Published in final edited form as:
Circ Heart Fail. 2021 February ; 14(2): e006799. doi:10.1161/CIRCHEARTFAILURE.119.006799.

The Association between Coffee Intake and Incident Heart Failure Risk – A Machine Learning Analysis of the Framingham Heart, Atherosclerosis Risk in Communities study, and Cardiovascular Health Studies

Laura M Stevens, BS[1,2], **Erik Linstead, PhD**[3], **Jennifer L Hall, PhD**[2], **David P Kao, MD**[1]

[1]University of Colorado Anschutz Medical School, Aurora CO

[2]Institute for Precision Cardiovascular Medicine at the American Heart Association, Dallas, TX

[3]Fowler School of Engineering, Chapman University Orange

Abstract

Background: Coronary heart disease (CHD), heart failure (HF) and stroke are complex diseases with multiple phenotypes. While many risk factors for these diseases are well known, investigation of as-yet unidentified risk factors may improve risk assessment and patient adherence to prevention guidelines. We investigated the diet domain in the Framingham Heart Study (FHS), Cardiovascular Heart Study (CHS), and Atherosclerosis Risk In Communities (ARIC) study to identify potential lifestyle and behavioral factors associated with CHD, HF, and stroke.

Methods: We used machine learning feature selection based on random forest analysis to identify potential risk factors associated with CHD, stroke, and HF in FHS. We evaluated the significance of selected variables using univariable and multivariable Cox proportional hazards analysis adjusted for known CV risks. Findings from FHS were then validated using CHS and ARIC.

Results: We identified multiple dietary and behavioral risk factors for CVD outcomes including marital status, red meat consumption, whole milk consumption, and coffee consumption. Amongst these dietary variables, increasing coffee consumption was associated with decreasing long-term risk of HF congruently in FHS, ARIC, and CHS.

Conclusions: Higher coffee intake was found to be associated with reduced risk of HF in all three studies. Further study is warranted to better define the role, possible causality, and potential mechanism of coffee consumption as a potential modifiable risk factor for HF.

BACKGROUND

Coronary heart disease (CHD), heart failure (HF), and stroke are amongst the top causes of death attributable to cardiovascular disease (CVD) in the United States[1]. Risk factors for CHD, HF and stroke have been previously identified and incorporated into predictive models

Correspondence: David Kao, 12700 East 19th Avenue. Campus box B-139, Aurora, CO 80045, david.kao@cuanschutz.edu, Phone: 303-724-8308.

DISCLOSURES: The authors have nothing to disclose.

to provide quantitative assessments for individual risk of developing disease and support development of personalized CVD prevention strategies[2–7]. Although widely used, these models consider a relatively limited set of patient characteristics, and there may remain as-yet unidentified risk markers, which could improve accuracy of risk prediction and possibly represent opportunities for improved CVD prevention[4,8,9]. Additional lifestyle and behavioral risk factors such as diet have been identified since the initial development of CVD risk models such as the Framingham heart score, which are used widely in primary care. Understanding and validation of the association of these factors in CVD has the potential to improve understanding of CV risk and aid in patient adherence to lifestyle and behavioral therapies[2].

Epidemiological studies like the Framingham Heart Study (FHS) collected thousands of patient characteristics. Validation and identification of risk factors for complex diseases such as CHD, HF, and stroke are made difficult by the large number and variety of potentially relevant patient characteristics such as comorbid conditions, lifestyle, and patient behavior. Traditionally, epidemiological statistical approaches use a hypothesis-driven framework to reduce the number of factors evaluated in a given model by selecting factors using clinical expertise. Although this approach is effective in focusing the analysis and reducing false discovery, clinical bias can impact the features evaluated, potentially excluding unanticipated predictors of disease.[10] Using machine learning-based feature selection to identify factors potentially important to disease risk can be advantageous because it allows for the assessment of large numbers of patient characteristics in a comparatively unbiased manner, reduces false positives, and can potentially pick up patterns that otherwise may be missed when using a hypothesis based approach.[11–14] The ability of machine learning methods to analyze very large sets of features in an automated fashion is one of the several advantages of machine learning that has fueled its adoption in data analytics. Longitudinal studies such as FHS, the Atherosclerosis Risk in Communities (ARIC) Study, and the Cardiovascular Health Study (CHS) enrolled thousands of patients with relatively high event rates over decades of follow-up. These study qualities provide an excellent foundation to explore utilizing machine learning to identify CVD risk factors beyond those used in current predictive models.

The objective of this analysis was to use supervised machine learning to identify potential variables important to assessing risk of incident CHD, stroke and HF in a hypothesis-free, data-driven manner. We then evaluated the significance of association between these features and CVD adjusting for known risk factors and validated these findings in the ARIC and CHS. We hypothesized that feature selection through supervised machine learning would identify potentially novel risk factors for CHD, HF, and stroke.

