Clinical and Genetic Determinants of Heart Failure: Optimized by Machine Learning and Mendelian Randomization

liao li zhen (liaolizhen@gdpu.edu.cn)
Guangdong Pharmaceutical University https://orcid.org/0000-0002-2806-3535
chen zhi chong
Sun Yat-sen University Sixth Affiliated Hospital
li wei dong
Guangdong Pharmaceutical University
liao xin xue
Sun Yat-sen University First Affiliated Hospital
zhuang xiao dong
Sun Yat-sen University First Affiliated Hospital

Original investigation

Keywords: Heart failure, Atherosclerosis Risk in Communities, Machine learning, Multivariable Cox regression, Mendelian randomization

Posted Date: July 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-670567/v1

Abstract

Background: Identifying unrecognized, potentially modifiable risk factors is essential for heart failure (HF) management.

Methods: The Atherosclerosis Risk in Communities (ARIC) study was used for machine learning (ML) to establish the top 20 important variables as potential risk factors for HF. Multivariable Cox regression analysis was performed in an explorative manner to find independent factors for HF and Mendelian randomization (MR) analysis to address causality.

Results: Of the 14,842 participants included in the ARIC analysis, 20.4% of participants (3,028) were identified as HF. The 20 variables with the highest importance selected by ML were creatinine, glucose, age, previous coronary artery disease (CAD), systolic blood pressure, fibrinogen, albumin, income, diabetes, magnesium, insulin, white blood cell, hemoglobin, sodium, education, phosphorus, diastolic blood pressure, protein-c, heart rate and body mass index (BMI). Cox regression analysis demonstrated 19 independently associated variables except sodium. MR analysis provided evidence supporting that genetically determined BMI, CAD, diabetes and education was causally associated with HF.

Conclusions: The ML plus MR framework was useful in identifying important causal factors of HF. BMI, CAD, diabetes, and education not only served as excellent prognostic factors for HF, but therapeutics targeted at these factors were likely to prevent HF effectively.

Background

Heart failure (HF) is considered an epidemic disease in modern world, affecting approximately 1–2% of the adult population.[1] It is the most common cardiovascular cause for hospital admission for people older than 60 year-old.[2] Traditional risk factors for HF mainly include coronary artery disease (CAD), diabetes, hypertension, valvular heart disease, arrhythmia, hypertrophic cardiomyopathy, and inflammatory diseases.[1] Recently, socioeconomic status is reported affects health outcomes.[3] Previous mendelian randomization (MR) study demonstrates that low education is a causal risk factor in the development of CAD,[4] which is the major cause of HF. It implies that some socioeconomic factors may also serve as novel risk factors for HF. Identifying unrecognized, potentially modifiable risk factors is essential for HF management, which is likely to improve the outcome of HF patients.

Machine learning (ML) methods, a field of computer science using algorithms to identify patterns, with lower variance and bias, is a useful approach to identify the best predictors among hundreds of complex phenotypic variables and build more accurate data-driven models.[5] MR analysis, using genetic variants as instrumental variables to test for causality, can infer credible causal associations.[6] In this study, the Atherosclerosis Risk in Communities (ARIC) study is used for ML to establish the top 20 variables as potential risk factors for HF. Multivariable Cox regression analysis is performed in an explorative manner to find independent factors for HF and MR analysis to address causality.

Methods

Study population

ARIC study is an ongoing prospective observational community-based study of the natural history of atherosclerotic diseases and cardiovascular risk factors.[7] In brief, the original cohort was recruited between 1987 and 1989 using probability sampling of 15,121 middle-aged (age 45 to 64 years) adults from 4 U.S. communities. Follow-up visits were carried out in 1990-1992 (93% return rate), 1993-1995 (86%), 1996-1998 (80%) and 2011-2013 (65%). Institutional review boards of all field centers approved the study protocol, and all participants gave informed consent.

The current study analyzed data with individuals who participated in ARIC visit 1 (1987-1989). HF was assessed after ARIC visit 5 with follow-up through December 31, 2015. The median (maximum) follow-up period for HF was 25.05 (28.12)

years. For this analysis, the adaptive tree imputation method was used for imputation of missing data, and variables in less than 40% of the population were excluded. Of 15,121 individuals who completed visit 1, we excluded 279 variables that had missing data more than 60%, and the remained 14,842 participants were included in this analysis.

Definitions of phenotypic variables

We included all phenotypic variables available from questionnaires, physical examination, laboratory biochemistry, electrocardiography (ECG) and ultrasonography tests at visit 1 as potential risk factors for HF.[8] **Supplement Table 1** provides a list of the variables used.

Definition of incident HF

In ARIC study, telephone interviewers contacted participants to inquire about all interim hospital admissions, outpatient diagnoses, and deaths every 6 to 9 months.[9] Two physicians reviewed all medical records for independent endpoint classification and assignment of event dates. Disagreement between discharge coding and computer algorithm were adjudicated by the ARIC Mortality and Morbidity Classification Committee. Incident HF including all subtypes was defined as the first occurrence of hospitalization records and death certificates for a HF diagnosis with an ICD-9 code of 428 (428.0-428.9) or ICD-10 code of 150.

