UNIVERSITAT DE BARCELONA

biobank uk

MACHINE LEARNING AND CAUSAL INFERENCE APPROACHES FOR SYSTEMIC MULTI-DISEASE ASSOCIATIONS IN UK Biobank

Alejandro González Álvarez

BCN AIM

1. Clinical problem statement

- Single disease approach. Classical diagnosis and treatment approaches of illnesses have been treated as differentiated and isolated specialties (i.e. cardiology or neurology).

- Multi-organ disease association. Many observational studies have supported the clinical relevance of multi-organ disease association.

- Neurocardiology: specialty which was born with the goal of studying and understanding the pathophysiological interplay of the nervous and cardiovascular systems.

2. Data Science objectives

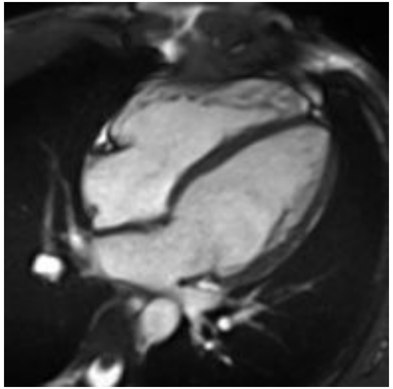
1. To integrate multiple data sources from different organs (heart and brain).

2. To facilitate the discovery of the causal links between heart and brain diseases by using both ML techniques and causal inference analysis.

3. Compare the performance of traditional machine learning techniques and causal inference approaches in multi-disease association studies.