METHODS

Data sources

This study was completed under an approved expedited Colorado Multiple Institution Review Board protocol (#15–1193). FHS, CHS, and ARIC clinical data were obtained from the National Heart, Lung, and Blood Institutes Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC, Calverton, MD). Because of the sensitive nature of the data collected for this study, requests to access the dataset from qualified researchers trained in human subject confidentiality protocols completed through BioLINCC. FHS, CHS, and ARIC were prospective longitudinal cohort studies designed to investigate the incidence, survival rate, and determinants of CVD. All studies were community-based, investigated multiple incident CVD endpoints, and included at least 10 years of follow-up from the exam used as baseline for this analysis. Study design, response rates, and methodologies of each study are reported in detail elsewhere[15,16,17]. Participants in FHS (n=5209) were between the ages of 30 and 62 and assessed every 2 years. Participants who attended FHS Exam 14 and who had not yet had a CVD event were used in this analysis (n=2732). Exam 14 was used as the reference date for all time-to-event analyses. CHS participants (n=5888) were above the age of 65 and assessed annually for approximately 10 years. Of the 5888 CHS participants, the 3704 participants without prior CVD that contained complete data for the dietary factors identified during feature selection were used in analysis. ARIC (n=15792) enrolled individuals age 45–64 without prior CVD and consisted of 4 exams conducted every 3 years and a 5th exam approximately 25 years after enrollment. 14925 participants from the ARIC study contained data for follow up CVD events and were used for analysis. FHS Exam 14 occurred during 1975–1978 with mean follow-up of 16.7±9.8 (max = 36.3 years) thereafter. ARIC began in 1987 and CHS in 1989, indicating a temporal overlap between all 3 studies. Patients in CHS on average were older than patients in FHS and ARIC and had higher rates of comorbid conditions and outcome events. Patients in ARIC, while on average 10 years younger than FHS patients, had similar rates of smoking status, hypertension, and diabetes, but trended lower in prevalence of outcome events. Baseline characteristics among participant subgroups were compared using the chi-square and Mann-Whitney U tests for categorical and continuous variables, respectively (Table 1). All analyses were performed using the R statistical package (version 3.5; R Foundation for Statistical Computing, Vienna, Austria). Random forest analysis was completed using the *caret* package, and Cox proportional hazards analysis was completed using the *survival* package[18,19].

Outcomes

Outcomes of interest were time-to-incident CHD, HF, and stroke. All outcomes were adjudicated per individual study protocols, and data transformations for harmonization are given in Supplement Tables I–IV. Incident HF in FHS was defined using the Framingham Heart Failure Diagnostic Criteria[20]. In ARIC, incident HF was defined using International Classification of Disease-Clinical Modification (ICD9-CM) Codes upon hospital discharge. In CHS, HF was adjudicated based on ejection fraction (when available), signs, symptoms, clinical tests, and medical therapy. Code for the analyses performed are available from the first author available from the first author (laura.stevens@ucdenver.edu) upon reasonable request.

Feature Selection

We used FHS Exam 14 because it was the first clinical exam to include dietary variables, and it had been used in the development of prior FHS risk models[7]. In total, 222 variables were recorded at Exam 14, which included conventional variables such as age, sex, blood pressure, and others (Supplement Table I). Variables with missing data and excluded variables with >15% missing values were excluded and samples with complete cases for the remaining variables were used in feature selection[21]. Of the remaining 204 variables, 16 were dietary factors and 13 were non-dietary lifestyle behaviors. Participants were included in the analysis only if they had no missing data from any of the 204 variables.

Patient characteristics potentially important for predicting incident CHD, HF, and stroke were identified using random forest analysis. For optimal feature selection, we used 10-fold cross validation with 5 repeats[14,18,22]. Candidate variables for use in time-to-event analysis were the top 20% predictors based on importance metrics across all outcomes in the random forest model (Supplement Table I). The majority of non-dietary and non-lifestyle variables in the top 20% were collinearly related to known risk factors.

Evaluation of Feature Significance

Significance, magnitude and direction of association between candidate dietary factors and outcomes of interest were assessed using multivariable Cox proportional hazards analysis. The randomized nature of both the random forest methods applied and importance metric calculated does control for potential confounding and collinearity in machine learning experiments. Variables with importance scores in the top 20% were assessed for collinearity and compared with known risk factors when assessing which variables to include as covariates in the Cox proportional hazards analysis. The association between variables in top 20% importance and the outcomes were assessed for collinearity and strength of association with the outcome. Given that the majority of non-dietary risk factors with importance scores in the top 20% were collinearly related to known risk factors and some known risk factors were only modestly a associated with the outcome, we chose to use the FHS CVD risk score to provide good coverage of the probabilities for known risk factors while also accounting for known risks with weaker association[23]. Models for individual CHD, HF, and stroke outcomes were calibrated as presented in the original FHS CVD risk score publication by D'Agostino et. al[5]. We chose to use a risk score when performing multivariable analysis to account for collinearity and the impact of the combinations of known risk factors over the impact of each individual factor alone[24]. A p-value < 0.05 was considered significant throughout.