Machine learning for variables selection

300 phenotypic variables were included after eliminating duplicate and meaningless variables. ML methods for variables selection using the Random Survival Forest (RSF) algorithm, an ensemble tree-based method for analysis of right-censored data.[10] While RSF is typically used for prediction, it is also an efficient variable selection technique.[11] The variable importance is ranked by the mean of the minimal depth of the maximal subtree over the entire forest, and variables appearing higher on the tree have a higher rank (and hence are more important).[12] By using the locally weighted scatter smoothing (LOWESS) curve and bar plot in non-parametric regression, the possible nonlinear associations between the survival probability calculated from the RSF method over the range of values for the top-20 variables were assessed.

Cox regression analysis for independent variables

Multivariable Cox regression analyses were performed in an explorative manner to find independent factors for HF. The top 20 variables selected by ML above were chosen for multivariable analysis. A backward stepwise procedure was performed using P>0.10 of the likelihood ratio test for exclusion.

Mendelian randomization for causal analysis

Based on the Cox regression analysis, the 19 independent variables other than serum sodium were selected for causal association analysis by MR method.

We included summary GWAS data from any array-based analysis, including targeted and untargeted arrays, with or without additional imputation for single nucleotide polymorphism (SNP). We also collected published GWAS associations that comprise only the significant hits of a GWAS after applying stringent p-value thresholds (e.g., $P<5*10^{-8}$, a conventional threshold for declaring statistical significance in GWAS), using the clumping algorithm (r^2 threshold=0.05 and window size=1 Mb). Data included in this study were the GWAS summary statistics from the MR-Base platform.[13] Details of studies and datasets used for analyses were presented in **Supplement Table 3**.

We then performed MR in a strategy known as two-sample MR by using results from GWAS.[13] We explored the associations in the following scenarios: 1) Causal associations between the potential factors and HF. We applied inverse-variance weighted (IVW) method for deriving causal estimates.[15] 2) Heterogeneity: We conducted heterogeneity tests in MR

analyses using IVW and MR-Egger approach. 3) Horizontal pleiotropy: it referred to when a genetic variant associated with traits on discrete pathways that were also causal in disease.[16] It was evaluated by P-value of the MR-Egger intercept.

The “causal” relationship was considered established if the observed association passed the IVW method and with no heterogeneity nor horizontal pleiotropy. We used another two MR methods including weighted median and MR-Egger for methodology sensitivity analysis.[15]

Statistical analysis

For all analyses, surviving patients were right censored to their follow-up date. Data transformation, indexing, and imputation were performed as necessary to generate data points to predict outcomes over the follow-up period. Using the imputed dataset, each continuous variable was centralized to the mean and scaled to the standard deviation, whereas categorical variables were coded into binary numbers (0 and 1). Descriptive data were presented as the mean ± SD for normally distributed variables and median (25th, 75th percentile) for non-normally distributed variables. P values were 2-sided, and evidence of association was declared at $P < 0.05$. Analyses were performed using R software (www.r-project.org) and Stata release 13.1 (StataCorp LP).

Results

Baseline characteristics

Baseline characteristics of the study sample are shown in **Table 1**. Of the 14,842 participants included in the analysis, the average age was 54.2 years with 45.2% male, 26.2% black. At visit 5, 20.4% of participants (3,028) were identified as HF.

Important variables selection and Cox regression analysis

As presented in **Figure 1**, the top-20 variables with the highest importance selected by RSF for HF were creatinine, glucose, age, previous CAD, systolic blood pressure, fibrinogen, albumin, income, diabetes, magnesium, insulin, white blood cell, hemoglobin, sodium, education level, phosphorus, diastolic blood pressure, protein-c, heart rate, and body mass index (BMI). The 20 important variables were then chosen for multivariable Cox analysis (**Table 2**) and non-linear analysis (**Supplement Figure 1**). Cox regression analysis demonstrated 19 independently associated variables except sodium. LOWESS curve revealed nonlinear associations between the survival probability over the range of values for heart rate, insulin and BMI (**Supplement Figure 1**).

Mendelian randomization

These 19 variables were then selected for causal association analysis by MR (**Table 3**). The IVW method estimate indicated that the odds ratio (OR) (95% confidence interval [CI]) for HF was 1.001 (1.001-1.002) per kg/m^2 increase in BMI. Results were consistent in weighted median method (OR, 1.001; 95% CI, 1.000-1.002; P=0.002). Both IVW and MR-Egger estimates indicated that there was no heterogeneity amongst the 304 SNPs in the causal effect between BMI and HF (P=0.165 and P=0.158 respectively). Moreover, there was no evidence of directional horizontal pleiotropy (MR-Egger intercept P=0.643). Our two-sample MR analysis provided evidence supporting that genetically predicted BMI was casual associated with HF.

Notedly, the IVW method estimate indicated that OR (95% CI) for HF was 0.999 (0.998-1.000) per standard deviation increase (3.6 years) in education. Results were consistent in weighted median method (OR, 0.998; 95% CI, 0.996-0.999; P=0.003) without evidence of heterogeneity and directional horizontal pleiotropy (MR-Egger intercept P=0.999). So, the genetically predicted education was negatively casual associated with HF. Similar results were also detected in CAD, diabetes and income.