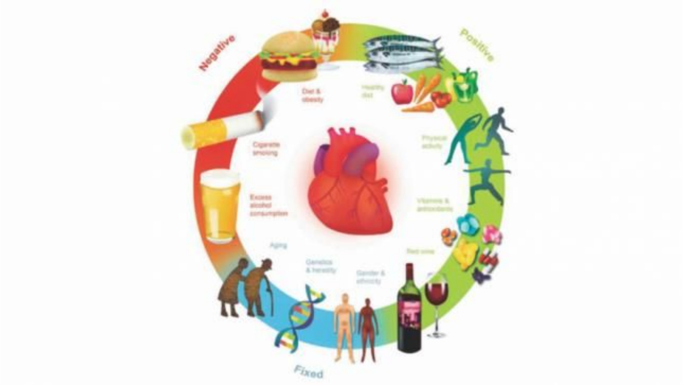
3. Datasets description

2065 UK Biobank Patients and 1416 variables



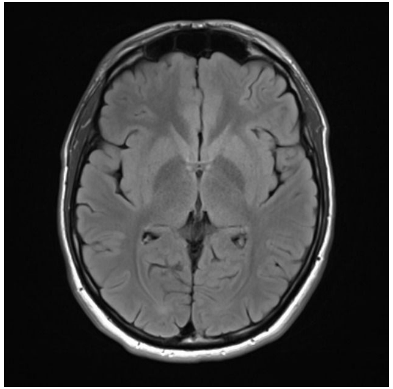
3.1 Heart CMR Radiomics

Heart imaging derived data that quantifies various changes in heart structures



3.2 VRFs

Cardiovascular risk factors that capture how well the heart works

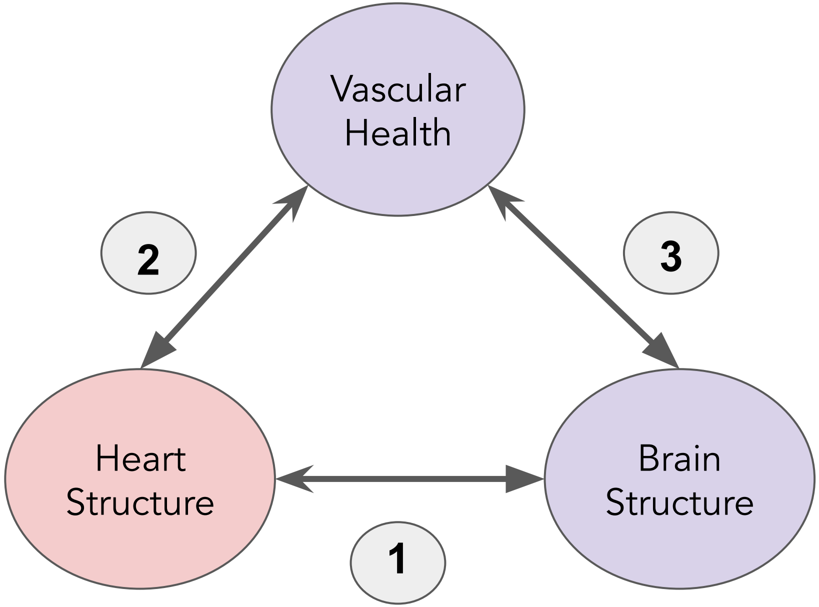


3.3 Brain MRI Indices

Brain structural imaging data that contains the structure of various brain regions

4. Related work

Plenty of studies have looked at the relationships between heart structure, brain structure and vascular health individually



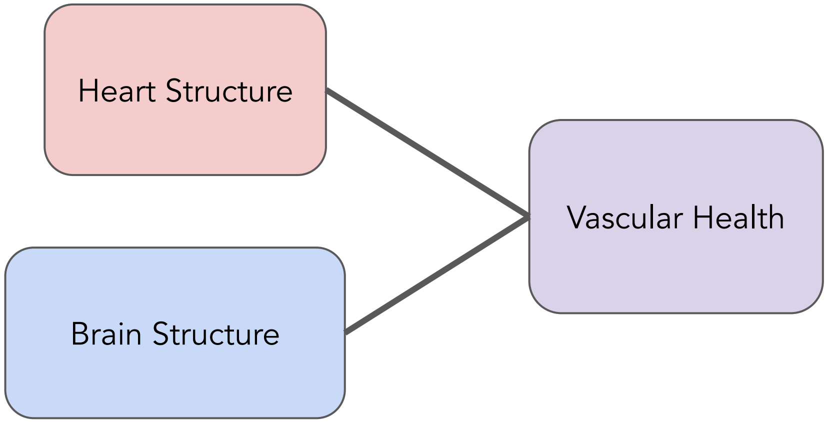
1. Phloyngam (2019)
2. Cetin et al. (2020)