Validation

Baseline exams of CHS and ARIC were used for validation of the findings from FHS. Where possible, dietary factors significantly associated with outcomes of interest in FHS were harmonized with comparable variables in CHS and ARIC (Supplement Tables II–IV). Associations between dietary factors and clinical outcomes were then validated in CHS and ARIC. All outcomes, traditional risk factors, and dietary factors harmonized between FHS, ARIC and CHS for validation are given in Supplement Tables II–IV. The first and senior authors (Stevens, Kao) each have full access to all relevant data and take responsibility for the integrity of the analysis.

RESULTS

Feature Selection

Baseline characteristics of analyzed participants for all 3 studies are summarized in Table 1. The decision trees from the random forest models containing the 204 potential data measurements at Exam 14 with CHD, stroke, and HF outcomes were investigated to assess the importance of potential risk factors of CHD, stroke, and HF. There were 35 common risk factors across all outcomes that were ranked in the top 20% of important features by random forest analysis for either CHD, HF, or stroke (Table 2). Among these features were known risk factors such as blood pressure, age, and cholesterol. Given the potential for behavioral modification with dietary factors over non-modifiable factors such as the number of dead siblings, we selected dietary factors in the top 20% of variables ranked by importance to be evaluated further. Dietary factors including consumption of whole milk, red meat, eggs, alcohol, cheese, coffee and decaffeinated coffee were also ranked in the top 20% most important to risk of all CVD outcomes as were other lifestyle factors including marital status (Table 2). To evaluate and validate the association of coffee consumption with HF and stroke, the coffee consumption values from CHS and ARIC were converted to cups per day. For ARIC, coffee consumption was reported as rarely/never, a few cups/month, 1 cup/week, 2–4 cups/week, 5–6 cups/week, 2–3 cups/day, 4–5 cups/day, and >6 cups/day, which were transformed to 0, 0.07, 0.14, 0.43, 0.79, 1, 2.5, 5, and 6.5 cups/day respectively. CHS frequencies were never, 5–10 cups/year, 1–3 cups/month, 1–4 cups/week, and nearly every day, which were transformed 0, 0.021, 0.07, 0.36, and 1 cup/day respectively. The definition of what constituted red meat and level of consumption in FHS were ambiguous and could not be satisfactorily harmonized with ARIC and CHS preventing confident validation. Therefore, red meat consumption was not investigated beyond initial importance. All other dietary factors were further investigated to assess magnitude and direction of risk using univariable and multivariable analysis.

Evaluation of Feature Significance

Of the dietary factors identified by random forest that were evaluated using Cox proportional hazards analysis, coffee consumption was the only factor that remained significantly associated with any of the outcomes. Increasing caffeinated coffee consumption was found to be significantly associated with reduced risk of HF (hazard ratio (HR) = 0.95/cup/day, [95% CI 0.91–0.99], p = 0.02) and stroke (HR = 0.94/cup/day [95% CI 0.89–0.99], p = 0.02) but not CHD (p = 0.21) or CVD (p = 0.59). Adjusted for the FHS CVD risk score, increasing caffeinated coffee consumption remained significantly associated with decreased risk of HF (HR = 0.95/cup/day [95% CI 0.90–1.00], p=0.03) but not stroke (p=0.33).

Validation

Results of univariable and multivariable survival analysis for HF in all 3 studies are shown in Figure 1. In univariable analysis, increasing coffee consumption was significantly associated with decreased risk of HF in both CHS (HR = 0.86/cup/day [95% CI 0.78–0.96], p = 0.005) and ARIC (HR = 0.98/cup/day [95% CI 0.96–0.99], p = 0.048). When adjusted for the FHS risk scores, coffee consumption remained significantly associated with HF in CHS (HR 0.88/cup/day [95% CI 0.79–0.97], p = 0.01). In ARIC, coffee consumption showed a trend towards multivariable association between coffee consumption and HF (HR 0.98/cup/day [95% CI 0.96–1.00], p=0.06). To investigate dose response, participant characteristics according to quartiles of caffeinated coffee consumption (0/day, 1/day, 2/day, ≥3/day) are shown in Table 3. Compared with no coffee consumption, risk of HF was similar in participants drinking 1 cup/day (p= 0.19) but reduced in participants drinking 2 cups/day (HR = 0.69 [95% CI 0.55–0.87], p<0.001) and ≥3 cups/day (HR = 0.71 [95% CI 0.58–0.89], p<0.001). A dose response threshold for reduction in risk was not found, however higher consumption rates did not yield high enough sample sizes to individually higher coffee consumption levels, and CHS did not report coffee consumption with enough granularity to categorize participants into consumption beyond 1 cup/day.

To investigate the possible role of caffeine in association between coffee consumption and HF risk, we performed additional analyses with respect to decaffeinated coffee consumption, (FHS and CHS) and caffeine intake (FHS, CHS, and ARIC). Caffeinated vs decaffeinated tea consumption was not separated in FHS Exam 14 or baseline CHS, and therefore was not considered. Data from FHS Exam 20 (n=867) and CHS Exam 8 (n=1903) and were used because they were the first to report estimated caffeine consumption.