Discussion

In this study, we used an ML approach in a hypothesis-free manner to identify important factors of HF in ARIC study. This powerful method confirmed several well-established relationships and identified a variety of novel factors which have not been previously reported. We then used Cox regression analysis in an explorative manner to find independent factors and MR analysis to address causality. Our findings revealed that BMI, CAD, diabetes, and education, not only served as prognostic factors for HF, but potential therapeutic targets for the treatment and prevention of HF.

Established and novel prognostic factors for HF

HF prevalence has increased exponentially over the last three decades. This increase is attributable to several factors, including an aging population, and recent advances in the treatment of cardiovascular disease, leading to increased survival following an acute cardiac event.[17] Prior studies have yielded inconsistencies in predictors of HF. Our ML results indicated that the 20 variables with the highest importance selected by RSF for HF are creatinine, glucose, age, previous CAD, systolic blood pressure, fibrinogen, albumin income, diabetes, magnesium, insulin, white blood cell, hemoglobin, sodium, education level, phosphorus, diastolic blood pressure, protein-c, heart rate and BMI. It confirmed several established risk factors as previously reported (e.g., age, history of CAD, diabetes, BMI, hemoglobin, total white blood cell, creatinine[18] and hypertension).[19-20] Yet to our surprises, a recent study in 20,254 US male veterans revealed that increased cardiorespiratory fitness was associated with progressively lower HF risk regardless of BMI, challenging BMI served as a well-established risk factors for HF.[21] Our result was different from the above research, and we analyzed that it might due to the different population included.

As for glucose impartment, it was reported that there existed a positive, continuous, and independent association between fasting plasma glucose and risk for HF.[22] The British Regional Heart Study carried out in older men demonstrated that serum magnesium was inversely related to risk of incident HF after adjustment for conventional CVD risk factors and incident MI.[23] Our results also support a positive association between glucose and HF and a negative association between magnesium. A number of other well-established factors that have been reported in literatures, including smoking, atrial fibrillation, chronic obstructive pulmonary disease,[1],[19] were not observed among the top 20 predictors by ML approach. Though not selected, it did not mean that these traditional risk factors were not important for HF, as most of risk factors had adverse effects on cardiac structure, which ultimately would result in HF.

It was noteworthy that several novel predictors of HF were identified, including fibrinogen, albumin, income, education, phosphorus, protein-c. Fibrinogen was suggestive associated with incident HF that had preserved ejection fraction (HR 1.12; 95% CI 1.03-1.22; P=0.01).[24] A prospective study of 3,366 men found that fibrinogen was associated with incident HF but this was abolished after adjustment for HF risk factors.[25] So fibrinogen as a risk factor for HF was still controversial. Many epidemiological studies have suggested an inverse association between serum albumin level and HF. In the Health ABC study of 2,907 elderly individuals with a 9.4 years follow-up, low serum albumin level was associated with the development of new-onset HF, mainly with preserved ejection fraction, regardless of inflammatory markers, BMI and CHD.[25] So low albumin level might serve as a novel predictor of increased risk of HF. In a very large population (N = 7,638,524) of chronic HF patients with access to universal healthcare, lower income was independently associated with higher mortality.[27] Another study conducted in 54 countries reported that greater income inequality was associated with worse HF outcomes, with an impact similar to those of major comorbidities.[28] More importantly, previous MR study demonstrated that genetic predisposition towards 3.6 years of additional education was associated with a one third lower risk of CAD, supporting that low education is a causal risk factor in the development of CAD,[29] which is the major cause of HF. These studies suggested that socioeconomic status (income and education) might also affect HF besides the traditional risk factors.

Low serum magnesium and high serum phosphorus were identified independently associated with greater risk of incident HF in ARIC cohort,[30] which was in accordance with our results. Our multivariable analysis revealed that protein C was slight negative associated with HF. Yet a prospective case-control study that involved 50 children demonstrated that there was a significant increase in plasma levels of cardiac myosin binding protein-C in patients with HF and this increase was associated with increased severity of HF, indicating positive association between cardiac myosin binding protein-C and HF. We analyzed the different associations might lie to different HF population and different kinds of protein-C.

In summary, by ML approach and multivariable analysis, we identified 13 traditional risk factors, including creatinine, glucose, age, previous CAD, systolic blood pressure, diabetes, magnesium, insulin, white blood cell, hemoglobin, sodium, diastolic blood pressure, heart rate and BMI, and 6 novel risk factors, including fibrinogen, albumin, income, education, phosphorus and protein-c.

Causality between the potential risk factors and HF

It would be of clinical value if the modifiable risk factors, such as BMI and education, were shown to causally lead to the development of HF.

As for causal factors of HF, hyperhomocysteinemia[30] and elevated lipoprotein(a) levels[32] were reported to be causally associated with HF. Because these two factors were not selected as top 20 variables, we did not analyze their causal estimate with HF. To our surprise, a recent MR study reported that though there was an observational association of CAD with HF, the genetically determined risk of CAD was significantly associated with HF with reduced ejection fraction but not with HF with preserved ejection fraction,[33] indicating that HF with reduced and preserved ejection fraction should be treated differentially.