1. Cox et al. (2019)
2. Gorelick and Sorond (2018)

1. Bruns (2017)
2. Jefferson et al. (2010)

5. Our approach

5.1 Traditional ML with Separate Models

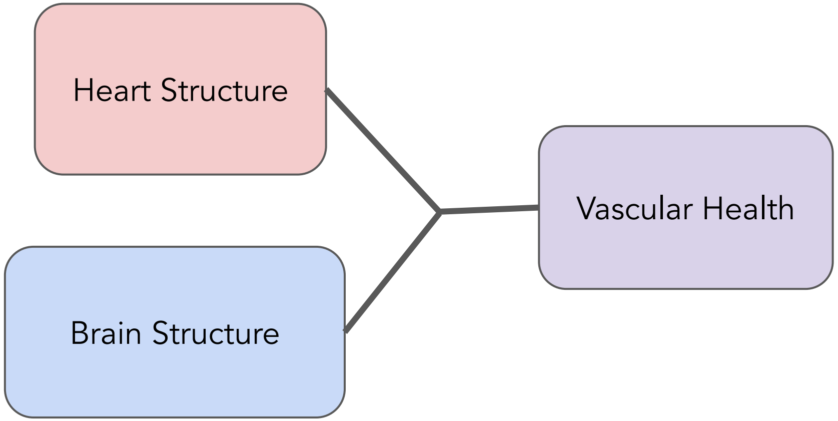


Heart Structure

Brain Structure

Vascular Health

5.2 Traditional ML with Joint Model



Heart Structure

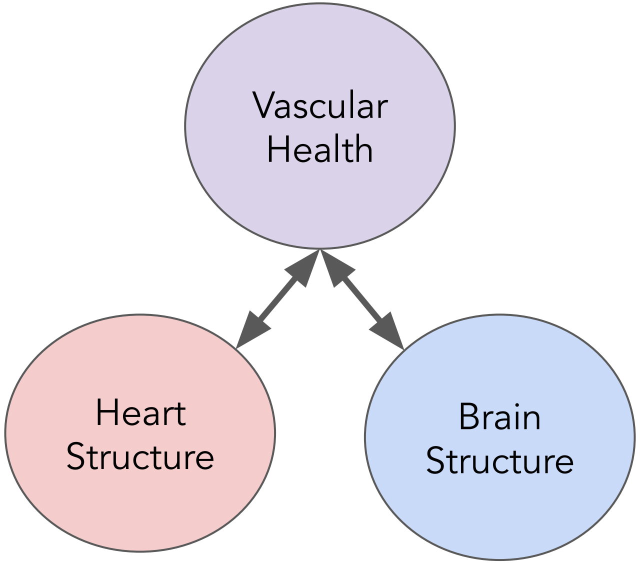
Brain Structure

Vascular Health

- In previous work, connections and associations were studied independently. Our approach attempts to study them simultaneously.

- Initial hypothesis: brain structure and heart structure are independent and provide unique information, therefore together they will improve performance.