Decaffeinated coffee consumption was significantly associated with *increased* risk of HF in multivariable analysis in FHS (HR 1.10/cup/day [95% CI 1.03–1.17], p=0.004) but not in CHS (p=0.63). All 3 studies showed a concordant inverse relationship between caffeine intake in 100 mg doses (1 cup coffee or 2 cups black tea) and risk of HF. In FHS, increased caffeine consumption was found to be significantly associated with reduced risk of HF in both univariable (HR 0.93/100 mg caffeine [95% CI 0.86–0.98], p=0.02) and multivariable analyses (HR = 0.92/100 mg caffeine [95% CI 0.86–0.98], p=0.01.) In CHS, caffeine consumption was significantly associated with reduced risk of HF in univariable analyses (HR 0.96/100 mg caffeine [95% CI 0.92–0.99], p=0.02) and showed a trend towards reduced HF risk in multivariable analysis (HR 0.97/100 mg caffeine [95% CI 0.93–1.00], p=0.07). In ARIC, increased caffeine consumption was also associated with significantly reduced risk of HF in univariable (HR 0.98/100 mg caffeine [95% CI 0.97–0.99], p=0.01) and multivariable analyses (HR 0.99/100 mg caffeine [95% CI 0.97–0.99], p=0.049).

DISCUSSION

HF incidence, HF hospitalizations and societal costs continue to increase despite decreasing CHD and stroke mortality rates[1,25]. Although much is known about modifiable risk factors for ischemic CVD, opportunities for reducing HF incidence are less clear, likely in part because a substantial fraction of HF is nonischemic in etiology. Using random forest feature selection applied to data from FHS, we found multiple dietary and behavioral risk factors including marital status, red meat consumption, whole milk consumption, and coffee consumption that may be associated with CHD, HF, or stroke. Evaluation of these features showed that people who reported higer coffee consumption rates were associated with decreased long-term risk of HF concordantly in FHS in ARIC, and CHS. Previous studies primarily focused on composite CVD outcomes or CHD and CVD mortality, whereas relatively few studies have reported an association between coffee consumption and HF risk. This analysis expands those findings to include the relationship between decreased risk of

HF and higher coffee consumption. The mechanism of this association is unclear, but limited analysis in FHS and CHS suggested that caffeine may be an important contributor. The pervasive consumption of coffee in modern society and the high potential for dietary modification that could reduce HF risk suggest further exploration of the role of caffeine and coffee in development of HF is warranted.

Caffeinated coffee consumption and reduced risk of CHD mortality has been previously reported in elderly participants without hypertension[26,27]. In FHS, elderly individuals who consumed any caffeinated coffee had a 43% reduction in CHD deaths compared with those who never consumed coffee. Similarly, an analysis of NHANES revealed that individuals ≥65 who did not have severe hypertension also had a dose-dependent decrease in CVD and CV mortality associated with higher coffee consumption[26,27]. Another prospective epidemiologic study found that consumption of coffee, green tea, and oolong tea, and total caffeine intake was associated with reduced stroke and mortality from CVD in Japanese men and women[28]. More recently, a review of 201 meta-analyses found that increasing daily coffee consumption was associated with decreased CVD mortality and all-cause mortality[29]. A systematic review of 351 observational studies of healthy adults, adolescents, and pregnant women found that consumption of 400 mg (4 cups of coffee or 8 cups of black tea) of caffeine/day or less was not associated with cardiovascular toxicity in adults[30]. A meta-analysis of 53 studies found a non-linear association between long-term coffee consumption and CVD risk, where 3–5 cups/day of coffee was significantly associated with decreased CVD risk, compared with none, light (1–2 cups/day), or heavy (≥6 cups/day) caffeinated coffee consumption. However, the authors speculated that heavy coffee consumption analysis may have been confounded by increased smoking rates, and decaffeinated coffee consumption was not associated with elevated CVD risk[31]. Finally, a meta-analysis incorporating 5 prospective studies comprised of 6,522 HF events and 140,220 participants, and investigating HF risk and coffee (caffeinated and decaffeinated) consumption, observed a statistically significant J-shaped relationship between coffee and HF. Compared with no consumption, the strongest inverse association was seen for 4 servings/day with a potentially higher risk above this level of consumption[32]. When considering doses of 1, 2, and ≥ 3 cups per day in this analysis, we did not observe a similar J-shaped curve. Coffee contains higher amounts of caffeine than any other dietary product in addition to containing many other constituents such as potassium, niacin, magnesium, or tocopherols that could contribute for this association[33]. Our results support caffeine is in fact an important contributor given that increased estimated caffeine consumption irrespective of source was associated with decreased HF risk in all three studies.