Our MR analysis showed that the genetically predicted BMI, CAD and diabetes was positive casually associated with HF, and education as a novel factor was also negative casually associated with HF. Among these four factors, education might serve as the source of some established risk factors, for education reflected socioeconomic circumstances and cognitive level. People with higher education level were usually with more self-management skills to maintain healthy status and access better health care. Previous findings also indicated a causal association between low educational attainment and increased risk of smoking,[34] which was risk factors for both CAD and diabetes.

Conclusions

BMI, CAD, diabetes and education not only serve as excellent prognostic factors for HF, but potential therapeutic targets for the treatment and prevention of HF. In other words, these four factors served as both “markers” and “makers” for HF.

Abbreviations

heart failure (HF)

Atherosclerosis Risk in Communities (ARIC)

machine learning (ML)

Mendelian randomization (MR)

coronary artery disease (CAD)

body mass index (BMI)

electrocardiography (ECG)

Random Survival Forest (RSF)

single nucleotide polymorphism (SNP)

inverse-variance weighted (IVW)

body mass index (BMI)

Declarations

Ethics approval and consent to participate

No applicable.

Consent for publication

The authors declared they consented for the publication of this manuscript.

Availability of data and materials

For more data and materials, please contact Xiao-dong Zhuang (zhuangxd3@mail.sysu.edu.cn).

Competing interests

None.

Funding

None.

Authors' contributions

L.L.Z, C.Z.C. performed statistical analysis; L.W.D. acquired the data; L.L.X. drafted the manuscript; Z.X.D. made critical revision of the manuscript for critical intellectual content.

Acknowledgements

None.

References

1. Tanai E, Frantz S. Pathophysiology of Heart Failure*Compr Physiol*. 2015;6:187–214.
2. Rossignol P, Hernandez AF, Solomon SD. F Zannad Heart failure drug treatment*Lancet*. 2019;393:1034–44.
3. DiPrete TA,CAP Burik PD, Koellinger. Genetic instrumental variable regression: Explaining socioeconomic and health outcomes in nonexperimental data. Proc Natl Acad Sci U S A. 2018;115:E4970–9.
4. Tillmann T, Vaucher J, Okbay A, et al. Education and coronary heart disease: mendelian randomisation study.*BMJ*. 2017;358:j3542.
5. Mandl KD, Manrai AK. Potential Excessive Testing at Scale: Biomarkers, Genomics, and Machine Learning.*JAMA*. 2019.
6. Burgess S, Labrecque JA. Mendelian randomization with a binary exposure variable: interpretation and presentation of causal estimates. Eur J Epidemiol. 2018;33:947–52.

7. A. R. Sharrett. The Atherosclerosis Risk in Communities (ARIC) Study. Introduction and objectives of the hemostasis component.*Ann Epidemiol*. 1992;2:467–469.
8. Norby FL, Soliman EZ, Chen LY, et al. Trajectories of Cardiovascular Risk Factors and Incidence of Atrial Fibrillation Over a 25-Year Follow-Up: The ARIC Study (Atherosclerosis Risk in Communities). Circulation. 2016;134:599–610.
9. Magnani JW, Norby FL, Agarwal SK, et al. Racial Differences in Atrial Fibrillation-Related Cardiovascular Disease and Mortality: The Atherosclerosis Risk in Communities (ARIC) Study. JAMA Cardiol. 2016;1:433–41.
10. Wright MN, Dankowski T, Ziegler A. Unbiased split variable selection for random survival forests using maximally selected rank statistics. Stat Med. 2017;36:1272–84.
11. Nasejje JB, Mwambi H, Dheda K, Lesosky M. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. BMC Med Res Methodol. 2017;17:115.
12. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circ Res. 2017;121:1092–101.
13. Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome.*Elife*. 2018;7.
14. Bowden,Greco J, Del MF, Minelli C, et al. Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption.*Int J Epidemiol*. 2018.
15. Pierce BL, Burgess S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. Am J Epidemiol. 2013;178:1177–84.
16. Hemani G, Tilling,Smith K, Davey G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLoS Genet. 2017;13:e1007081.
17. Benjamin EJ, Virani SS, Callaway CW, et al. Heart Disease and Stroke Statistics-2018 Update: A Report From the American. Heart Association*Circulation*. 2018;137:e67–492.
18. Metra M, Cotter G, Senger S, et al. Prognostic Significance of Creatinine Increases During an Acute Heart Failure Admission in Patients With and Without Residual Congestion: A Post Hoc Analysis of the PROTECT Data. Circ Heart Fail. 2018;11:e4644.
19. Uijl A, Koudstaal S, Direk K, et al. Risk factors for incident heart failure in age- and sex-specific strata: a population-based cohort using linked electronic health records.*Eur J Heart Fail*. 2019.
20. Chatterjee NA, Chae CU, Kim E, et al. Modifiable Risk Factors for Incident Heart Failure in Atrial Fibrillation. JACC Heart Fail. 2017;5:552–60.
21. Kokkinos P, Faselis C, Franklin B, et al. Cardiorespiratory fitness, body mass index and heart failure incidence. Eur J Heart Fail. 2019;21:436–44.
22. Khan H, Kunutsor SK, Kauhanen J, et al. Fasting plasma glucose and incident heart failure risk: a population-based cohort study and new meta-analysis. J Card Fail. 2014;20:584–92.
23. Wannamethee SG, Papacosta O, Lennon L, Whincup PH. Serum magnesium and risk of incident heart failure in older men: The British Regional Heart Study. Eur J Epidemiol. 2018;33:873–82.
24. de Boer RA, Nayor M, DeFilippi CR, et al. Association of Cardiovascular Biomarkers With Incident Heart Failure With Preserved and Reduced Ejection Fraction. JAMA Cardiol. 2018;3:215–24.
25. Wannamethee SG, Whincup PH, Papacosta O, Lennon L, Lowe GD. Associations between blood coagulation markers, NT-proBNP and risk of incident heart failure in older men: The British Regional Heart Study. Int J Cardiol. 2017;230:567–71.
26. Gopal DM, Kalogeropoulos AP, Georgiopoulou VV, et al. Serum albumin concentration and heart failure risk The Health, Aging, and Body Composition Study. Am Heart J. 2010;160:279–85.

