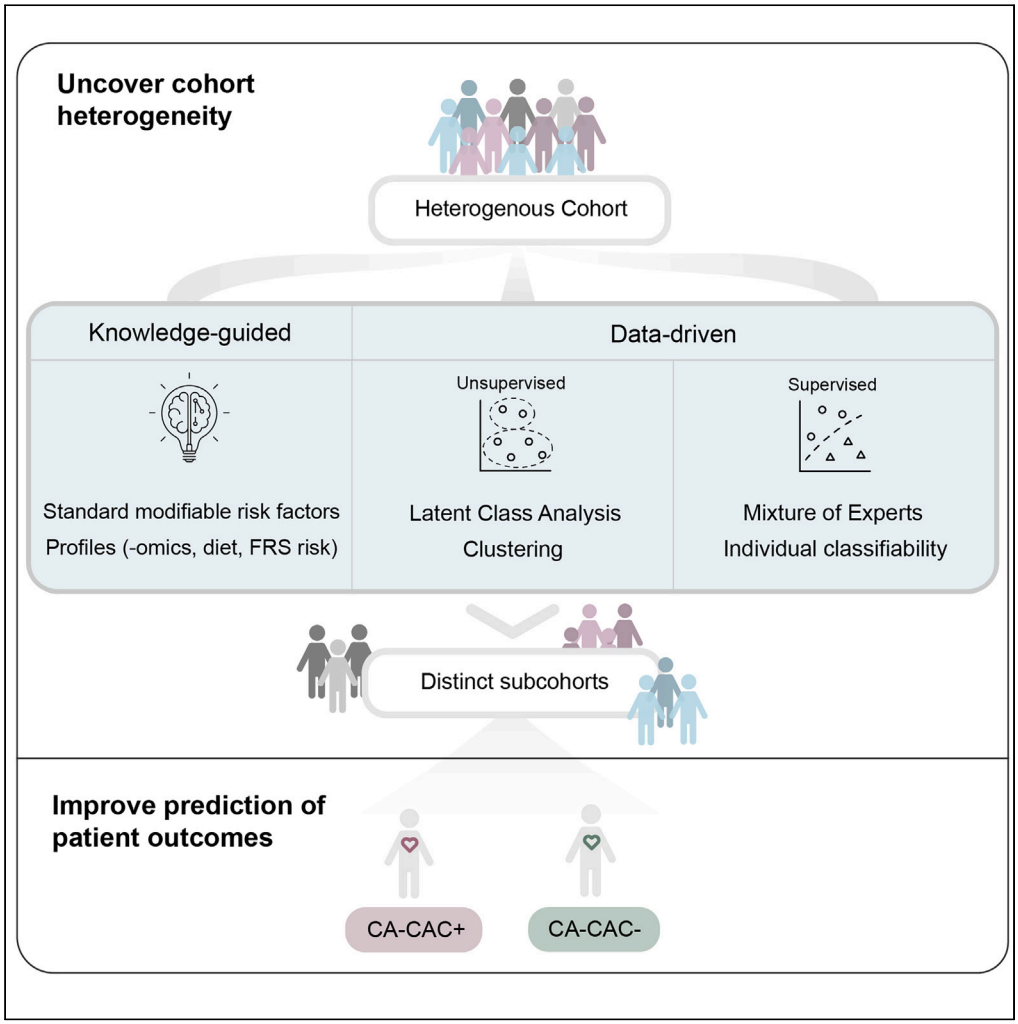
iScience

CellPress
OPEN ACCESS

Article

Overcoming cohort heterogeneity for the prediction of subclinical cardiovascular disease risk



Adam S. Chan, Songhua Wu, Stephen T. Vernon, ..., Tongliang Liu, Jean Y.H. Yang, Ellis Patrick

jean.yang@sydney.edu.au (J.Y.H.Y.)
ellis.patrick@sydney.edu.au (E.P.)

Highlights

Metabolomes of 837 subjects from the BioHEART-CT study were profiled for CAD risk

We used knowledge-guided and data-driven methods to identify distinct subcohorts

Modeling of subclinical CAD within subcohorts improves overall prediction accuracy

Chan et al., iScience 26, 106633
May 19, 2023
https://doi.org/10.1016/j.isci.2023.106633

iScience

CellPress
OPEN ACCESS

Article

Overcoming cohort heterogeneity for the prediction of subclinical cardiovascular disease risk

Adam S. Chan,[1,2,3] Songhua Wu,[4] Stephen T. Vernon,[6] Owen Tang,[2,6] Gemma A. Figtree,[2,6] Tongliang Liu,[3,4] Jean Y.H. Yang,[1,2,3,7,*] and Ellis Patrick[1,3,5,7,8,*]

SUMMARY

Cardiovascular disease remains a leading cause of mortality with an estimated half a billion people affected in 2019. However, detecting signals between specific pathophysiology and coronary plaque phenotypes using complex multi-omic discovery datasets remains challenging due to the diversity of individuals and their risk factors. Given the complex cohort heterogeneity present in those with coronary artery disease (CAD), we illustrate several different methods, both knowledge-guided and data-driven approaches, for identifying subcohorts of individuals with subclinical CAD and distinct metabolomic signatures. We then demonstrate that utilizing these subcohorts can improve the prediction of subclinical CAD and can facilitate the discovery of novel biomarkers of subclinical disease. Analyses acknowledging cohort heterogeneity through identifying and utilizing these subcohorts may be able to advance our understanding of CVD and provide more effective preventative treatments to reduce the burden of this disease in individuals and in society as a whole.

INTRODUCTION

Cardiovascular diseases (CVD) remain the leading cause of death globally, with an estimated 523 million prevalent cases globally in 2019.[1] While physicians and individuals have traditionally relied on quantifying cardiovascular risk through identifying and targeting well-recognized standard modifiable cardiovascular risk factors (SMuRFs: hypertension, diabetes mellitus, dyslipidemia, and smoking), there is still a substantial unmet need to elucidate the biological mechanisms of individual susceptibility to these risk factors. Recently, the BioHEART-CT study team has aimed to address the challenge of discovering standard blood-based biomarkers that reflect an individual's vascular response to risk factors or signals the early development of atherosclerosis,[2–4] through quantifying subclinical coronary artery disease (CAD) by CT coronary angiography and pairing this with analysis of blood samples and outcome data.[4] This is possible with the development of high-throughput technologies that have facilitated the detailed characterization of metabolite signatures and their associations with cardiovascular disease in large cohorts of individuals.[5,6]

While we have begun to unravel metabolic mechanisms and detect associations with cardiovascular risk using large cohort studies, the heterogeneity embedded in these cohorts makes detection of these relationships a challenging task. The effective analysis of such large heterogeneous cohorts necessitates the consideration of distinctive subcohorts of individuals present within the larger cohort,[7,8] resulting in improved understanding of biological mechanisms and prediction of CVD. This approach is supported by recent studies such as the detection of race-specific metabolite associations with incident coronary heart disease[9] and the improvement in therapies for heart failure through the detection of distinct phenotypes differing significantly in outcomes.[10] In addition, improved invasive and non-invasive imaging technologies have allowed increasing appreciation of the different subtypes of atherosclerotic plaque that develop, each with likely distinct dysregulated signaling pathways, highlighting the importance of not "bucketing" CAD as a singular entity.

Recognition of cohort heterogeneity within study populations is already an existing idea, but only more recently have investigators sought to define cohort heterogeneity[11,12] and utilize it in their analyses. For

[1]School of Mathematics and Statistics, The University of Sydney, Sydney, NSW, Australia

[2]Charles Perkins Centre, The University of Sydney, Sydney, NSW, Australia

[3]Sydney Precision Data Science Centre, The University of Sydney, Sydney, NSW, Australia

[4]School of Computer Science, The University of Sydney, Sydney, NSW, Australia

[5]Westmead Medical Institute, Sydney, NSW, Australia

[6]Kolling Institute of Medical Research, Royal North Shore Hospital, Sydney, NSW, Australia

[7]Senior author

[8]Lead contact

*Correspondence: jean.yang@sydney.edu.au (J.Y.H.Y.), ellis.patrick@sydney.edu.au (E.P.)

https://doi.org/10.1016/j.isci.2023.106633

instance, clustering analysis has been used to determine phenotypic heterogeneity in patients with unique characteristics to better inform patient care for a broad range of diseases,[13–15] and data mining techniques were used to discover homogeneous subgroups within a disease population to make more precise predictions.[16–18] Given the complexity of CVD, including interactions with many comorbidities and predispositions, it is not surprising that the pool of individuals that develop CVD are diverse with differing risk factors and characteristics. This diversity, which can be framed as cohort heterogeneity, is important to elucidate as it has large implications on downstream metabolomic analyses.[19]

Several different approaches can be useful to help understand and unravel heterogeneity present in complex datasets. We categorize these approaches as *knowledge-guided* or *data-driven* approaches. Knowledge-guided approaches, having prior contextual understanding of heterogeneity, include defining cohorts using clinical characteristics of individuals such as sex[20–23] and age[6,24]; and stratification using established or coexisting individual profiles such as diet profiles[25] and omic profiles.[26,27] Data-driven approaches for uncovering cohort heterogeneity can be divided into *unsupervised* techniques which ignore the outcome being studied, and *supervised* techniques, which include information about the outcome. Unsupervised techniques include clustering[10,28,29] and latent class analysis[30,31] of clinical phenotypes; supervised techniques include mixture of experts (MoE)[32–35] and subgroup discovery algorithms[17,36] to subdivide large cohorts. Following stratification of individuals and identification of heterogeneity present, techniques including hierarchical modeling of subcohorts and ensemble learning can then be employed to improve the prediction of CAD in the whole cohort.

