent performance (i.e. using the same data which has been used to derive the model). Apparent performance estimates are known to be overly optimistic due to overfitting occurring when model selection/fitting as well as performance evaluation are based on the same parts of data. To quantify the expected optimism in the apparent prediction performance the tenfold cross-validation (1) was applied. For each fold, a prediction model was selected as described above based on all observations but within this fold. Fitted values were then calculated as described above for all observations within the training set as well as for those observations from the validation set. Then, an AUC was calculated for the training and validation set, separately. Results from both the training as well as from the validation sets were averaged over the different folds. The difference between AUC(training) and AUC(validation) was interpreted as an estimate for the optimism in the apparent performance, and was used to calculate an optimism corrected AUC (=apparent AUC-(AUC(training)-AUC(validation)). The performance is compared with the performance of the standard model.

Computations were performed with the statistical computational environment R. For the stability selection the stabsel function of the R-package stabs7 was applied, imputations were performed with the R-package amelia (8), and the combination of the imputed estimates was done with the mi.inference function of the R-package mix (9). The imputation via the k-nearest-neighbour method was done with the R-package impute (10).

3. Post hoc analysis of troponin levels

A prediction model including initial troponin levels measured in the individuals of the study population was performed as a post hoc analysis: in this model the estimated standard models (i.e. including sex, age, heparin, systolic blood pressure and heart rate) were extended to include initial troponin. Troponin I levels had been measured by Architect STAT High Sensitive Assay (Abbott, USA). The average troponin I level of the study population was 3.09 μg/l (standard deviation: 8.98 μg/l). Missing values were imputed by an additional multiple imputation procedure based on the clinical covariates listed in Supplementary Table 1 above.

4. References

1. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning*, Second Edition, Springer Series in Statistics, 2009.

2. Rubin DB: *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc. New York Chichester, Brisbane, Toronto, Singapore, 1987.

3. Benjamini Y and Hochberg Y: Controlling the false discovery rate a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol* 57, 289-300, 1995.

4. Meinshausen N, Bühlmann P: Stability selection (with discussion). *J R Stat Soc: Series B* 72:417-73, 2010.

5. Shah RD and Richard J. Samworth RJ: Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society, B* 75, 1, 55-80, 2013.

6. Tibshirani R: Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*  58, 1, 267-288, 1996.

7. Hofner B and Hothorn T. stabs: Stability Selection with Error Control, R package version 0.6-3, 2017 (https://CRAN.R-project.org/package=stabs).

8. Honaker J, King G, Blackwell M: Amelia II: A Program for Missing Data. *Journal of Statistical Software* 45(7), 1-47, 2011.

9. Schafer JL: mix: Estimation/Multiple Imputation for Mixed Categorical and Continuous Data. R package version 1.0-10. 2017 (https://CRAN.R-project.org/package=mix).

10. Hastie T, Tibshirani R, Narasimhan B, Chu G: impute: impute: Imputation for microarray data. R package version 1.60.0. 2019. https://www.bioconductor.org/packages/release/bioc/html/impute.html

1. STEMI: ST-elevation myocardial infarction [↑](#footnote-ref-1)
2. CS: cardiogenic shock [↑](#footnote-ref-2)
3. FDR: false discovery rate [↑](#footnote-ref-3)
4. ENSP: expected number of selected variables with low selection probability [↑](#footnote-ref-4)
5. lasso: least absolute shrinkage and selection operator [↑](#footnote-ref-5)
6. AUC: area under the curve [↑](#footnote-ref-6)