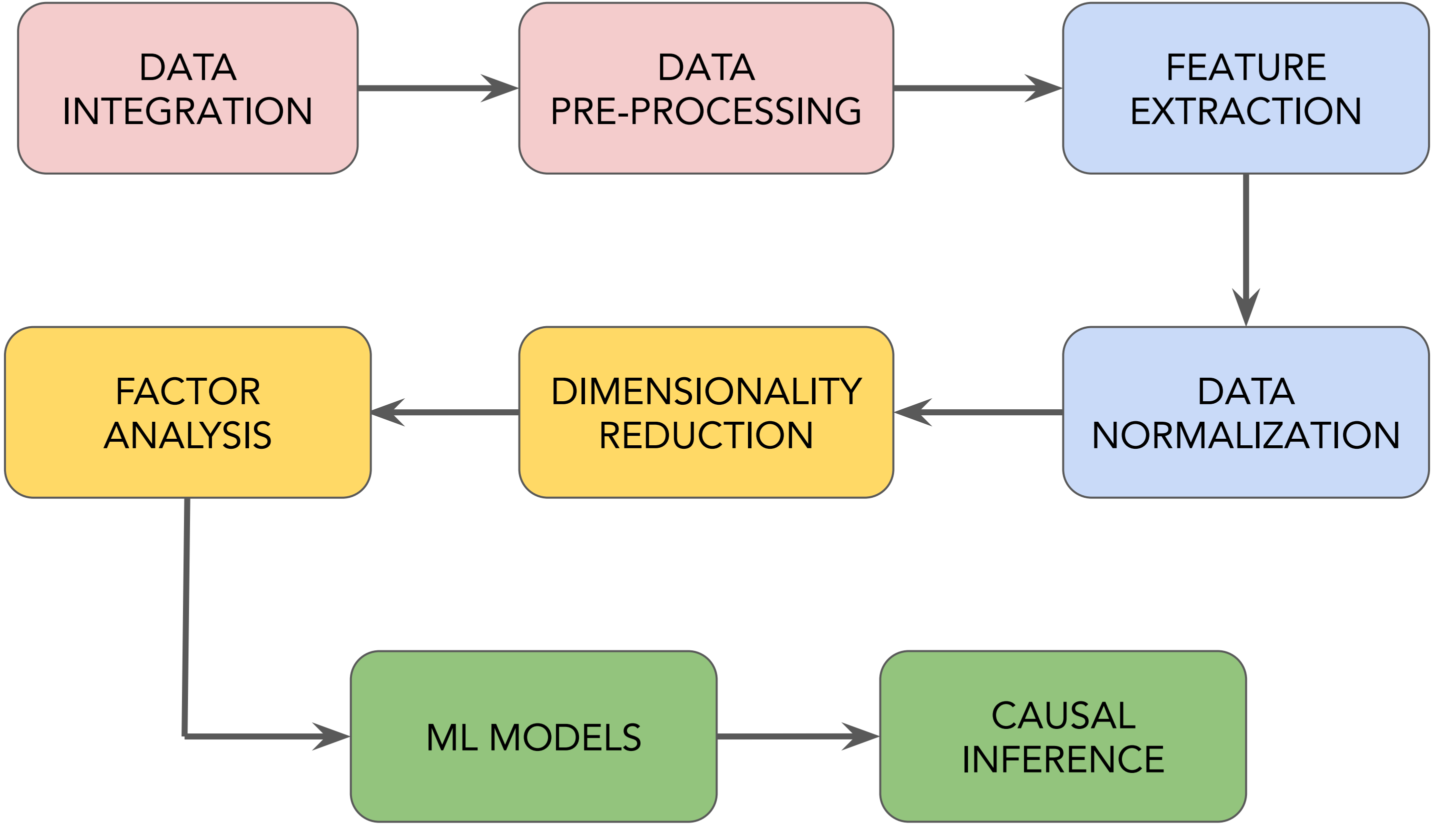
5.3 Causal Inference techniques



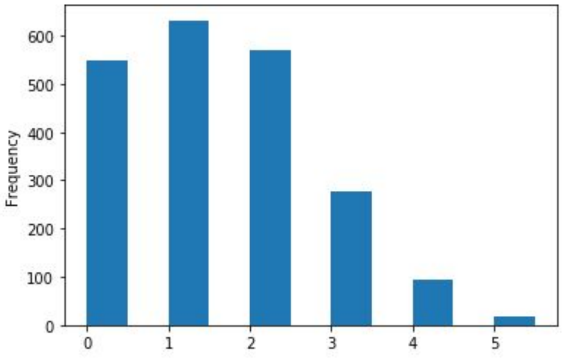
- If our initial hypothesis fails and performance does not improve, it will mean that both datasets do not provide unique information and as a result they might be similar and somehow correlated.

- Therefore, the relationship between them will be high. Causal Inference techniques pretend to find this link between them.

