Supplementary Table 1. Detailed variables from the SPRINT (Systolic Blood Pressure Intervention Trial) data.

Variable List
MASKID, RZ_AGE, CVDHISTORY, GENDER, BPMEDS, SBPAVG, HRAVG, CIGSMOKER, RACE_WHITE, RACE_BLACK, RACE_INDIAN, RACE_HAWAIIAN, RACE_ASIAN, RACE_OTHER, LIVEWITHOTHERS, EDUCATION, ATRIALFIB, ANGINA, HEARTATT, CONHEARTFAIL, IRRHEARTBEAT, ULCER, CROHNS, DIVERTICULITIS, HEPATITIS, GALLBLADDER, KIDINFECT, BPH, PROSTATITIS, OSTEOARTHRITIS, RHEARTHRITIS, GOUT, OTHARTHRITIS, HIPPROB, CANCER, SKINCANCER, PVD, SEIZURE, STROKE, TIA, THYROIDDIS, ANEMIA, DIABETES, HYPERTENS, LOWBKPAIN, CATARACTS, SCHIZO, DEPRESS, BIPOLAR, ANXIETY, PTSD, ALCOHOL, FAMHST, ALCOHOLDRK, SMOKED100, NOWSMOKE, VIGACTIV, LESSVIGACTIV, ASPIRIN, OTHMED, MEDICARE, MEDICAID, VA, PRIVOTHER, UNINSURED, FULLTIME, RETIRED, PARTTIME, HOUSE, UNEMPLOYED, LOOKING, SEATSYS, SEATHEART, STANDSYS, STANDHEART, DIZZY, MARITALSTATUS, WEIGHT, HEIGHT, RX1, GEN_HEALTH, MOD_ACTS, STAIRS, ACC_LESS_PHYS, LIMIT_WORK_PHYS, ACC_LESS_EMOTION, WORK_CARE_EMOTION, PAIN_INTERFERE, CALM, ENERGY, DOWNHEARTED, SOC_ACTIVITIES, PHYS_HEALTH, EMOT_PROB, FAINT, LITTLE_INTEREST, FEEL_DOWN, SLEEPING, TIRED, EATING, FEEL_BAD, CONCENTRATING, MOVING, THOUGHTS, DIFFICULTY, RESULT_BUN, RESULT_CHR, RESULT_CL, RESULT_CO2, RESULT_CRDUR, RESULT_GLUR, RESULT_HDL, RESULT_K, RESULT_LDLR, RESULT_NA, RESULT_TRR, RESULT_UMALCR, RESULT_UMALI, RESULT_CREATR, RESULT_GFR, QRSDURATION, QUAL, AFIB, AFLUTTER, CV, CVP, SL, AFIBFLUTTER, LVHCV, LVHCVP, LVHMC, LVHSL, LVHANY3, NO_AFIBFLUTTER, NO_LVHCV, NO_LVHCVP, NO_LVHMC, NO_LVHSL, NO_LVHANY3

Supplementary Methods

All baseline information collected in the SPRINT trial was merged together from different datasets by patient id. Baseline information included demographics, medical history, clinical status, anthropometry, laboratory, and ECG data, with more than 120 variables (Supplementary Table 1). Information related to race, insurance, and working status were combined and analyzed as dummy variables, respectively. Medication history was defined as a categorical variable according to whether at least one record name medication existed or not. As for missing, seven variables with more than 20% missing data were deleted, and the remained were imputed with mean and mode for continuous and categorical variables, respectively. The outcome was defined as a composite CVD endpoint, including myocardial infarction, stroke, heart failure, non-MI acute coronary syndrome, or CVD death. The outcome is defined as 1 when either

endpoint is observed, otherwise, 0. When multiple endpoints occurred, the time to the outcome was the earliest observed endpoint time. If no endpoint occurred, the time to the outcome was the maximum follow-up time. The composite CVD outcome was categorical data to apply the methods mentioned in this paper. Survival analysis was additionally performed to display the extension of RF (random forests).

After data preparation, the training dataset with 1,524 participants (70%) was sampled randomly and the remainder was the test dataset. R packages "randomForest", "e1071", and "nnet" were utilized to apply RF, SVM (support vector machine), and NNs (neural networks), respectively. Additionally, the R package "randomForestSRC" was used for the random survival forests application. The training dataset was utilized for model training and then applied the model to the test dataset for prediction. The prediction accuracy was calculated to display the performance of different methods. The top 10 important variables of RF were outputted to display the potential risk factor identification.