27. Cainzos-Achirica M, Capdevila C, Vela E, et al. Individual income, mortality and healthcare resource use in patients with chronic heart failure living in a universal healthcare system: A population-based study in Catalonia. Spain*Int J Cardiol*. 2019;277:250–7.
28. Dewan P, Rorth R, Jhund PS, et al. Income Inequality and Outcomes in Heart Failure: A Global Between-Country Analysis. JACC Heart Fail. 2019;7:336–46.
29. Tillmann T, Vaucher J, Okbay A, et al. Education and coronary heart disease: mendelian randomisation study.*BMJ*. 2017;358:j3542.
30. Lutsey PL, Alonso A, Michos ED, et al. Serum magnesium, phosphorus, and calcium are associated with risk of incident heart failure: the Atherosclerosis Risk in Communities (ARIC) Study. Am J Clin Nutr. 2014;100:756–64.
31. Strauss E, Supinski W, Radziemski A, Oszkinis G, Pawlak AL, Gluszek J. Is hyperhomocysteinemia a causal factor for heart failure? The impact of the functional variants of MTHFR and PON1 on ischemic and non-ischemic etiology. Int J Cardiol. 2017;228:37–44.
32. Kamstrup PR. B. G. Nordestgaard. Elevated Lipoprotein(a) Levels, LPA Risk Genotypes, and Increased Risk of Heart Failure in the General Population. JACC Heart Fail. 2016;4:78–87.
33. Mordi IR, Pearson ER, Palmer CNA, Doney ASF. C. C. Lang. Differential Association of Genetic Risk of Coronary Artery Disease With Development of Heart Failure With Reduced Versus Preserved Ejection Fraction. Circulation. 2019;139:986–8.
34. Gage SH, Bowden,Smith J, Davey G. M. R. Munafo. Investigating causality in associations between education and smoking: a two-sample Mendelian randomization study. Int J Epidemiol. 2018;47:1131–40.

Tables

Table 1. Baseline Characteristics of Study Population in Atherosclerosis Risk in Communities at Visit 1 (from 1987 to 1989).

Variable (n= 14,842)	Value
Age, year	54.2 ± 5.8
Male, %	45.2
Race, %	
Black	26.2
Non-black	73.8
BMI, m/kg^2	27.7 ± 5.4
Heart rate	66.7 ± 10.3
SBP, mm Hg	121.4 ± 19.0
DBP, mm Hg	73.7 ± 11.3
Smoking, %	
Current smoker	26.2
Former smoker	32.3
Ever smoked	58.5
Drinking, %	
Current Drinker	6.2
Former Drinker	19.5
Ever Drinker	74.3
Education Level, %	
Basic Education	23.9
Intermediate Education	40.7
Advanced Education	35.4
Sport index	2.4 ± 0.8
Leisure time index	2.4 ± 0.6
Previous history, %	
Heart failure	5.1
Coronary Heart Disease	4.9
Diabetes	9.8
Hypertension	35.3
Stroke or TIA	6.2
Medication history, %	
Anti-Hypertension	31.4
Cholesterol Lowering	3.3
Stain	1.4

Aspirin	47.3
Anticoagulants	1.2
Laboratory markers	
HDL-c, mmol/l	1.3 ± 0.4
LDL-c, mmol/l	3.6 ± 1.0
TG, mmol/l	1.5 ± 1.0
Glucose, mg/dl	108.9 ± 40.5
NT-proBNP, pg/ml	138.9 (70.7, 281.8)
Hs-CRP, mg/l	2.05 (0.9, 4.4)
Hs-TNT, ng/l	0.011 (0.007, 0.016)
Cardiovascular outcomes, %	
All-cause Mortality	42.8
Cardiovascular Disease	32.3
Coronary Heart Disease	16.4
Stroke	8.7
Heart Failure	20.4
Atrial Fibrillation	17.5

Values are mean ± SD, %, or median (25th, 75th percentiles). SBP, Systolic blood pressure; DBP, Diastolic blood pressure; TIA, transient ischemic attack; HDL-c, high-density lipoprotein cholesterol; LDL-c, low-density lipoprotein cholesterol; TG, Triglyceride; hs-CRP, high-sensitivity C-reactive protein; hs-cTnT, high-sensitivity cardiac troponin T; NT-proBNP, N-terminal pro-B-type natriuretic peptide.