In this paper, we: (1) demonstrate that modeling distinct subcohorts of individuals improves prediction of CAD, (2) elucidate and discover several subcohorts existing in the data and compare the methods to derive these using clinical variables, (3) and further suggest techniques which incorporate heterogeneity in modeling of metabolomics and discuss how these can be utilized.

RESULTS

BioHEART study for identifying CAD risk using metabolomics

Complex heterogeneity in large patient cohorts can be uncovered through a range of approaches, generating subcohorts for subsequent modeling of subclinical CAD. The metabolomics profile of 837 individuals from the BioHEART study was measured along with their clinical variables (including demographics, risk factors, and medications) and CAD status (Figure 1A). We defined the CAD status of a patient using coronary artery calcium scores (CACS), whether they had clinically actionable CACS (CA-CAC+) or not (CA-CAC-). This is a biobanked cohort study designed for discovery of new omics signatures of subclinical CAD, trained off CT coronary angiography images of plaque burden.[4] Clinical variables were used to determine subcohorts of individuals in both knowledge-guided and data-driven approaches, with supervised data-driven approaches also utilizing individual CAD status and metabolomics to subcohort individuals (Figure 1B). We compared six different approaches for subcohort identification from the various knowledge-guided and data-driven approaches outlined in Table S1. Distinct subcohorts found by these approaches can then be used to model CA-CAC+ using metabolites to increase overall performance (Figure 1C).

We constructed models predicting CA-CAC using metabolomics data in a cohort of 837 individuals from the BioHEART-CT study. To assess the performance of the machine learning models, the modeling was performed in a discovery cohort (n = 512) and assessed in a validation cohort (n = 325). Individual demographics, risk factors, and regular medication use of these cohorts are presented in Table 1. Both cohorts had similar proportions of: CA-CAC+ (32.8% in discovery vs. 36.6% in validation, p = 0.3), age (mean = 60, sd = 12 vs. mean = 62, sd = 12, p = 0.11), and SMuRFs (p = 0.07). The discovery cohort had a higher proportion of males (58.6% vs. 49.5%, p = 0.01) than the validation cohort. The discovery cohort also had lower proportions of statin users (30.9% vs. 37.5%, p = 0.046), hypertension (35.7% vs. 43.1%, p = 0.034), and diabetes mellitus (7.2% vs. 11.4%, p = 0.039). Ideally, all models constructed should be robust to the similarities and differences between the two cohorts.

In the overall cohort, multivariate models constructed using clinical and demographic features to predict CA-CAC+ performed better than models constructed with metabolomics data only. When all clinical and demographic features were modeled using a lasso logistic regression model in the discovery cohort (without subgroups) to predict CA-CAC+, a cross-validated mean AUC of 0.74 was obtained in the

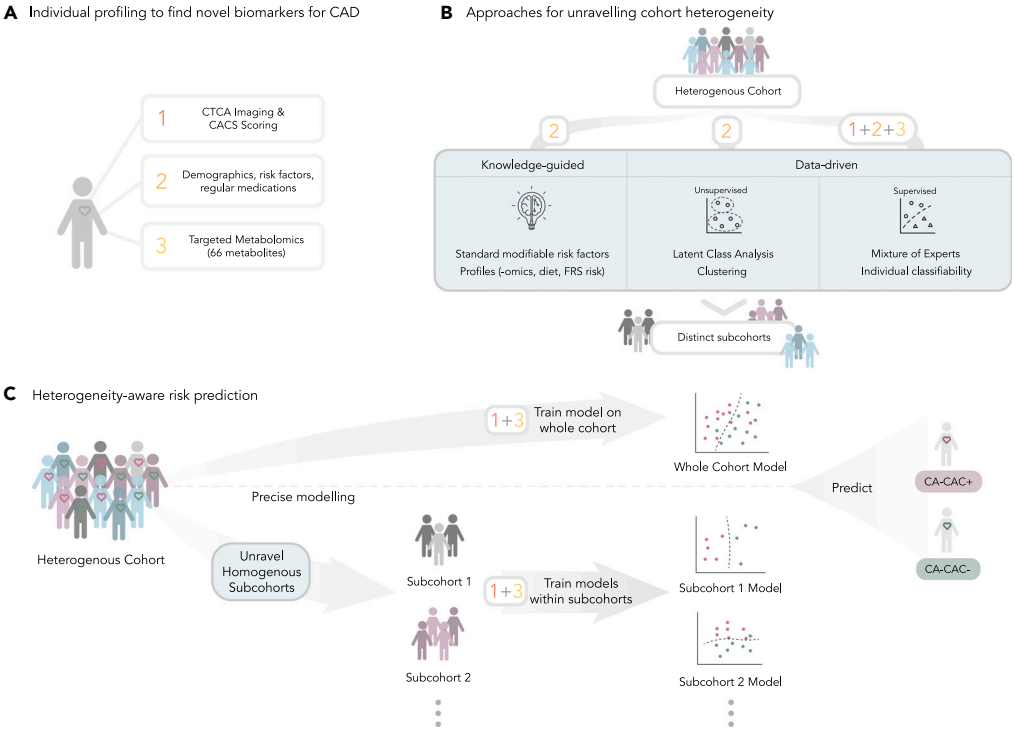


Figure 1. Schematic of approaches for unraveling cohort heterogeneity and subsequent heterogeneity-aware prediction

(A) Data variables measured for analyzed individuals.

(B) Methods for extracting subcohorts categorized by knowledge-guided and data-driven approaches.

(C) Workflow for heterogeneity-aware prediction of CA-CAC+ through modeling within subcohorts. Note: for Mixture of experts, clinical variables (2) are additionally required for model training along with CA-CAC status (1) and metabolomics (3).

discovery cohort and an AUC of 0.69 in the validation cohort (Figure S1A). The most predictive features were age, number of SMuRFs, use of statin, and use of angiotensin-converting enzyme (ACE) inhibitors or angiotensin receptor blockers (ARB) medications (Figure S1B). Alternatively, models constructed using metabolite-to-metabolite ratios had a cross-validated mean AUC of 0.58 in the discovery set and an AUC of 0.57 in the validation set (Figure S1A). While it is clear that models constructed with clinical and demographic features are more predictive of CA-CAC+, their primary features of age and sex are unmodifiable. Models using metabolomics data accurately predict a different subset of individuals and classify different patients as CA-CAC+ compared to the clinical model (Figure S1C), indicating they may be complementary. This provided an appropriate context for interpreting more complex models built solely on metabolomics data.

Case study: Modeling CAD by sex improves prediction using metabolomics

Building separate models using metabolomics data from distinct subcohorts can improve the overall classification of CA-CAC in individuals. To demonstrate the positive impact of modeling separate groups of individuals, we stratified the cohort by sex, a well-known risk factor for CAD. Models trained on males only (AUC = 0.75 in discovery and AUC = 0.63 validation) and females only (AUC = 0.72 in discovery and AUC = 0.65 in validation) outperformed a model which ignored this heterogeneity (AUC = 0.64 in discovery and AUC = 0.58 in validation; Figure 2A). Differences in model performance were explained by differences in the underlying models, as the most predictive metabolite-to-metabolites ratios in males were distinct from those in females (Figure 2B). Distinctions between the two models included serotonin present exclusively in the model built on females, and ratios with cAMP and tryptophan were present in the male model only. Ratios with choline were common across models in both sexes and the whole cohort. For example, the coefficients can be interpreted as a decrease in the ratio of histamine/serotonin corresponded to an increase in risk of CA-CAC+ in females and an increase in the ratio of 3HK/tryptophan corresponded to

Table 1. Clinical characteristics of discovery and validation cohort of individuals