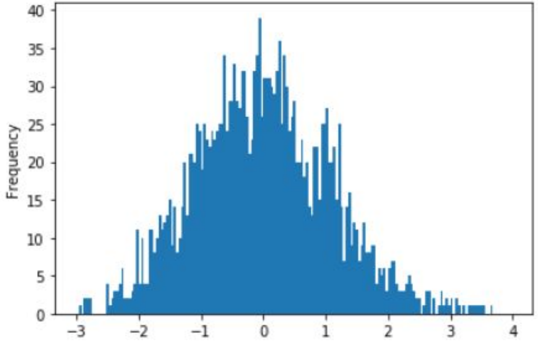
6. Data pipeline



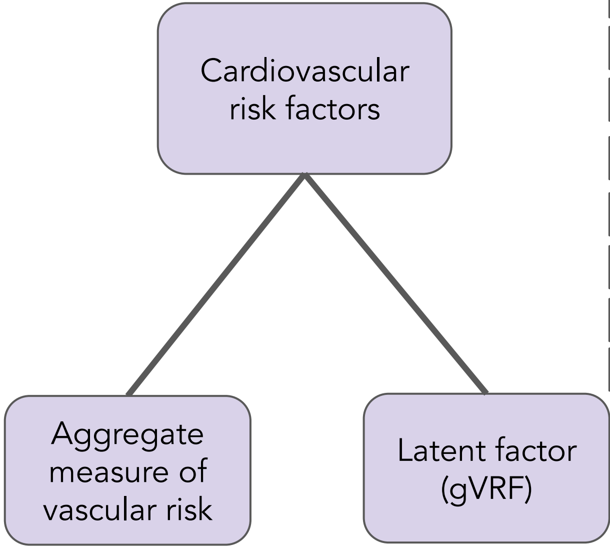
7. Dimensionality reduction



7.1 Aggregate measure

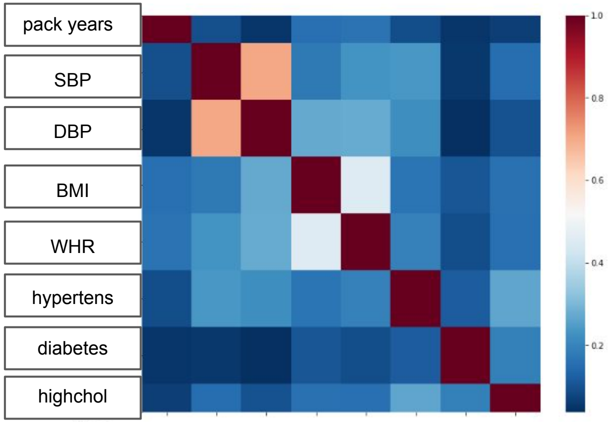


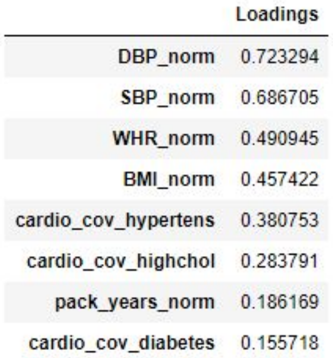
7.2 gVRF


Cardiovascular risk factors
Aggregate measure of vascular risk
Latent factor (gVRF)

Counting instances of a self-reported diagnosis of hypertension, diabetes, or hypercholesterolaemia, having ever smoked, having a BMI >25, and having a high WHR (>0.85 for females and >0.90 for males).

We also derived a latent factor of general vascular risk (gVRF). It was derived from smoking pack years, diastolic and systolic blood pressure, BMI, WHR, self-reported hypertension, diabetes and hypercholesterolaemia.

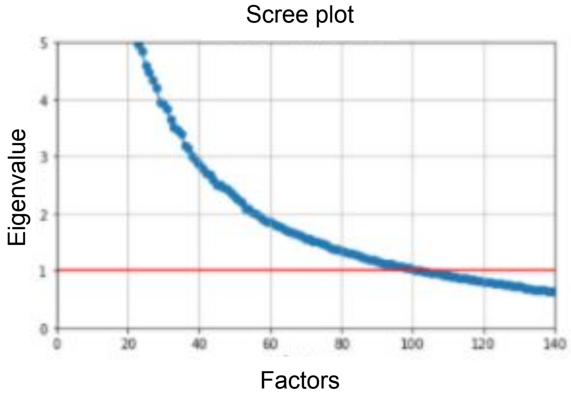


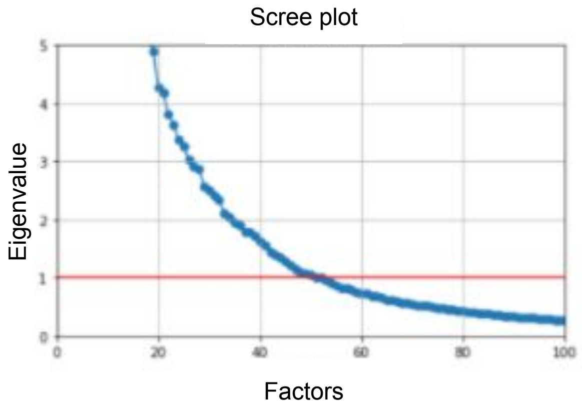


	Loadings
DBP_norm	0.723294
SBP_norm	0.686705
WHR_norm	0.490945
BMI_norm	0.457422
cardio_cov_hypertens	0.380753
cardio_cov_highchol	0.283791
pack_years_norm	0.186169
cardio_cov_diabetes	0.155718