Table 2. The Number of Variables and the Performance (Concordance Index and Brier Score) for Each of the Models

Discriminative power (C index)			**Model accuracy (Brier score)**			**Model interpretability (Number of Variables)**		
All variables	RSF	0.807	**All variables**	RSF	0.036	**All variables**	RSF	299
	Cox-lasso	0.804		Cox-lasso	0.095		Cox-lasso	102
	Cox-mcp	0.799		Cox-mcp	0.096		Cox-mcp	47
Top-20 by RSF	RSF	0.795	**Top-20 by RSF**	RSF	0.038	**Top-20 by RSF**	RSF	20
	Cox-lasso	0.778		Cox-lasso	0.105		Cox-lasso	19
	Cox-mcp	0.778		Cox-mcp	0.106		Cox-mcp	20
	Cox-AIC	0.777		Cox-AIC	0.106		Cox-AIC	18
	Cox-PH	0.778		Cox-PH	0.106		Cox-PH	20

RSF, random survival forest; Cox-lasso, least absolute shrinkage and selection operator for Cox regression; Cox-mcp, minimax concave penalty for Cox regression; Cox-AIC, Akaike Information Criteria for Cox regression; Cox-PH, Cox proportional hazards regression.

Table 3. Multivariable Cox regression analysis of the top 20 predictors for HF.

Ranking	Variables	Multi-beta	multi-HR (95% CI for HR)	P
1	CREATININE	0.26308	1.3009 (1.2423-1.3623)	<<0.0001
2	GLU	0.051188	1.0525 (1.0351-1.0702)	1.76E-09
3	AGE_v1	0.064022	1.0661 (1.0583-1.074)	<<0.0001
4	PRVCHD	1.1922	3.2942 (2.9054-3.7351)	<<0.0001
5	SBP	0.015323	1.0154 (1.0129-1.018)	<<0.0001
6	FIBRINOGEN	0.0022969	1.0023 (1.0018-1.0028)	1.11E-16
7	ALB	-0.57163	0.56461 (0.48271-0.6604)	8.74E-13
8	INCMOE	-0.072163	0.93038 (0.91133-0.94983)	8.06E-12
9	DIABTS	0.40199	1.4948 (1.3055-1.7115)	5.88E-09
10	MG	-0.88347	0.41335 (0.32632-0.52358)	2.39E-13
11	INSULIN	0.0016487	1.0017 (1.0008-1.0025)	0.00021251
12	WBC	0.05509	1.0566 (1.0444-1.069)	<<0.0001
13	HB	0.13976	1.15 (1.1154-1.1856)	<<0.0001
14	NA.	-0.0021617	0.99784 (0.9817-1.0142)	0.795
15	ELEVEL01	-0.092631	0.91153 (0.8867-0.93705)	4.90E-11
16	P	0.15997	1.1735 (1.0874-1.2664)	3.85E-05
17	DBP	-0.007178	0.99285 (0.9884-0.99732)	0.0017461
18	PROTEIN_C	-0.089702	0.9142 (0.85955-0.97233)	0.0043427
19	HR	0.0061942	1.0062 (1.0026-1.0099)	0.00086084
20	BMI	0.030836	1.0313 (1.0243-1.0384)	<<0.0001

HF, Heart failure; CREATININE, Creatinine; GLU, Glucose; AGE_v1, Age in v1; PRVCHD, Previous coronary heart disease; SBP, Systolic blood pressure; FIBRINOGEN, Fibrinogen; ALB, Albumin; INCMOE, Income; DIABTS, Diabetes; MG, Magnesium; INSULIN, Insulin; WBC, White blood cell; HB, Hemoglobin; NA, Sodium; ELEVEL01, Education level; P, Phosphorus; DBP, Diastolic blood pressure; PROTEIN_C , Protein-c; HR, Heart rate; BMI, Body mass index.