Cohort	N	Overall, N = 837[a]	Discovery, N = 512[a]	Validation, N = 325[a]	p value[b]
Age	837	61 (12)	60 (12)	62 (12)	0.11
Sex	837				0.010
Male		461.0 (55.1%)	300.0 (58.6%)	161.0 (49.5%)	
Female		376.0 (44.9%)	212.0 (41.4%)	164.0 (50.5%)	
SMuRFs	837				0.070
0		177.0 (21.1%)	121.0 (23.6%)	56.0 (17.2%)	
1		360.0 (43.0%)	218.0 (42.6%)	142.0 (43.7%)	
2		220.0 (26.3%)	133.0 (26.0%)	87.0 (26.8%)	
3		68.0 (8.1%)	35.0 (6.8%)	33.0 (10.2%)	
4		12.0 (1.4%)	5.0 (1.0%)	7.0 (2.2%)	
CACS	837	216 (645)	213 (717)	220 (511)	0.3
FRS	837	0.08 (0.06)	0.08 (0.06)	0.09 (0.07)	0.025
BMI	834	26.9 (4.8)	26.8 (4.9)	27.1 (4.8)	0.4
Hypertension	837	323.0 (38.6%)	183.0 (35.7%)	140.0 (43.1%)	0.034
Diabetes	837	74.0 (8.8%)	37.0 (7.2%)	37.0 (11.4%)	0.039
Hypercholesterolemia	837	466.0 (55.7%)	269.0 (52.5%)	197.0 (60.6%)	0.022
Family history of ischemic heart disease	837	196.0 (23.4%)	143.0 (27.9%)	53.0 (16.3%)	<0.001
Antiplatelet	837	143.0 (17.1%)	85.0 (16.6%)	58.0 (17.8%)	0.6
ACE-inhibitor/Angiotensin receptor blocker	837	263.0 (31.4%)	154.0 (30.1%)	109.0 (33.5%)	0.3
Beta-blocker	837	126.0 (15.1%)	81.0 (15.8%)	45.0 (13.8%)	0.4
Diuretic	837	69.0 (8.2%)	38.0 (7.4%)	31.0 (9.5%)	0.3
Statin	837	280.0 (33.5%)	158.0 (30.9%)	122.0 (37.5%)	0.046
Smoking status	837				0.3
Current smoker		55.0 (6.6%)	37.0 (7.2%)	18.0 (5.5%)	
Never smoked		467.0 (55.8%)	291.0 (56.8%)	176.0 (54.2%)	
Ex-smoker		315.0 (37.6%)	184.0 (35.9%)	131.0 (40.3%)	
CA-CAC+	837	287.0 (34.3%)	168.0 (32.8%)	119.0 (36.6%)	0.3
SMuRFless	837	177.0 (21.1%)	121.0 (23.6%)	56.0 (17.2%)	0.027

N represents the total number of individuals and n represents the number of individuals with the corresponding clinical characteristic.

[a]Mean (SD); n (%).

[b]Wilcoxon rank-sum test; Pearson's Chi-squared test; Fisher's exact test.

an increase in risk of CA-CAC+ in males. The benefit of incorporating knowledge-guided subcohorts in modeling was evident as the performance of both sex-specific models outperformed the model trained on the whole cohort with improved interpretability of metabolite relationships with CA-CAC+.

Different approaches reveal distinct subcohorts and metabolomic signatures with CA-CAC+

Given the size and complexity of the BioHEART cohort, we compared a series of methods that are designed to identify distinct subcohorts. We found that each approach identified different subcohorts with varying clinical characteristics (Figure 3A). Under the knowledge-guided approaches, when we stratified by sex, males and females had distinct proportions of suggestive symptoms for CAD (47.3% in males and 66.8% in females), osteoarthritis (26.2% in males and 46.5% in females), and shortness of breath (18.7% in males and 34% in females). Comparing between statin and non-statin users, we found differences in the number of SMuRFs, aspirin, and anti-platelets. Using data-guided approaches, latent class analysis (LCA) identified five unique subcohorts differing on most clinical variables, with the most discriminating variables being hypertension, arthritis, aspirin, anti-platelets, and ACE inhibitor or ARB use. Unsupervised k-means clustering on the clinical characteristics produced subcohorts distinguished by age, number of SMuRFs,

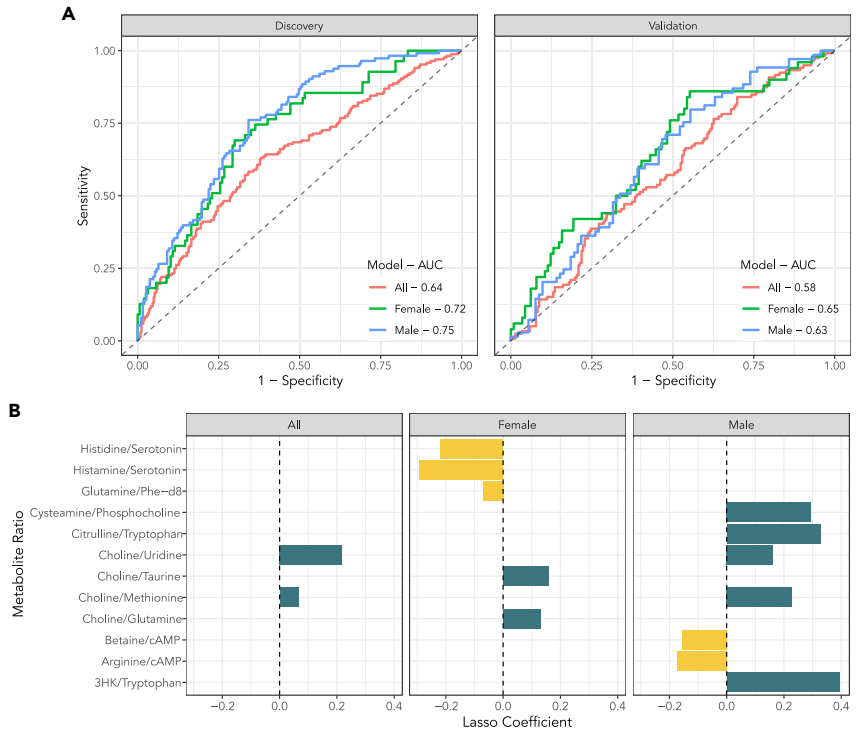


Figure 2. Modeling of CAD by sex gives rise to distinct models and improves performance

(A) Performance of models trained on both sexes, males only and females only. Lasso logistic regression models were trained within the three groups in the discovery cohort, and then ROC curves were generated for the predictions on the discovery cohort (left panel) and the validation cohort (right panel).

(B) Coefficients in each lasso regression model trained on both sexes, males only and females only.

hypertension, osteoarthritis, CA-CAC, previous medical history of heart-related issues, and use of cardiovascular-related medications (Figure S2), using rMoE distinguished cohorts characterized by old/young, SMuRFs, hypertension, and statin use. Evidently, many unique subcohorts exist, driven by different facets of the data, and can be elaborated through use of different approaches.

To investigate whether metabolite signatures differ between subcohorts, we assessed the association between each metabolite-to-metabolite ratio and CA-CAC+. Associations of each metabolite-to-metabolite ratio with CA-CAC+ were obtained for each subgroup and across the whole cohort using t-tests. There were distinct differences in the metabolite ratios associated with CA-CAC+ in each subcohort (Figure 3B). For example, from the knowledge-based approaches, citrulline/histidine was highly associated in non-statin users ($p = 3.07 \times 10^{-7}$, FDR $= 7.83 \times 10^{-4}$) yet not associated in statin users ($p = 0.443$, FDR $= 0.76$). Under the data-driven approaches, choline/histamine associated with CA-CAC+ in k-means cluster 1 ($p = 5.32 \times 10^{-7}$, FDR $= 1.36 \times 10^{-3}$) and was not associated in k-means cluster 2 ($p = 0.63$, FDR $= 0.91$); histamine/tryptophan was positively associated with CA-CAC+ in k-means cluster 2 ($p = 0.018$, FDR $= 0.24$) and negatively associated in k-means cluster 1 ($p = 2.34 \times 10^{-4}$, FDR $= 0.026$); and 1-methylhistamine/choline was negatively associated with CA-CAC+ in classifiability group 2 ($p = 7.67 \times 10^{-7}$, FDR $= 1.96 \times 10^{-3}$) and in the whole cohort ($p = 5.82 \times 10^{-5}$, FDR $= 9.63 \times 10^{-3}$), but had no association in classifiability group 1 ($p = 0.11$, FDR $= 0.47$).

This implies that between subcohorts there can be different risk factors and varying metabolite signatures associated with CAD, and so accounting for these may improve overall classification.