Increasing decaffeinated coffee consumption was associated with increased risk of HF, although this was only shown in FHS. As with prior studies, interpretation of the association between decaffeinated coffee and incident CVD was limited by much lower reported coffee consumption[31]. Association between increasing decaffeinated coffee consumption and increased CVD risk could also be due to unobserved latent or confounding factors, such as individuals with other CVD-risk factors switching from regular to decaffeinated coffee or concomitant high risk behaviors such as smoking[31,34]. Additionally, methods for decaffeinating coffee can involve the addition harmful chemicals, which could be affecting the association between increased decaffeinated coffee consumption and increased risk of HF[35].

The potential of intentional higher coffee consumption as a means of reducing HF risk cannot be determined from this analysis. It remains possible that coffee consumption is a marker or proxy for another behavior or dietary factor that reduces HF risk. Consequently, intentional or prescribed increase in coffee intake for the purposes of reducing HF risk cannot be recommended based on our results. However, the pervasive popularity of coffee worldwide, suggests great potential for reducing CVD risk through dietary modification if the association is true and highlights the importance of future clinical studies to validate these observations.

Limitations

The observational and retrospective nature of the data, much of which relies on patient recall introduces significant uncertainty regarding data quality and the strong possibility of unmeasured confounders. For example, the data specifically does not distinguish between the type of coffee consumed (e.g. type of bean/organic/mold content), use of additives (e.g. sugar or creamer), brewing method (e.g. drip vs. espresso), or timing of consumption (e.g. with breakfast vs. after dinner/before bedtime), which could further impact the associations between coffee and clinical outcomes. Associations between coffee consumptions and key CVD risk factors not present in the Framingham risk models could also impact results. Data regarding caffeine intake was estimated based on patient-reported dietary intake and not collected uniformly. Because correlation does not imply causality, a prospective randomized or cohort control trial would ideally be used to validate these findings.

CONCLUSION

Machine learning feature selection identified coffee consumption as an important risk factor for subsequent development of HF. Higher coffee consumption and caffeine intake were associated with reduced risk of HF in 3 large, well-known epidemiologic studies, although decaffeinated coffee was not. Further study is warranted to better define the mechanism and role of coffee consumption as a potential modifiable risk factor for HF.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS:

The authors would like to thank Carsten Görg and Chris Roeder for there assistance in preparing this manuscript.

FUNDING: DPK: NHLBI K08HL125725; AHA 17IG33660301.

Abbreviations

ARIC	Atherosclerosis Risk In Communities
CHD	Coronary Heart Disease

CVD	Cardiovascular Disease
CHS	Cardiovascular Health Study
FHS	Framingham Heart Study
HF	Heart Failure

REFERENCES

1. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Das SR, et al. Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. Circulation. 2019;139:e56–e528. [PubMed: 30700139]
2. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2014;129:S49–S73. [PubMed: 24222018]
3. Brindle PM, McConnachie A, Upton MN, Hart CL, Smith GD, Watt GC. The accuracy of the Framingham risk-score in different socioeconomic groups: a prospective study. Br J Gen Pract. 2005;55:838–845. [PubMed: 16281999]
4. Lloyd-Jones DM. Strengths and Limitations of the ASCVD Risk Score and What Should Go in the Risk Discussion. Am Coll Cardiol [Internet]. 2014 [cited 2020 Dec 30];Available from: https://www.acc.org/latest-in-cardiology/articles/2014/08/25/14/48/strengths-and-limitations-of-the-ascvd-risk-score-and-what-should-go-in-the-risk-discussion
5. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117:743–753. [PubMed: 18212285]
6. Kannel WB, D'Agostino RB, Silbershatz H, Belanger AJ, Wilson PWF, Levy D. Profile for Estimating Risk of Heart Failure. Arch Intern Med. 1999;159:1197. [PubMed: 10371227]
7. D'Agostino RB, Wolf PA, Belanger AJ, Kannel WB. Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study. Stroke. 1994;25:40–43. [PubMed: 8266381]
8. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, Lassale CM, Siontis GCM, Chiocchia V, Roberts C, Schlüssel MM, Gerry S, Black JA, Heus P, Schouw YT van der, Peelen LM, Moons KGM. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416. [PubMed: 27184143]
9. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc JAMIA. 2017;24:198–208. [PubMed: 27189013]
10. Oakden-Rayner L, Carneiro G, Bessen T, Nascimento JC, Bradley AP, Palmer LJ. Precision Radiology: Predicting longevity using feature engineering and deep learning methods in a radiomics framework. Sci Rep. 2017;7:1–13. [PubMed: 28127051]
11. Awada W, Khoshgoftaar TM, Dittman D, Wald R, Napolitano A. A review of the stability of feature selection techniques for bioinformatics data. In: 2012 IEEE 13th International Conference on Information Reuse Integration (IRI). 2012. p. 356–363.
12. Breiman L Random Forests. Mach Learn. 2001;45:5–32.
13. Syarif I, Zaluska E, Prugel-Bennett A, Wills G. Application of Bagging, Boosting and Stacking to Intrusion Detection. In: Perner P, editor. Machine Learning and Data Mining in Pattern Recognition. Berlin, Heidelberg: Springer; 2012. p. 593–602.
14. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23:2507–2517. [PubMed: 17720704]
15. Framingham Heart Study: https://www.framinghamheartstudy.org/
16. The Atherosclerosis Risk in Communities Study (ARIC), NHLBI Obesity Research - NHLBI, NIH. [cited 2017 Nov 9];Available from: https://www.nhlbi.nih.gov/research/resources/obesity/population/aric.htm

17. Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A, et al. The cardiovascular health study: Design and rationale. Ann Epidemiol. 1991;1:263–276. [PubMed: 1669507]

18. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, Core Team R, et al. caret: Classification and Regression Training [Internet]. 2017 [cited 2018 Mar 5]. Available from: https://CRAN.R-project.org/package=caret

19. Therneau TM, until 2009) TL (original S->R port and maintainer. survival: Survival Analysis [Internet]. 2017 [cited 2018 Mar 5]. Available from: https://CRAN.R-project.org/package=survival

20. McKee PA, Castelli WP, McNamara PM, Kannel WB. The Natural History of Congestive Heart Failure: The Framingham Study. N Engl J Med. 1971;285:1441–1446. [PubMed: 5122894]

21. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, Petersen I. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol. 2017;9:157–166. [PubMed: 28352203]

22. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016;18:e323. [PubMed: 27986644]

23. Arbogast PG, Ray WA. Performance of Disease Risk Scores, Propensity Scores, and Traditional Multivariable Outcome Regression in the Presence of Multiple Confounders. Am J Epidemiol. 2011;174:613–620. [PubMed: 21749976]

24. Concato J The Risk of Determining Risk with Multivariable Models. Ann Intern Med. 1993;118:201. [PubMed: 8417638]

25. Folsom AR, Yamagishi K, Hozawa A, Chambless LE. Absolute and Attributable Risks of Heart Failure Incidence in Relation to Optimal Risk Factors. Circ Heart Fail. 2009;2:11–17. [PubMed: 19808310]

26. A Greenberg J, Chow G, C Ziegelstein R. Caffeinated Coffee Consumption, Cardiovascular Disease, and Heart Valve Disease in the Elderly (from the Framingham Study). Am J Cardiol. 2009;102:1502–8.

27. Greenberg JA, Dunbar CC, Schnoll R, Kokolis R, Kokolis S, Kassotis J. Caffeinated beverage intake and the risk of heart disease mortality in the elderly: a prospective analysis. Am J Clin Nutr. 2007;85:392–398. [PubMed: 17284734]

28. Mineharu Y, Koizumi A, Wada Y, Iso H, Watanabe Y, Date C, Yamamoto A, Kikuchi S, Inaba Y, Toyoshima H, et al. Coffee, green tea, black tea and oolong tea consumption and risk of mortality from cardiovascular disease in Japanese men and women. J Epidemiol Community Health. 2011;65:230–240. [PubMed: 19996359]

29. Poole R, Kennedy OJ, Roderick P, Fallowfield JA, Hayes PC, Parkes J. Coffee consumption and health: umbrella review of meta-analyses of multiple health outcomes. BMJ. 2017;359:j5024. [PubMed: 29167102]

30. Wikoff D, Welsh BT, Henderson R, Brorby GP, Britt J, Myers E, Goldberger J, Lieberman HR, O’Brien C, Peck J, et al. Systematic review of the potential adverse effects of caffeine consumption in healthy adults, pregnant women, adolescents, and children. Food Chem Toxicol. 2017;109:585–648. [PubMed: 28438661]

31. Ding M, Bhupathiraju SN, Satija A, van Dam RM, Hu FB. Long-Term Coffee Consumption and Risk of Cardiovascular Disease: A Systematic Review and a Dose-Response Meta-Analysis of Prospective Cohort Studies. Circulation. 2014;129:643–659. [PubMed: 24201300]

32. Mostofsky E, Rice MS, Levitan EB, Mittleman MA. Habitual Coffee Consumption and Risk of Heart Failure: A Dose–Response Meta-Analysis. Circ Heart Fail. 2012;5:401–405. [PubMed: 22740040]

33. You D-C, Kim Y-S, Ha A-W, Lee Y-N, Kim S-M, Kim C-H, Lee S-H, Choi D, Lee J-M. Possible Health Effects of Caffeinated Coffee Consumption on Alzheimer’s Disease and Cardiovascular Disease. Toxicol Res. 2011;27:7–10. [PubMed: 24278543]

34. Dusseldorp M van, Smits P, Thien T, Katan MB. Effect of decaffeinated versus regular coffee on blood pressure. A 12-week, double-blind trial. Hypertension. 1989;14:563–569. [PubMed: 2680964]

35. Ramalakshmi K, Raghavan B. Caffeine in Coffee: Its Removal. Why and How? Crit Rev Food Sci Nutr. 1999;39:441–456. [PubMed: 10516914]

Summary:

- Little is known about the risk of developing heart failure associated with dietary components and the potential benefits of changing intake of specific foods.
- By using machine learning in the Framingham study, we identified several dietary factors that might be associated risk of heart failure.
- We found in three large, well-known studies (the Framingham, Cardiovascular Health and Atherosclerosis Risk in Communities studies) that increased coffee consumption appeared to correlate with reduced risk of developing heart failure later in life.
- Additional work is needed to determine whether modulating coffee intake could affect future risk of developing heart failure.