Brain Structural MRI

Heart CMR Radiomics

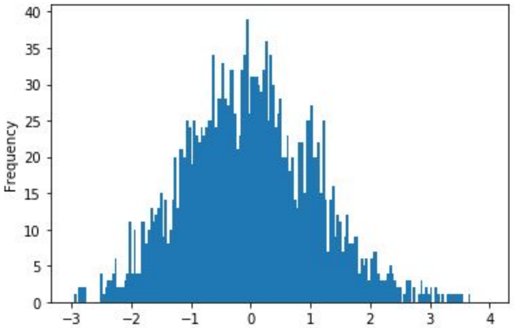




Features	# Variables	KMO Score	Scree test
gVRF	8	0,6418	3
Brain MRI Indices	744	0.9526	100
Heart CMR Radiomics	639	0.9781	50

8. Machine Learning approaches

8.1 Predicting gVRF

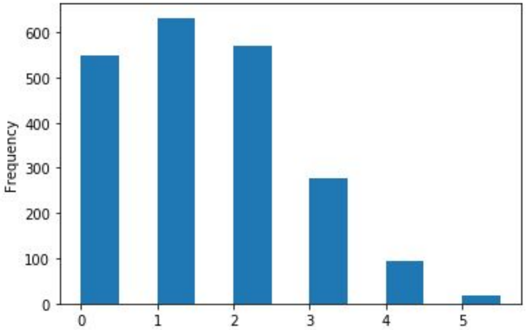


from sklearn.linear_model import LinearRegression

Evaluation of performance

- R^2: coefficient of determination, regression score function.
- MAPE: mean absolute percentage error.
- MAE: mean of the absolute value of the errors.
- MSE: mean of the squared errors.
- RMSE: square root of the mean of the squared errors.

8.2 Predicting Aggregate measure



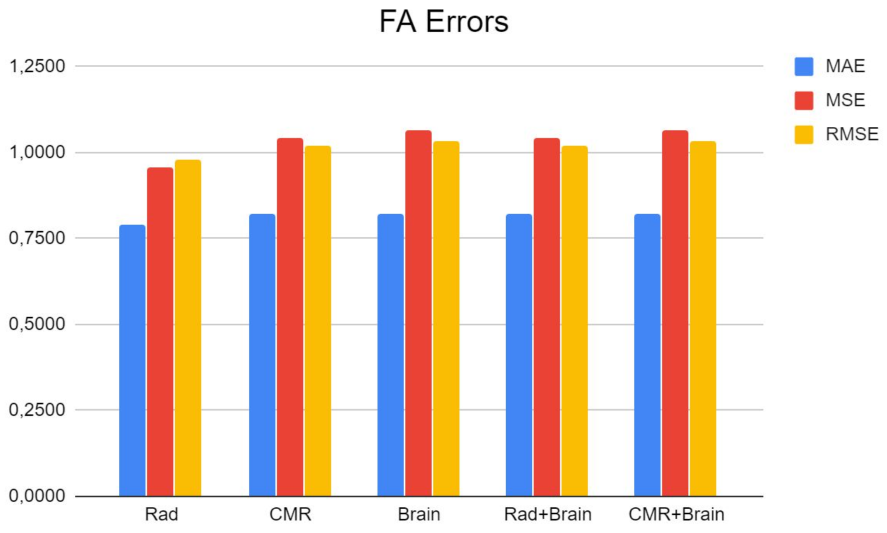
from sklearn.ensemble import RandomForest