Subcohort-specific modeling improves overall prediction

To investigate whether incorporating subcohorts in modeling can improve classification of the overall cohort, we compared the performance of the model built on the whole cohort to combining models built within each subcohort. Within each approach, predictions from subcohort models were combined to

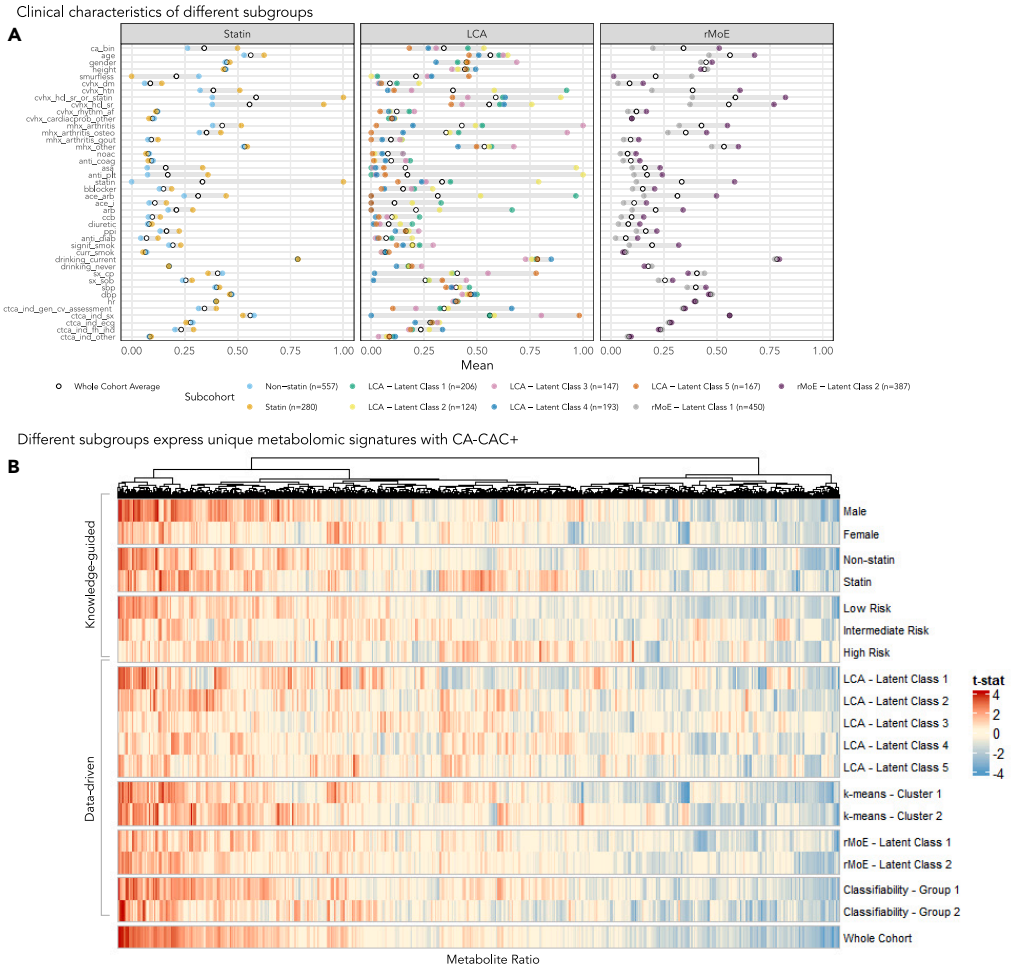


Figure 3. Various approaches for uncovering cohort heterogeneity produce subcohorts with distinct clinical characteristics and associations with CAD

(A) Plot of the mean values of each clinical variable across both discovery and validation cohorts within each subcohort, where numerical variables are scaled between 0 and 1 and categorical variables are binarized to 0 or 1. Plot is faceted by the approach used for deriving the subcohorts. The average value of each variable across the whole cohort is plotted in each facet as a white dot with black outline, and gray bars are plotted between the highest and lowest value in each facet.

(B) Heatmap of test statistics from individual t-tests testing between CA-CAC+/CA-CAC- and each metabolite-to-metabolite ratio within the respective subcohorts. Red values indicate positively associated and blue values indicate negatively associated metabolite ratios with CA-CAC+.

calculate overall balanced accuracies for the whole validation cohort (Figure 4A). The data-driven approaches rMoE, k-means, LCA, and classifiability had overall balanced accuracies of 0.63, 0.61, 0.6, and 0.59, respectively (Figure 4B). This was 0.01–0.05 higher than the whole cohort balanced accuracy of 0.58. Stratification by gender had the highest overall balanced accuracy of 0.63 of the knowledge-guided approaches. The lowest overall balanced accuracy of 0.57 was from splitting the cohort by FRS category, which was similar to the whole cohort model. The benefit of incorporating cohort heterogeneity in predicting CA-CAC+ with metabolomics was clear as the majority of approaches had resulted in similar or greater performance than the whole cohort model which did not consider heterogeneity.

Finally, we investigated whether each subcohort was more effective in predicting CA-CAC+, or only whether one subcohort improved the classification. Each approach in Table S1 was able to identify a subcohort of individuals where metabolomics predicted CA-CAC+ more accurately than the whole cohort model (Figure 4B), though there were situations where all subcohorts improved or only some improved prediction. In the latter case, classifiability modeling had one subcohort that improved prediction (bal.

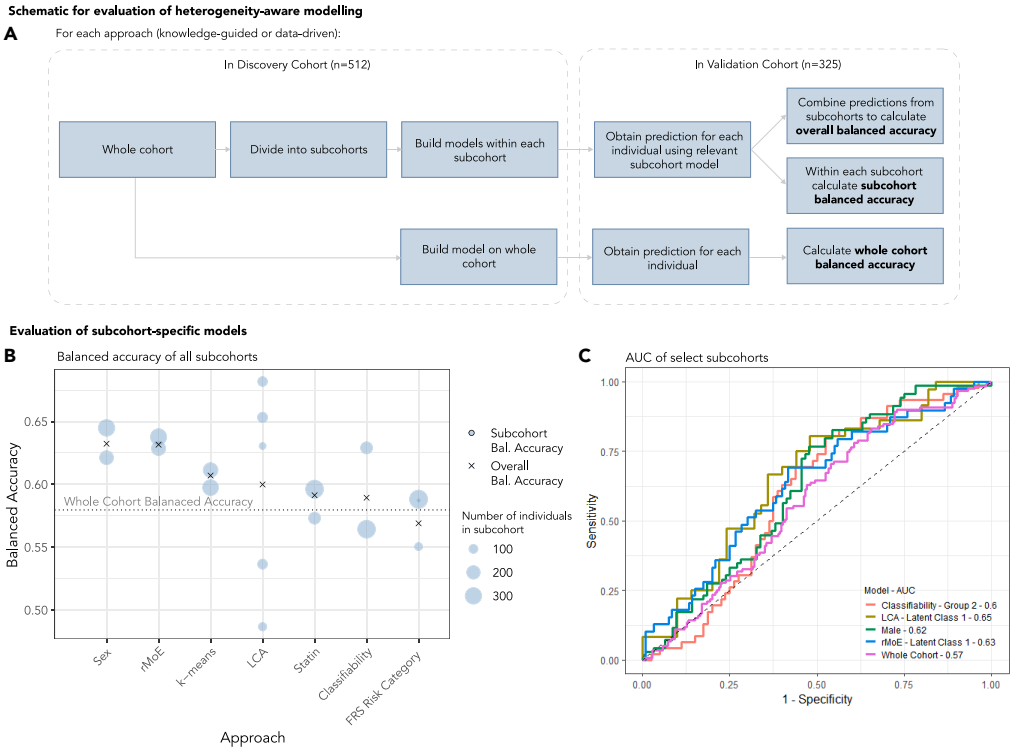


Figure 4. Acknowledging subcohorts in modeling of CAD improves prediction of the overall cohort

(A) Workflow of validation of subcohort models and calculated metrics.

(B) Dot plot of balanced accuracies in the validation cohort for each subcohort and approach. Approaches are ordered from highest to lowest overall balanced accuracy (left to right). Crosses indicate the overall balanced accuracies which combine the balanced accuracies across each subcohort, and dashed line indicates balanced accuracy from the model trained on the whole cohort.

(C) ROC curves of models from selected subcohorts from each category of approach, trained on the discovery cohort and tested on the validation cohort.

accuracy = 0.63, AUC = 0.6) and one that did poorly (bal. accuracy = 0.56, AUC = 0.54). This suggests that while we still perform better overall, we need to further understand whether the sub-optimal subcohort may need more data (see discussion). The subcohorts where metabolite ratios were most predictive were derived from LCA (bal. accuracy = 0.68, AUC = 0.65), stratification by males (bal. accuracy = 0.64, AUC = 0.62), and rMoE (bal. accuracy = 0.64, AUC = 0.63). Respective AUCs and corresponding ROC curves are shown in Figure 4C. Thus, by modeling distinct subcohorts, it is possible to identify groups of individuals for whom metabolomics signatures are most strongly predictive of CA-CAC+.

DISCUSSION

In this study, we investigated a large patient cohort where the cohort heterogeneity and its effect on the CVD risk were unknown. We compared several approaches for uncovering cohort heterogeneity present using clinical variables which allowed for a more effective analysis of CVD risk using metabolomics. We illustrated that both knowledge-guided and data-driven approaches reduced cohort heterogeneity present through identifying distinct subcohorts of individuals, and that predicting CA-CAC+ risk using metabolomics within these subcohorts produced different models with distinct performance. Overall, we highlighted several advantages of heterogeneity-aware modeling over modeling without consideration for cohort heterogeneity, including the improvement of overall accuracy and identification of subcohorts where we can be confident in predicting using metabolomics data.