Table 4. Causal association between genetically determined established risk factors and HF

Classification	Established risk factors	Method	nsnp	OR	95%□CI□		P value	Heterogeneity P	MR-Egger intercept P
Tradional	CREATININE	MR Egger	43	0.999	0.988	1.011	0.871	0.044	0.726
		Weighted median	43	1.001	0.997	1.005	0.650		
		Inverse variance weighted	43	1.001	0.998	1.004	0.582	0.053	
	GLU	MR Egger	14	1.000	0.997	1.002	0.830	0.125	0.461
		Weighted median	14	1.000	0.999	1.001	0.945		
		Inverse variance weighted	14	1.001	0.999	1.002	0.296	0.137	
	AGE_v1	NA							
	PRVCHD	MR Egger	39	1.001	1.000	1.001	0.175	0.017	0.997
		Weighted median	39	1.001	1.000	1.001	0.000		
		Inverse variance weighted	39	1.001	1.000	1.001	0.001	0.022	
	SBP	MR Egger	164	1.000	0.998	1.002	0.803	0.292	0.372
		Weighted median	164	1.001	1.000	1.002	0.016		
		Inverse variance weighted	164	1.001	1.000	1.001	0.051	0.295	
	DIABTS	MR Egger	36	1.000	1.000	1.001	0.156	0.915	0.644
		Weighted median	36	1.000	1.000	1.001	0.121		
		Inverse variance weighted	36	1.000	1.000	1.001	0.014	0.929	
	MG	MR Egger	5	0.999	0.996	1.002	0.479	0.956	0.666
		Weighted median	5	0.999	0.996	1.001	0.160		
		Inverse variance weighted	5	0.998	0.996	1.000	0.127	0.977	
	INSULIN	Inverse	2	0.999	0.998	1.001	0.347	0.324	NA

		variance weighted							
	DBP	MR Egger	177	1.001	0.999	1.003	0.537	0.248	0.883
		Weighted median	177	1.001	1.000	1.002	0.069		
		Inverse variance weighted	177	1.000	1.000	1.001	0.096	0.265	
	HR	MR Egger	15	1.000	1.000	1.001	0.610	0.146	0.631
		Weighted median	15	1.000	1.000	1.000	0.337		
		Inverse variance weighted	15	1.000	1.000	1.000	0.849	0.179	
	BMI	MR Egger	304	1.001	1.000	1.002	0.151	0.158	0.643
		Weighted median	304	1.001	1.000	1.002	0.002		
		Inverse variance weighted	304	1.001	1.001	1.002	0.000	0.165	
	WBC	Wald ratio	1	1.000	0.997	1.004	0.791	NA	NA
	HB	MR Egger	5	1.001	0.998	1.003	0.656	0.785	0.816
		Weighted median	5	1.000	0.999	1.001	0.581		
		Inverse variance weighted	5	1.000	0.999	1.001	0.478	0.890	
Novel	FIBRINOGEN	MR Egger	10	0.999	0.983	1.016	0.951	0.347	0.784
		Weighted median	10	1.003	0.996	1.010	0.468		
		Inverse variance weighted	10	1.002	0.996	1.007	0.530	0.434	
	ALB	Inverse variance weighted	2	1.000	0.999	1.001	0.546	0.475	NA
	INCMOE	Wald ratio	1	1.005	1.000	1.010	0.044	NA	NA
	ELEVEL01	MR Egger	70	0.999	0.993	1.004	0.584	0.270	0.999
		Weighted median	70	0.998	0.996	0.999	0.003		

	Inverse variance weighted	70	0.999	0.998	1.000	0.005	0.298	
P	MR Egger	4	1.004	0.992	1.017	0.574	0.961	0.584
	Weighted median	4	1.000	0.998	1.002	1.000		
	Inverse variance weighted	4	1.000	0.998	1.002	0.848	0.971	
PROTEIN_C	NA							

HF, Heart failure; CREATININE, Creatinine; GLU, Glucose; AGE_v1, Age in v1; PRVCHD, Previous coronary heart disease; SBP, Systolic blood pressure; FIBRINOGEN, Fibrinogen; ALB, Albumin; INCMOE, Income; DIABTS, Diabetes; MG, Magnesium; INSULIN, Insulin; WBC, White blood cell; HB, Hemoglobin; ELEVEL01, Education level; P, Phosphorus; DBP, Diastolic blood pressure; PROTEIN_C, Protein-c; HR, Heart rate; BMI, Body mass index.