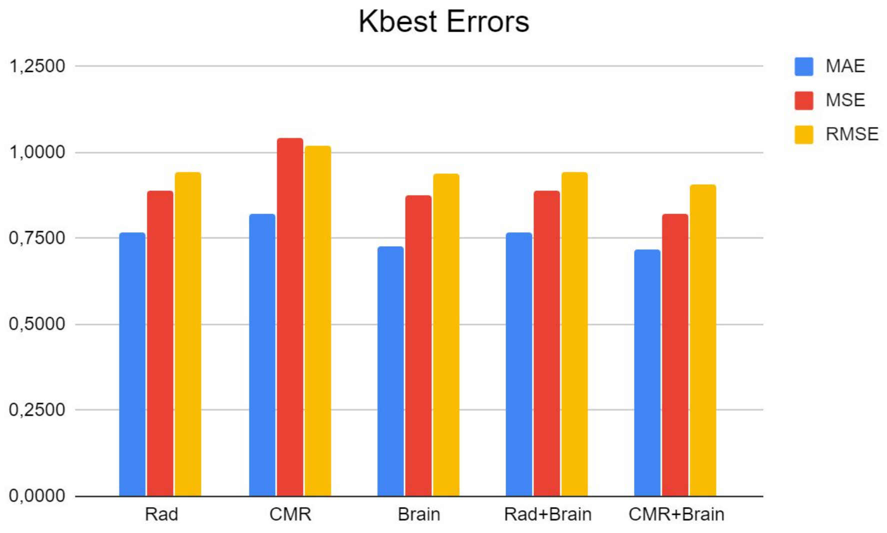
Evaluation of performance

- Accuracy.
- Confusion matrix.
- ROC curve.
- AUC.

8.1 Predicting gVRF

Best results are obtained with heart radiomics in case of FA, and brain MRI indices and its combination with cardio CMR in case of SelectKbest feature selection



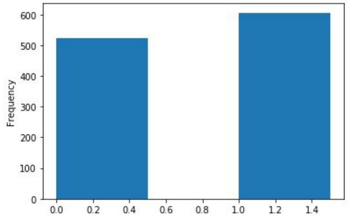


However, overall, the combination of these datasets does not seem to improve at all, or improve very little our performance metrics, which may lead us to reject our initial hypothesis

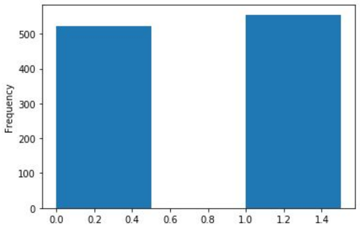
8.2 Predicting aggregate measure of vascular risk

- Multiclass classification problem ✗
- Binarize classes:
  - Random oversampling ✓
  - Random undersampling
  - SMOTE-Tomek