The identification and acknowledgment of subcohorts in modeling enables more accurate and personalized predictions, an essential component in precision medicine for cardiovascular risk detection. By dividing the cohort into five latent classes using LCA, we observed subcohorts where the performance

of models was either much better or much worse than the whole cohort model. While the overall balanced accuracy of modeling the LCA subcohorts separately was similar to the whole cohort balanced accuracy, we were able to more insightfully detect the types of individuals where using metabolomics was either helpful or unhelpful in risk prediction. The next step of these findings would be to identify biomarkers through other data media (e.g. proteomics, lipidomics, or single-cell -omics) in the individuals where metabolomics failed to predict well, and incorporate these to further improve overall accuracy. Although metabolomics data were not able to outperform CA-CAC prediction through clinical variables, we believe that investigating more subcohorts and data media can begin to bridge this gap and allow for the identification of new biomarkers. These potential biomarkers across multiple platforms would improve the number of measurable, targetable, and modifiable risk factors of cardiovascular disease. When applied to the appropriate subcohort of individuals in clinical decision models, they can provide increasingly personalized prevention measures for harmful cardiovascular events.

In our analysis, we identified different subcohorts that had distinct signatures with CAD. Models that were built in some of these subcohorts — LCA - Latent Class 1, and Classifiability - Group 2 (Figure S2) — outperformed the model built on the whole cohort. The subcohort LCA - Latent Class 1 was mostly characterized by individuals with hypertension and use of ACE inhibitor or ARB; and Classifiability - Group 2 was relatively healthier individuals not using proton pump inhibitors, ACE inhibitors, anti-coagulants, or diuretics and had one or more SMuRFs. The top predictive features of CA-CAC+ in both of these models included ratios with choline.[37,38] In addition, the LCA - Latent Class 1 model had a higher ratio of cysteine/spermine and lower ratio of 3-deaazadenosine/cysteine correlating to increased risk of CA-CAC+, and the Classifiability - Group 2 model had a higher ratio of choline/α-keto-β-methylvaleric acid and a lower ratio of betaine/choline correlating to increased risk of CA-CAC+. These metabolites have been studied previously to determine associations with cardiovascular disease.[37–43] Since these ratios were employed to reduce the batch variation and additionally used in multivariate predictive models, the underlying biological mechanisms of each metabolite cannot be directly ascertained from these results. However, they do suggest that cohort heterogeneity plays a role in cardiovascular disease risk, providing insight into the categories of individuals who may be more at risk of CAD based on the aforementioned metabolites. Thus, cohort heterogeneity should be considered in the investigation and discovery of biomarkers of CAD.

A limitation of dividing a cohort into subcohorts is that for smaller groups of individuals there may be insufficient data to reliably assess associations with CA-CAC+. Given that distinct individual subcohorts may have unique drivers for disease risk, analyzing subcohorts with lower numbers of individuals may lead to lower statistical power and inaccurate representations of risk factors for these groups. In our case, separating individuals by FRS risk category meant that a small number of high-FRS risk individuals were separated for analysis. This subcohort on average had lower classification performance by AUC with much higher variability, possibly due to the lack of data points. Determining if the data population contains under-represented groups helps to decide whether a subcohort should be recognized as its own cohort for analysis or combined with others. This may be best determined from the contextual understanding of the clinical and biological experts; however, data-driven techniques, particularly methods for addressing fairness in machine learning,[44–46] can be adapted for the identification of cohort heterogeneity and the determination of under-represented subcohorts (or representation bias)[47] in the data. It is important to determine the characteristics of heterogeneity in datasets to begin to understand whether the trends found in them represent the broader public in general.

The data-driven approach of rMoE, in theory, is a modeling framework aligned to our aim of utilizing cohort heterogeneity for the improved prediction of CA-CAC+. rMoE achieves this by jointly optimizing a gating network using clinical variables and an experts network using metabolomics to predict CA-CAC+. While this method performed quite well (Figure 4B), a deeper investigation discovered that, in order to maximize the AUC, the clinical variables had a large impact on the rMoE predictions. This was evident as the algorithm used clinical variables to decide for each individual the probability of belonging to each latent class, where one of the latent classes were all classified as one outcome. This was contrary to our motivation to understand CA-CAC+ risk via metabolomics and it was clear that the clinical variables contained much more signal for predictive models. Although this approach optimizes the overall AUC with the data provided, further optimization within each latent class could be implemented to avoid having one latent class be predicted solely as one outcome, so that we may find distinct metabolomic signatures with CA-CAC+ for different subcohorts of individuals.

Previous analyses[5] of this metabolomic data have attempted to measure the average effect of each metabolite on the whole cohort with and without adjustment for age, sex, hypertension, hypercholesterolemia, diabetes, and significant smoking. In this analysis, we have been able to uncover subcohort-specific associations, as well as detect subcohorts using known and unrecognized contributors to heterogeneity. The case study for modeling for CA-CAC+ risk split by sex is not a groundbreaking finding, but demonstrates the benefits for subcohort modeling. While this may suggest that analysts should not analyze cohorts together to begin with, the heterogeneity existing in datasets may not be necessarily known before analysis, thus providing the need for methods to identify subcohorts. In addition, the subcohorts could have been modeled as interaction terms in models. We instead modeled the subcohorts separately to improve interpretability of the results, avoid the "curse of dimensionality",[48] and to allow the use of more complex models[49–51] which do not let the user define interaction terms such as in logistic regression.

In summary, we demonstrate the importance for incorporating cohort heterogeneity in predicting cardiovascular risk and review several different approaches that can be employed for untangling cohort heterogeneity present. This is particularly relevant when aiming to discover novel biological relationships with disease risk in complex datasets where the heterogeneity is unknown. Precision medicine remains a field with potential for substantial individual, public health, and economic impacts, through the discovery of more personalized prevention pathways for harmful cardiovascular events. As the diversity of individual characteristics and size of our datasets become available with the advent of newer technologies, we believe that addressing complex cohort heterogeneity will play a key role in the advancement of precision health and will lead to avenues for reducing the overall burden of cardiovascular disease in our comminutes.

Limitations of the study

A limitation of this study was that the individuals recruited into BioHEART were referred to for CT coronary angiography due to suspected CAD, thus the results presented in this paper may not be representative of the general population.

Throughout most of the models, we used metabolite-to-metabolite ratios as features to identify more complex relationships with disease and reduce some of the batch variation present in the data. Although the use of ratios decreased the interpretability of the relationships between metabolites and CA-CAC+, our primary aim was to improve prediction performance as opposed to maintaining high interpretability.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Clinical samples
- METHOD DETAILS
  - CT coronary imaging acquisition
  - Study cohort and design
  - Metabolomics data acquisition and processing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Data normalisation
  - Marker ratio calculation
  - Multivariate modeling and validation in whole cohort
  - Identification of subcohorts
  - Univariate associations
  - Multivariate modelling and validation within subcohorts
  - Calculation of balanced accuracies
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106633.

ACKNOWLEDGMENTS

The authors thank all their colleagues, particularly at The University of Sydney, Sydney Precision Data Science Center, and Charles Perkins Center for their support and intellectual engagement. The following sources of funding for each author are gratefully acknowledged. The AIR@innoHK programme of the Innovation and Technology Commission of Hong Kong to EP and JYHY. An Australian Research Council Discovery Early Career Researcher Award (DE200100944) funded by the Australian Government (E.P.). Research Training Program Stipend Scholarship to AC. G.F. is supported by a National Health and Medical Research Council Practitioner Fellowship (grant number APP11359290), Heart Research Australia, and the New South Wales Office of Health and Medical Research. The BioHEART study has received support from a combination of grants including from the Ramsay Teaching and Research Foundation, BioPlatforms Australia, the Vonwiller Foundation and Heart Research Australia. All funding sources had no role in the study design; in the collection, analysis, and interpretation of data, in the writing of the manuscript, and in the decision to submit the manuscript for publication.

AUTHOR CONTRIBUTIONS

E.P. and J.Y.H.Y. conceived the study and designed the investigation. S.W. and A.C. performed the data investigation and extended comparison study with input from T.L., E.P., and J.Y.H.Y. A.C. compiled and formulated the final data analysis with input from all authors. G.F. and S.V. jointly examined and analyzed the clinical relevance. All authors wrote the manuscript and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 3, 2022
Revised: February 3, 2023
Accepted: April 4, 2023
Published: April 11, 2023

REFERENCES

1. Roth, G.A., Mensah, G.A., Johnson, C.O., Addolorato, G., Ammirati, E., Baddour, L.M., Barengo, N.C., Beaton, A.Z., Benjamin, E.J., Benziger, C.P., et al. (2020). Global burden of cardiovascular diseases and risk factors, 1990-2019: update from the GBD 2019 study. J. Am. Coll. Cardiol. *76*, 2982–3021.

2. Figtree, G.A., and Vernon, S.T. (2021). Coronary artery disease patients without standard modifiable risk factors (SMuRFs)- a forgotten group calling out for new discoveries. Cardiovasc. Res. *117*, e76–e78.