Clinical Impact Statement:

- Controlling for known risk factors, increased coffee consumption was found to be associated with reduced risk of HF in three large longitudinal epidemiologic studies (FHS, CHS and ARIC). The mechanism of this association is unclear, but limited analysis in FHS and CHS suggest caffeine may be an important contributor. The high prevelance of coffee consumption in society suggests further study is warranted to better define the role, possible causality, and potential mechanism of coffee and caffeine consumption as potential modifiable risk factors for HF.

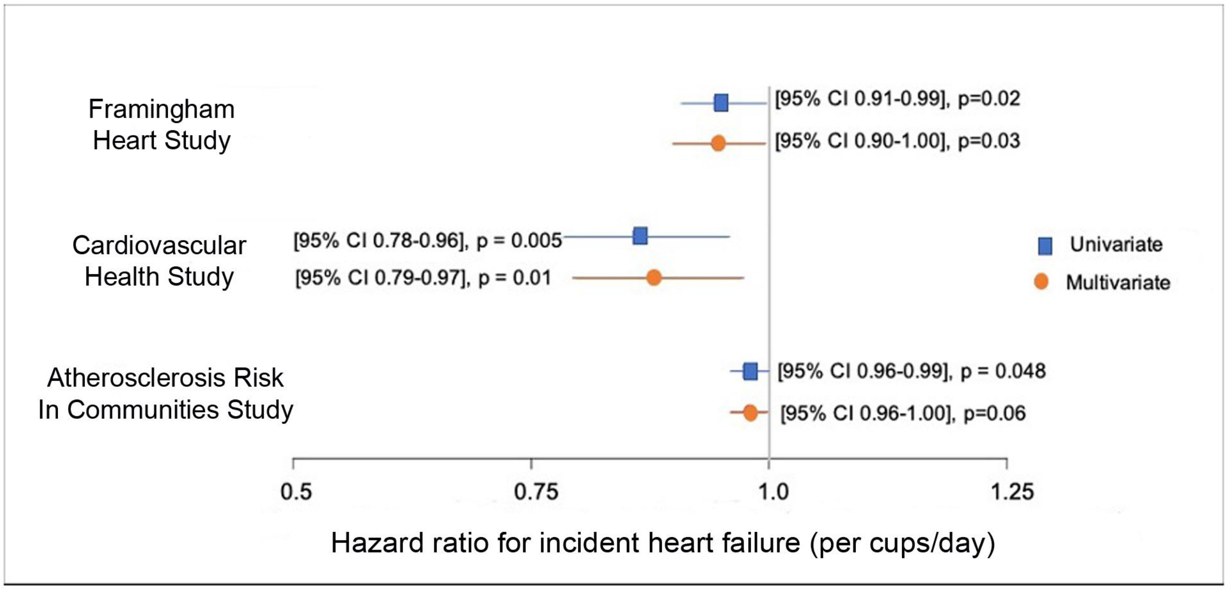

Figure 1:

Association between coffee and incident heart failure in the Framingham Heart Study (FHS), Cardiovascular Health Study (CHS), Atherosclerosis Risk in Communities (ARIC) study.

Table 1 –

Overview statistics of coffee consumption and known risk factors of CVD, N (%) and median [quartile 1-quartile 3]

	Framingham Exam 14 (N=2732)	ARIC Baseline (N=14925)	CHS Baseline (N=3704)
Female	1602 (59)	8153 (55)	1625 (49)
Age, years	66 [61–73]	54 [49–59]	71 [67–75]
Systolic BP, mg/dL	137 [125–150]	119 [108–131]	134 [121–149]
HDL cholesterol, mg/dL	48 [39–58]	48 [39–61]	50 [42–62]
Total Cholesterol, mg/dL	227 [202–258]	212 [186–239]	210 [185–235]
BMI, kg/m^2	25.9 [23.5–28.7]	26.89 [24.0–30.4]	26.07 [23.6–29.0]
CVD Risk Score, N (%)	10.2 [5.9–17.5]	4.18 [2.2–7.8]	11.03 [6.7–17.6)
Hypertension, N (%)	880 (32)	3812 (26)	1772 (47)
Current smoking, N (%)	820 (30)	3951 (26)	385 (10)
Diabetes mellitus, N (%)	176 (7)	1509 (10)	604 (16)
Coffee Intake, cups/day- median (Q1-Q3)	2 (1–3)	1 (0–2.5)	0.36 (0–1)
Incident outcomes			
CVD, N (%)	1172 (43)	4401 (30)	2736 (73)
HF, N (%)	625 (23)	2324 (17)	1698 (46)
Stroke, N (%)	461 (17)	1127 (8)	1147 (31)
CHD, N (%)	706 (31)	3033 (20)	2199 (59)