Figures

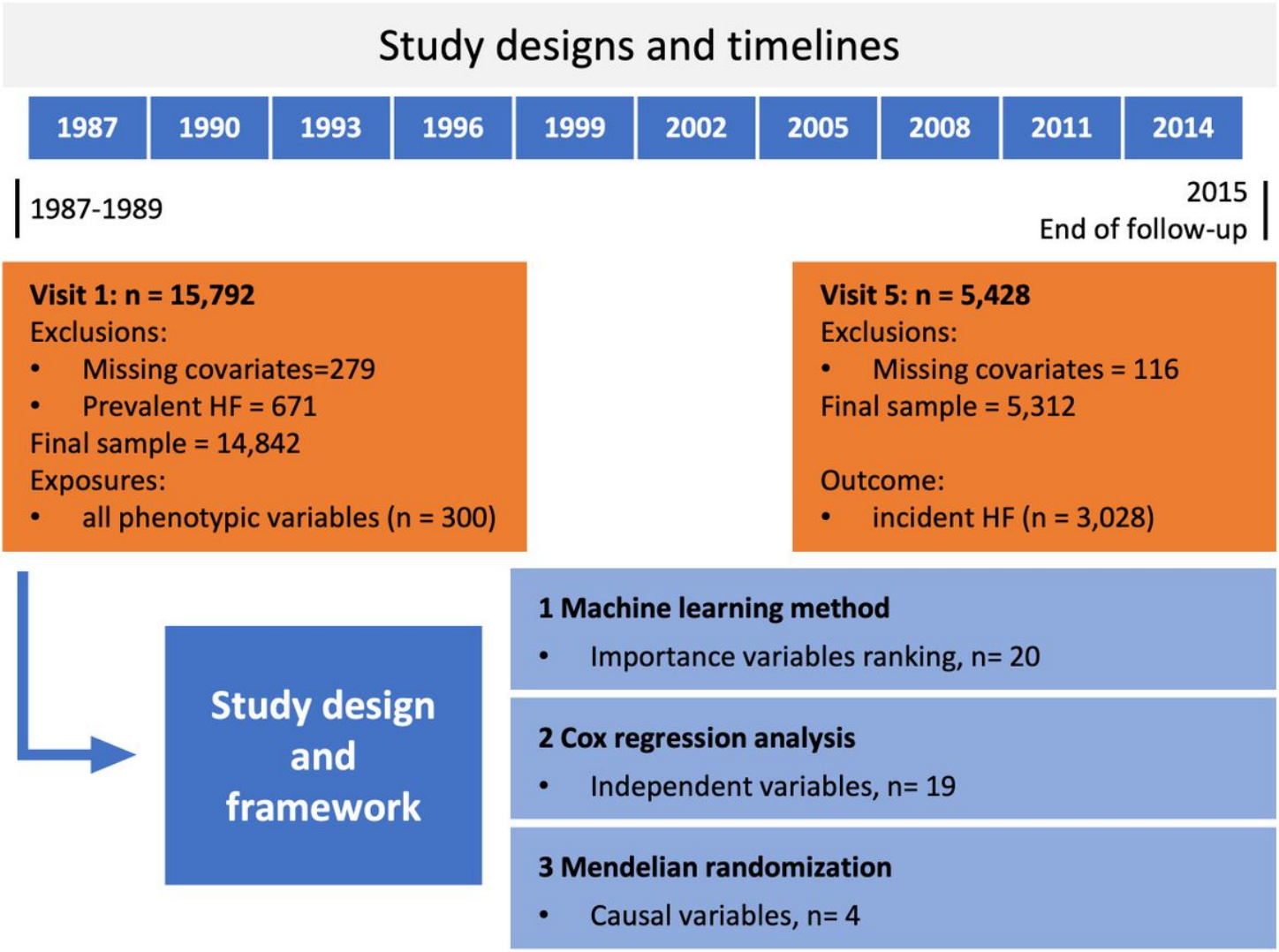

Figure 1

Timeline, sample size and study framework in this analysis. HF, heart failure.

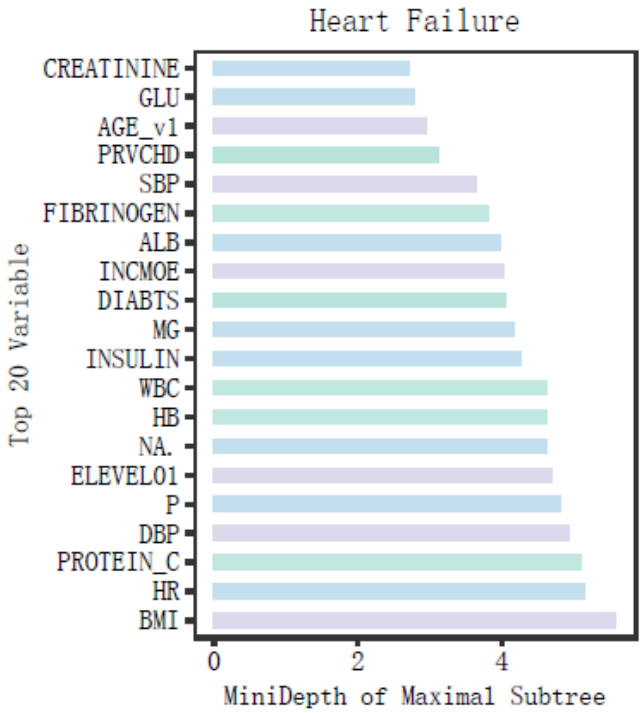

Figure 2

Variable importance ranking by random survival forest model for heart failure. The variable importance is measured using the minimum depth of the maximal subtree (Y-axis), with lower values representing greater importance of corresponding variable (top to bottom). CREATININE, Creatinine; GLU, Glucose; AGE_v1, Age in v1; PRVCHD, Previous coronary heart disease; SBP, Systolic blood pressure; FIBRINOGEN, Fibrinogen; ALB, Albumin; INCMOE, Income; DIABTS, Diabetes; MG, Magnesium; INSULIN, Insulin; WBC, White blood cell; HB, Hemoglobin; NA, Sodium; ELEVEL01, Education level; P, Phosphorus; DBP, Diastolic blood pressure; PROTEIN_C, Protein-c; HR, Heart rate; BMI, Body mass index.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementFigure1Linearanalysis.pdf
- SuppTables.doc