3. Vernon, S.T., Hansen, T., Kott, K.A., Yang, J.Y., O'Sullivan, J.F., and Figtree, G.A. (2019). Utilizing state-of-the-art "omics" technology and bioinformatics to identify new biological mechanisms and biomarkers for coronary artery disease. Microcirculation *26*, e12488.

4. Kott, K.A., Vernon, S.T., Hansen, T., Yu, C., Bubb, K.J., Coffey, S., Sullivan, D., Yang, J., O'Sullivan, J., Chow, C., et al. (2019). Biobanking for discovery of novel cardiovascular biomarkers using imaging-quantified disease burden: protocol for the longitudinal, prospective, BioHEART-CT cohort study. BMJ Open *9*, e028649.

5. Vernon, S.T., Tang, O., Kim, T., Chan, A.S., Kott, K.A., Park, J., Hansen, T., Koay, Y.C., Grieve, S.M., O'Sullivan, J.F., et al. (2021). Metabolic signatures in coronary artery disease: results from the BioHEART-CT study. Cells *10*, 980. https://doi.org/10.3390/cells10050980.

6. Würtz, P., Raiko, J.R., Magnussen, C.G., Soininen, P., Kangas, A.J., Tynkkynen, T., Thomson, R., Laatikainen, R., Savolainen, M.J., Laurikka, J., et al. (2012). High-throughput quantification of circulating metabolites improves prediction of subclinical atherosclerosis. Eur. Heart J. *33*, 2307–2316.

7. Kent, D.M., Rothwell, P.M., Ioannidis, J.P.A., Altman, D.G., and Hayward, R.A. (2010). Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials *11*, 85.

8. Schork, N.J. (2015). Personalized medicine: time for one-person trials. Nature *520*, 609–611.

9. Cruz, D.E., Tahir, U.A., Hu, J., Ngo, D., Chen, Z.-Z., Robbins, J.M., Katz, D., Balasubramanian, R., Peterson, B., Deng, S., et al. (2022). Metabolomic analysis of coronary heart disease in an african American cohort from the jackson heart study. JAMA Cardiol. *7*, 184–194. https://doi.org/10.1001/jamacardio.2021.4925.

10. Ahmad, T., Lund, L.H., Rao, P., Ghosh, R., Warier, P., Vaccaro, B., Dahlström, U., O'Connor, C.M., Felker, G.M., and Desai, N.R. (2018). Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. J. Am. Heart Assoc. *7*, e008081. https://doi.org/10.1161/jaha.117.008081.

11. Higgins, J.P.T., and Green, S. (2008). Cochrane Handbook for Systematic Reviews of Interventions (Wiley).

12. Velentgas, P., Dreyer, N.A., Nourjah, P., and Smith, S.R. (2013). In Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide P, N.A.D. Velentgas, P. Nourjah, S.R. Smith, and M.M. Torchia, eds. (Agency for Healthcare Research and Quality (US)).

13. Sharma, A., Zheng, Y., Ezekowitz, J.A., Westerhout, C.M., Udell, J.A., Goodman, S.G., Armstrong, P.W., Buse, J.B., Green, J.B.,

Josse, R.G., et al. (2022). Cluster analysis of cardiovascular phenotypes in patients with type 2 diabetes and established atherosclerotic cardiovascular disease: a potential approach to precision medicine. Diabetes Care *45*, 204–212.

14. Deng, K., Zhang, X., Liu, Y., Zhang, L., Wang, G., Feng, M., Oliver, B.G., Wang, L., Hansbro, P.M., Qin, L., et al. (2021). Heterogeneity of paucigranulocytic asthma: a prospective cohort study with hierarchical cluster analysis. J. Allergy Clin. Immunol. Pract. *9*, 2344–2355.

15. Zinchuk, A., and Yaggi, H.K. (2020). Phenotypic subtypes of osa: a challenge and opportunity for precision medicine. Chest *157*, 403–420.

16. Al-Taie, Z., Liu, D., Mitchem, J.B., Papageorgiou, C., Kaifi, J.T., Warren, W.C., and Shyu, C.-R. (2021). Explainable artificial intelligence in high-throughput drug repositioning for subgroup stratifications with interventionable potential. J. Biomed. Inf. *118*, 103792.

17. Queyrel, M., Templier, A., and Zucker, J.-D. (2021). Reject and cascade classifier with subgroup discovery for interpretable metagenomic signatures. Communications in Computer and Information Science, 49–66. https://doi.org/10.1007/978-3-030-93736-2_5.

18. Liu, D., Baskett, W., Beversdorf, D., and Shyu, C.-R. (2020). Exploratory data mining for subgroup cohort discoveries and prioritization. IEEE J. Biomed. Health Inform. *24*, 1456–1468.

19. Tolstikov, V., Moser, A.J., Sarangarajan, R., Narain, N.R., and Kiebish, M.A. (2020). Current status of metabolomic biomarker discovery: impact of study design and demographic characteristics. Metabolites *10*, 224. https://doi.org/10.3390/metabo10060224.

20. Mosca, L., Barrett-Connor, E., and Wenger, N.K. (2011). Sex/gender differences in cardiovascular disease prevention: what a difference a decade makes. Circulation *124*, 2145–2154.

21. Meyer, M.R., Haas, E., and Barton, M. (2006). Gender differences of cardiovascular disease: new perspectives for estrogen receptor signaling. Hypertension *47*, 1019–1026.

22. Vaura, F., Palmu, J., Aittokallio, J., Kauko, A., and Niiranen, T. (2022). Genetic, molecular, and cellular determinants of sex-specific cardiovascular traits. Circ. Res. *130*, 611–631.

23. Leopold, J.A., and Antman, E.M. (2021). A precision medicine approach to sex-based differences in ideal cardiovascular health. Sci. Rep. *11*, 14848.

24. Barden, A.E., Huang, R.-C., Beilin, L.J., Rauschert, S., Tsai, I.-J., Oddy, W.H., and Mori, T.A. (2022). Identifying young adults at high risk of cardiometabolic disease using cluster analysis and the Framingham 30-yr risk score. Nutr. Metabol. Cardiovasc. Dis. *32*, 429–435.

25. Akbaraly, T., Würtz, P., Singh-Manoux, A., Shipley, M.J., Haapakoski, R., Lehto, M., Desrumaux, C., Kähönen, M., Lehtimäki, T., Mikkilä, V., et al. (2018). Association of circulating metabolites with healthy diet and risk of cardiovascular disease: analysis of two cohort studies. Sci. Rep. *8*, 8620.

26. Talmor-Barkan, Y., Bar, N., Shaul, A.A., Shahaf, N., Godneva, A., Bussi, Y., Lotan-Pompan, M., Weinberger, A., Shechter, A., Chezar-Azerrad, C., et al. (2022). Metabolomic and microbiome profiling reveals personalized risk factors for coronary artery disease. Nat. Med. *28*, 295–302.

27. Mundra, P.A., Barlow, C.K., Nestel, P.J., Barnes, E.H., Kirby, A., Thompson, P., Sullivan, D.R., Alshehry, Z.H., Mellett, N.A., Huynh, K., et al. (2018). Large-scale plasma lipidomic profiling identifies lipids that predict cardiovascular events in secondary prevention. JCI Insight *3*, e121326. https://doi.org/10.1172/jci.insight.121326.

28. Vázquez-Fresno, R., Llorach, R., Perera, A., Mandal, R., Feliz, M., Tinahones, F.J., Wishart, D.S., and Andres-Lacueva, C. (2016). Clinical phenotype clustering in cardiovascular risk patients for the identification of responsive metabotypes after red wine polyphenol intake. J. Nutr. Biochem. *28*, 114–120.

29. Flores, A.M., Schuler, A., Eberhard, A.V., Olin, J.W., Cooke, J.P., Leeper, N.J., Shah, N.H., and Ross, E.G. (2021). Unsupervised learning for automated detection of coronary artery disease subgroups. J. Am. Heart Assoc. *10*, e021976.

30. Jeong, A., Imboden, M., Hansen, S., Zemp, E., Bridevaux, P.-O., Lovison, G., Schindler, C., and Probst-Hensch, N. (2017). Heterogeneity of obesity-asthma association disentangled by latent class analysis, the SAPALDIA cohort. Respir. Med. *125*, 25–32.

31. Mori, M., Krumholz, H.M., and Allore, H.G. (2020). Using latent class analysis to identify hidden clinical phenotypes. JAMA *324*, 700–701.

32. Hurley, N.C., Berkowitz, A., Masoudi, F., Ross, J., Desai, N., Shah, N., Dhruva, S., and Mortazavi, B.J. (2021). Outcomes-driven clinical phenotyping in cardiogenic shock using a mixture of experts. In 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) (IEEE). https://doi.org/10.1109/bhi50953.2021.9508568.