[*] $p < 0.001$ for all characteristics

BP = Blood pressure. HDL = High density lipoprotein BMI = Body Mass Index CVD = Cardiovascular disease HF = Heart failure CHD = Coronary heart disease

Table 2 -

Variables ranked in top 20% from random forest decision trees and importance score for all outcomes: CVD, CHD, stroke and HF (displayed in ranked order)

Variable	Description
FG311	SUGAR-EXAM 14
FG313	CHOLESTEROL-EXAM 14
FG72	SBP-PHYSICIAN-2ND-EXAM 14
FG62	WEIGHT-EXAM 14
FG70	SBP-PHYSICIAN-1ST-EXAM 14
FG68	SBP-NURSE-EXAM 14
FG234	VENTRICULAR-RATE-MIN-EXAM 14
FG69	DBP-NURSE-EXAM 14
FG53	AGE-EXAM 14
FG312	CREATININE-EXAM 14
FG73	DBP-PHYSICIAN-2ND-EXAM 14
FG310	HEMATOCRIT-EXAM 14
FG63	HEIGHT IN INCHES
FG239	AQRS-EXAM 14
FG71	DBP-PHYSICIAN-1ST-EXAM 14
FG235	P-R-INTERVAL-EXAM 14
FG122	**RED-MEAT-WEEK-EXAM 14**
FG114	**COFFEE-CUPS-DAY-EXAM 14**
FG120	**COCKTAILS-WEEK-EXAM 14**
FG237	QT-INTERVAL-EXAM 14
FG115	**COFFEE-DECAF-CUPS-DAY-EXAM 14**
FG121	**EGGS-WEEK-EXAM 14**
FG320	**NO-OF-BROTHER-DEAD-EXAM 14**
FG124	**WHOLE-MILK-WEEK-EXAM 14**
FG116	**TEA-CUPS-DAY-EXAM 14**
FG123	**CHEESE-WEEK-EXAM 14**
FG236	QRS-INTERVAL-EXAM 14
FG271	FUNCTIONAL-CLASS-EXAM 14
FG119	**WINE-WEEK-EXAM 14**
FG257	ECG-CLINICAL-READING-EXAM 14
FG321	NO-OF-SISTER-DEAD-EXAM 14
FG99	OTHER-MEDICINES-EXAM 14
FG190	SYSTOLIC-MUR-VALVE-EXAM 14
FG170	CORNEAL-ARCUS-EXAM 14
FG118	**BEER-WEEK-EXAM 14**
FG319	PH PH 8 or 9
FG104	CIGARETTES-DAY-EXAM 14
FG138	CHEST-DISCOMFORT-EXAM 14

Variable	Description
FG125	**MARGARINE-VS-BUTTER-EXAM 14**
FG258	HYPERTENSIVE-STATUS-EXAM 14
FG58	MARITAL-STATUS-EXAM 14

Table 3 –

Clinical characteristics and outcomes of coffee consumption and known risk factors of CVD by quartile of coffee consumption, N (%) and median [quartile 1-quartile 3]

	0 cups/day 8809 (41)	1 cup/day 5130 (24)	2 cups/day 4056 (19)	≥3 cups/day 3347 (16)
Female	4874 (55)	2719 (53)	2106 (52)	1648 (49)
Age, years	59 (51–67)	62 (54–69)	56 (50–61)	56 [50–61]
Systolic BP, mg/dL	125 (112–139)	127 (115–143)	121 (109–134)	119 [108–133]
HDL cholesterol, mg/dL	49 (39–61)	50 (40–62)	49 (39–61)	48 [38–59]
Total Cholesterol, mg/dL	212 (187–239)	214 (188–242)	214 (188–241)	217 [190–244]
BMI, kg/m^2	26.8 (24.0–30.4)	26.7 (23.9–30.1)	26.4 (23.8–29.5)	26.1 [23.5–29.2]
CVD Risk Score, (%)	5.6 (2.8–11.0)	6.9 (3.5–12.8)	4.5 (2.4–8.9)	5.1 [2.6–9.3]
Hypertension, (%)	3040 (35)	1847 (36)	959 (24)	592 (18)
Current smoking, (%)	1529 (17)	1139 (22)	1134 (28)	1339 (40)
Diabetes mellitus, (%)	1112 (13)	644 (13)	329 (8)	193 (6)
Incident outcomes				
CVD, (%)	3546 (40)	2305 (45)	1262 (31)	1145 (34)
HF, (%)	2037 (23)	1387 (27)	627 (15)	566 (17)
Stroke, (%)	1242 (14)	824 (16)	355 (9)	296 (9)
CHD, (%)	2636 (30)	1643 (32)	865 (21)	757 (23)

* p < 0.001 across all quartiles

BP = Blood pressure. HDL = High density lipoprotein BMI = Body Mass Index CVD = Cardiovascular disease HF = Heart failure. CHD = Coronary heart disease