33. Xu, X., Lubomski, M., Holmes, A.J., Sue, C.M., Davis, R.L., Muller, S., and Yang, J.Y.H. (2021). NEMoE: a nutrition aware regularized mixture of experts model addressing diet-cohort heterogeneity of gut microbiota in Parkinson's Disease. Preprint at medRxiv. https://doi.org/10.1101/2021.11.10.21266194.

34. Huo, Z., Zhang, L., Khera, R., Huang, S., Qian, X., Wang, Z., and Mortazavi, B.J. (2021). Sparse gated mixture-of-experts to separate and interpret patient heterogeneity in EHR data. In 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) (IEEE). https://doi.org/10.1109/bhi50953.2021.9508549.

35. Shou, X., Mavroudeas, G., Magdon-Ismail, M., Figueroa, J., Kuruzovich, J.N., and Bennett, K.P. (2020). Supervised mixture of experts models for population health. Methods *179*, 101–110.

36. Patrick, E., Schramm, S.-J., Ormerod, J.T., Scolyer, R.A., Mann, G.J., Mueller, S., and Yang, J.Y.H. (2017). A multi-step classifier addressing cohort heterogeneity improves performance of prognostic biomarkers in three cancer types. Oncotarget *8*, 2807–2815.

37. Ueland, P.M. (2011). Choline and betaine in health and disease. J. Inherit. Metab. Dis. *34*, 3–15.

38. Millard, H.R., Musani, S.K., Dibaba, D.T., Talegawkar, S.A., Taylor, H.A., Tucker, K.L., and Bidulescu, A. (2018). Dietary choline and betaine; associations with subclinical markers of cardiovascular disease risk and incidence of CVD, coronary heart disease and stroke: the Jackson Heart Study. Eur. J. Nutr. *57*, 51–60.

39. Madeo, F., Eisenberg, T., Pietrocola, F., and Kroemer, G. (2018). Spermidine in health and disease. Science *359*, eaan2788. https://doi.org/10.1126/science.aan2788.

40. Eisenberg, T., Abdellatif, M., Zimmermann, A., Schroeder, S., Pendl, T., Harger, A., Stekovic, S., Schipke, J., Magnes, C., Schmidt, A., et al. (2017). Dietary spermidine for lowering high blood pressure. Autophagy *13*, 767–769.

41. Walker, G., Langheinrich, A.C., Dennhauser, E., Bohle, R.M., Dreyer, T., Kreuzer, J., Tillmanns, H., Braun-Dullaeus, R.C., and Haberbosch, W. (1999). 3-deazaadenosine prevents adhesion molecule expression and atherosclerotic lesion formation in the aortas of C57BL/6J mice. Arterioscler. Thromb. Vasc. Biol. *19*, 2673–2679.

42. Liu, C.-L., Guo, J., Zhang, X., Sukhova, G.K., Libby, P., and Shi, G.-P. (2018). Cysteine protease cathepsins in cardiovascular disease: from basic research to clinical trials. Nat. Rev. Cardiol. *15*, 351–370.

43. Lutgens, S.P.M., Cleutjens, K.B.J.M., Daemen, M.J.A.P., and Heeneman, S. (2007). Cathepsin cysteine proteases in cardiovascular disease. Faseb. J. *21*, 3029–3041.

44. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Comput. Surv. *54*, 1–35.

45. Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G., and Chin, M.H. (2018). Ensuring fairness in machine learning to advance health equity. Ann. Intern. Med. *169*, 866–872.

46. Hardt, P.; Srebro (2016). Equality of opportunity in supervised learning. Adv. Neural Inf. Process. Syst. *29*.

47. Suresh, H., and Guttag, J.V. (2019). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. Preprint at arXiv. https://doi.org/10.1145/3465416.3483305.

48. Berisha, V., Krantsevich, C., Hahn, P.R., Hahn, S., Dasarathy, G., Turaga, P., and Liss, J. (2021). Digital medicine and the curse of dimensionality. NPJ Digit. Med. *4*, 153.

49. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

50. Boser, B.E., Guyon, I.M., and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory COLT '92 (Association for Computing Machinery), pp. 144–152.

51. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., et al. (2015). Xgboost: extreme gradient boosting. R package version 0. 4-2 *1*, 1–4.

52. adam2o1o (2023). Adam2o1o/metabolomicscohortheterogeneity_analysis: cohort heterogeneity manuscript analysis. Preprint at Zenodo. https://doi.org/10.5281/ZENODO.7707308.

53. Harden, S.P., Bull, R.K., Bury, R.W., Castellano, E.A., Clayton, B., Hamilton, M.C.K., Morgan-Hughes, G.J., O'Regan, D., Padley, S.P.G., Roditi, G.H., et al. (2016). The safe practice of CT coronary angiography in adult patients in UK imaging departments. Clin. Radiol. *71*, 722–728.

54. Gensini, G.G. (1983). A more meaningful scoring system for determining the severity of coronary heart disease. Am. J. Cardiol. *51*, 606.

55. Agatston, A.S., Janowitz, W.R., Hildner, F.J., Zusmer, N.R., Viamonte, M., Jr., and Detrano, R. (1990). Quantification of coronary artery calcium using ultrafast computed tomography. J. Am. Coll. Cardiol. *15*, 827–832.

56. Koay, Y.C., Stanton, K., Kienzle, V., Li, M., Yang, J., Celermajer, D.S., and O'Sullivan, J.F. (2021). Effect of chronic exercise in healthy young male adults: a metabolomic analysis. Cardiovasc. Res. *117*, 613–622.

57. Koay, Y.C., Wali, J.A., Luk, A.W.S., Macia, L., Cogger, V.C., Pulpitel, T.J., Wahl, D., Solon-Biet, S.M., Holmes, A., Simpson, S.J., and O'Sullivan, J.F. (2019). Ingestion of resistant starch by mice markedly increases microbiome-derived metabolites. Faseb. J. *33*, 8033–8042.

58. Kim, T., Tang, O., Vernon, S.T., Kott, K.A., Koay, Y.C., Park, J., James, D.E., Grieve, S.M., Speed, T.P., Yang, P., et al. (2021). A hierarchical approach to removal of unwanted variation for large-scale metabolomics data. Nat. Commun. *12*, 4992.

59. Wang, K.Y.X., Pupo, G.M., Tembe, V., Patrick, E., Strbenac, D., Schramm, S.-J., Thompson, J.F., Scolyer, R.A., Mueller, S., Tarr, G., et al. (2020). Cross-Platform Omics Prediction procedure: a game changer for implementing precision medicine in patients with stage-III melanoma. Pre Print at. bioRxiv. https://doi.org/10.1101/2020.12.09.415927.

60. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. J. Stat. Software *33*, 1–22.

61. Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A. (2009). The e1071 Package. https://www.cs.upc.edu/~belanche/Docencia/mineria/Practiques/R/e1071.pdf.

62. Anderson, K.M., Odell, P.M., Wilson, P.W., and Kannel, W.B. (1991). Cardiovascular disease risk profiles. Am. Heart J. *121*, 293–298.

63. Linzer, D.A., and Lewis, J.B. (2011). poLCA: an R package for polytomous variable latent class analysis. J. Stat. Software *42*, 1–29.

64. R Core Team (2020). R: A Language and Environment for Statistical Computing.

65. Kaufman, L., and Rousseeuw, P.J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis (John Wiley & Sons).

66. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., and Hinton, G.E. (1991). Adaptive mixtures of local experts. Neural Comput. *3*, 79–87.

67. Atzmueller, M., and Puppe, F. (2006). SD-map – a fast algorithm for exhaustive subgroup discovery. In Lecture Notes in Computer Science Lecture notes in computer science (Springer Berlin Heidelberg), pp. 6–17.

68. Ganin, U. (2016). Ajakan, and Germain Domain-adversarial training of neural networks. J. Mach.

69. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer Science & Business Media).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Adult patients undergoing clinically indicated CTCA	BioHEART-CT Study[4]	https://www.australianclinicaltrials.gov.au/anzctr/trial/ACTRN12618001322224
Deposited data		
Metabolomics and Clinical Data	This paper	https://zenodo.org/badge/latestdoi/542345826
Original Code	This paper	https://zenodo.org/badge/latestdoi/542345826
Software and algorithms		
R version 4.2.1	R foundation	https://www.R-project.org/
glmnet	Friedman et al.[60]	https://cran.r-project.org/web/packages/glmnet/index.html
NEMoE	Xu et al.[33]	https://github.com/SydneyBioX/NEMoE
poLCA	Linzer et al.[63]	https://cran.r-project.org/web/packages/poLCA/index.html
rsubgroup	Atzmueller et al.[67]	https://cran.r-project.org/web/packages/rsubgroup/index.html
hRUV	Kim et al.[58]	https://github.com/SydneyBioX/hRUV
ggplot2	N/A	https://cran.r-project.org/web/packages/ggplot2/

RESOURCE AVAILABILITY

Lead contact

Further information and any related requests should be directed to and will be fulfilled by the lead contact, Ellis Patrick (ellis.patrick@sydney.edu.au).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The version of source code used for the preparation of the manuscript and the data generated in this study is available on Zenodo: https://doi.org/10.5281/zenodo.7707308.[52]
- Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Clinical samples

The samples used were from the BioHEART-CT Study.[44] The study was approved by the Northern Sydney Local Health District Human Research Ethics Committee (HREC/17/HAWKE/343) and all participants provided informed written consent. Briefly, individuals undergoing clinically indicated CT coronary angiogram (CTCA) for suspected coronary artery disease were recruited from multiple sites in Sydney, Australia. Blood samples were taken at the time of recruitment, and after appropriate processing, plasma samples including replicates were aliquoted and stored at -80°C until analysis. Individuals were advised to fast for $\geq$2 hours prior to their CTCA as per standard clinical practice.

METHOD DETAILS

CT coronary imaging acquisition

CTCA images were acquired on a 256-slice scanner using standard clinical protocols, overseen and dual-reported by accredited cardiologists and radiologists. Heart rate was optimised using oral metoprolol or

ivabradine based on body weight, as clinically appropriate. Prospective studies were preferentially performed unless heart rate control was suboptimal, in which case retrospective acquisition was utilised. Vasodilation was achieved using sublingual nitroglycerine (600-800 micrograms) given immediately prior to intravenous contrast delivery. Radiation doses were minimised as per guidelines[53] and reconstructions were performed using vendor-specific software. CTCAs were analysed using the validated 17-segment Gensini score to identify those with and without CAD.[54] Coronary artery calcium scores (CACS) were assessed using vendor-specific software utilising the Agatston method[55] as a measure of CAD burden. In brief, hyperattunated areas of at least 1 mm^2 or $\geq$ 3 adjacent pixels with >130 Hounsfield units (HU) were incorporated into the CACS using the well described and validated Agatston method.

Study cohort and design

This study included 837 from the initial 1000 individuals of the BioHEART-CT study. The discovery set consists of batches 3-10 and the validation set consists of batches 11-15. There were 512 individuals in the discovery set and 325 individuals in the validation set. Calcium scores are calculated as an absolute numerical value using the well-validated standardised tool (Agatson score) and also expressed as a percentile based on sex, age and ethnicity based on large well-characterised datasets (Multi-Ethnic Study of Atherosclerosis). Individuals with a raw CACS greater than 100 Agatston units or calcium score age/sex percentile in the top quartile (> 75%) are considered to have clinically actionable CACS (CA-CAC+) and all others are considered as being free of clinically actionable CACS (CA-CAC-).

Metabolomics data acquisition and processing

Targeted metabolomics based on scheduled multiple reaction monitoring optimised to the metabolite of interest using authentic standards was applied in this study[5] 10 µl plasma was mixed with 90µl HILIC sample buffer, an acetonitrile: methanol: formic acid mix (75:25:0.2, v:v:v). The resulting mixture was vortexed and spun at 14,000 rpm for 20 minutes to remove plasma protein. The metabolite containing supernatant was then transferred to a glass HPLC sample vial and resolved on an Agilent 1260 Infinity HPLC System, and m/z was determined by Qtrap5500 (Sciex).[56,57] Each sample was eluted over a 25-minute period, and each batch of samples took 40 hours to complete. A total of 14 batches were completed over 44 days. Metabolite elution characteristics were pre-determined using pure standards. Metabolite abundance peaks were integrated using the area under the curve for calibrated peaks from MultiQuant (SCIEX), with manual adjustments to the curves as appropriate. This ensures the consistency of all the peaks integrated. The metabolites that were not present in at least 30% of the samples were filtered out. Missing values were then imputed using the minimum value of each metabolite. This resulted in a total of 71 metabolites (see Table S3).

QUANTIFICATION AND STATISTICAL ANALYSIS

Data normalisation

For the metabolites analysed, the hRUV[58] was then performed on the log2 transformed processed metabolomics data. For intra batch normalisation, loess smoothing on samples and RUV-III using short replicates with parameter k set to 15 was performed. For inter batch normalisation, a concatenating hierarchical structure using batch replicate samples was used, with the RUV-III parameter k was set to 3. Following hRUV normalisation, the mean within each batch was subtracted from the normalised value to remove any remaining batch variation as shown in Figure S6.

Marker ratio calculation

Metabolites analysed in this paper were calculated as metabolite-to-metabolite ratios for each metabolite in the dataset. This was done following the transferability concept proposed in[59] and self-normalising aspect of calculating ratios to reduce variation between batches (Figure S6). The raw imputed metabolite data was first log transformed and then ratios between each metabolite were taken. The mean within each batch was subtracted from the normalised value to remove any remaining batch variation. This resulted in 2485 metabolite ratios considered in analysis.

Multivariate modeling and validation in whole cohort

A lasso logistic regression model was used to predict CA-CAC+ through the glmnet package[60] on the whole cohort using clinical and demographic variables(including age, sex, and number of SMuRFs). The top 15 features by absolute difference in AUC from 0.5 as independent variables in the model, and the

tuning parameter lambda was selected by the cv.glmnet function from the glmnet package.[60] A naive bayes model, as implemented by the e1701 R package,[61] was built on all metabolite-to-metabolite ratios. Ratios with variance greater than 0.1 were filtered and then the top 15 by absolute difference in AUC from 0.5 were used as independent variables in the model. Both models were validated in the discovery cohort using repeated 3-fold cross validation (CV) with 5 repeats, and validated through the validation cohort.

Identification of subcohorts

Several approaches were employed to uncover subgroups in the data. Each of these approaches used different criterion to select the number of subgroups and resulted in varying numbers of subgroups. The results generated in this study are summarised in Table S2.

Framingham Risk Score (FRS) risk category

FRS scores were calculated using the formula described in.[62] Individuals scoring a CVD risk score of 0.15 or greater were categorised as high risk, below 0.15 and 0.1 or greater were categorised as intermediate risk and below 0.1 were categorised as low risk.

Latent class analysis (LCA)

LCA was performed using the poLCA package[63] using the clinical variables as input. The number of classes, 5, was chosen using the Bayesian information criterion.

k-means Clustering

The kmeans function from the R stats[64] package was used to perform k-means clustering using the clinical variables as input. The average silhouette of observations was computed for different values of k, with the optimal number of clusters, 2, chosen by the maximum average silhouette.[65]

Regularised mixture of experts (rMoE)

rMoE was performed using the implementation from the NEMoE package.[33] rMoE comprises a mixture of experts model[66] with added elastic-net penalties to both the gating and experts networks. Clinical variables were used in the gating network and the metabolite ratios were used in the experts network with the penalty regulariser values of 0.08 and 0.012 respectively. The number of latent classes was chosen to be 2 and the initialisation method was k-means clustering. Latent classes determined by rMoE were determined by rounding the gating probability, and predictions used the overall probability calculated by the dot product of the gating probabilities and expert probabilities.

Classifiability

Repeated 5-fold CV with 5 repeats was performed in the discovery cohort using metabolite ratios as features to predict CA-CAC with naive bayes models. Individual classifiability scores[36] were then obtained by calculating the proportion of correct classifications for each individual. The subgroup discovery algorithm, SD-Map,[67] was then used with the clinical variables as input and CA-CAC as the target to determine a subcohort defined by clinical variables to have higher classifiability scores. The group with higher classifiability determined by SD-Map were called group 1 and the rest called group 2.

Univariate associations

P-values in Table 1 were calculated using Wilcoxon rank-sum test for numerical variables; Pearson's Chi-squared test for categorical variables; and Fisher's exact test for binary variables. Test statistics between metabolite ratios and CA-CAC in Figure 3 were calculated using two sample Welch t-tests.

Multivariate modelling and validation within subcohorts

Lasso logistic regression models were used in Figure 2 sex-sepecific models. Feature selection used |AUC-0.5| to select the top 15 features, and the tuning parameter lambda was selected by the cv.glmnet function from the glmnet package.[60]

Subcohort models were constructed using naive bayes models on metabolite ratios following feature selection. A variance cutoff of 0.1 and then the top 15 features by absolute difference in AUC from 0.5 were

used as criteria for feature selection. All subcohort models were validated within the individuals in the respective subcohort in both.

Adversarial Domain Generalization (ADG) approaches employ a domain discriminator combined with a feature extractor to adversarially learn domain-invariant features, which are representative while their domain cannot be recognized by the domain discriminator.[68] In this work, different batches are treated as different domains and we use ADG to learn the batch-invariant features for downstreaming analysis. Performance of using metabolite-to-metabolite ratios was compared with ADG to validate the idea of transferrable markers.

Calculation of balanced accuracies

In order to calculate balanced accuracy, a cutoff had to be chosen to classify CA-CAC+/CA-CAC- from model predictions. For each model, we computed ROC curves, and determined the optimal cutoff as the one that produced the smallest distance to the point (0,1) on ROC space.

All data was analysed in R version 4.2.1 [64] and visualised using ggplot2.[69]

ADDITIONAL RESOURCES

The Trial Registration Number for the BioHEART-CT Study: ACTRN12618001322224.