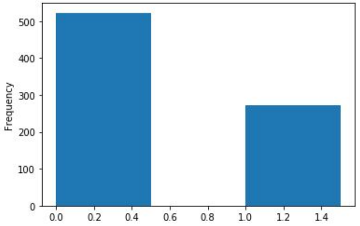
0vs1



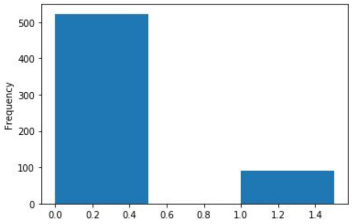
0vs2



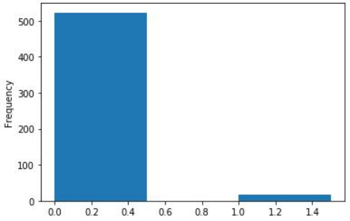
0vs3



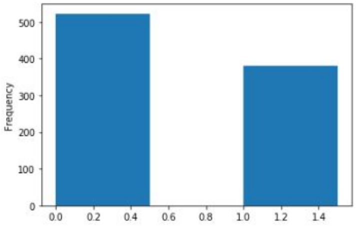
0vs4



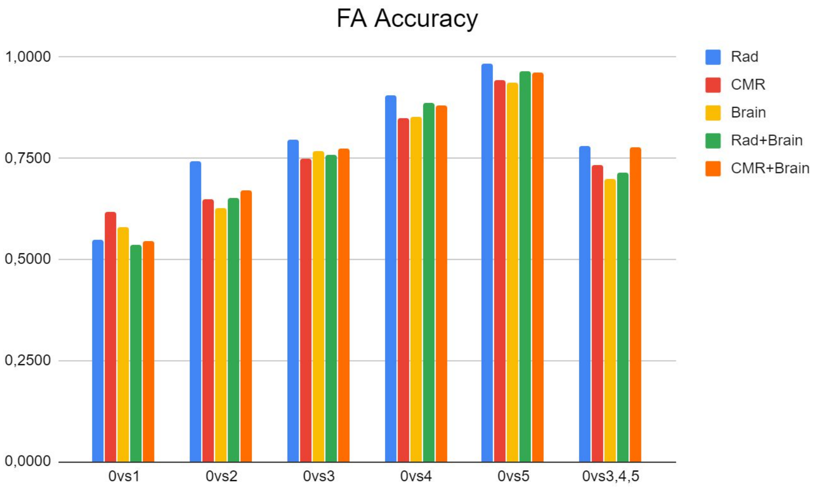
0vs5

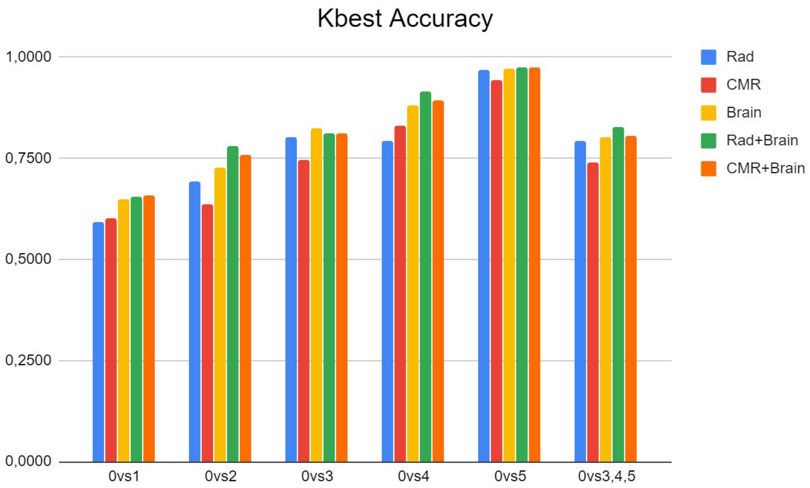


0vs3,4,5



The main trend is that model accuracy and AUC improves as VRFs burden increases.

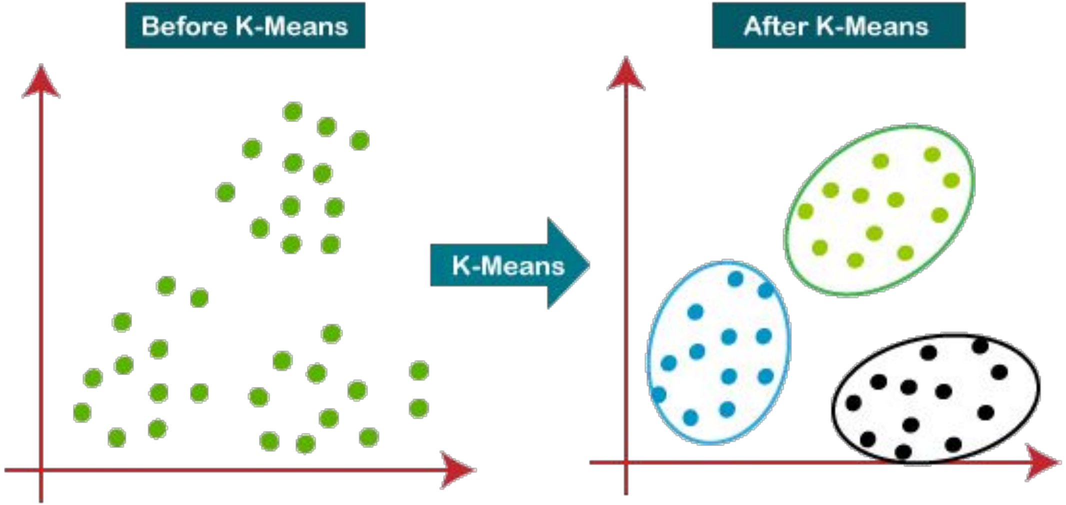




Again, as it happened when predicting gVRF, the combination of heart and brain datasets does not seem to improve at all, or improve very little our performance metrics, which may lead us to reject our initial hypothesis.

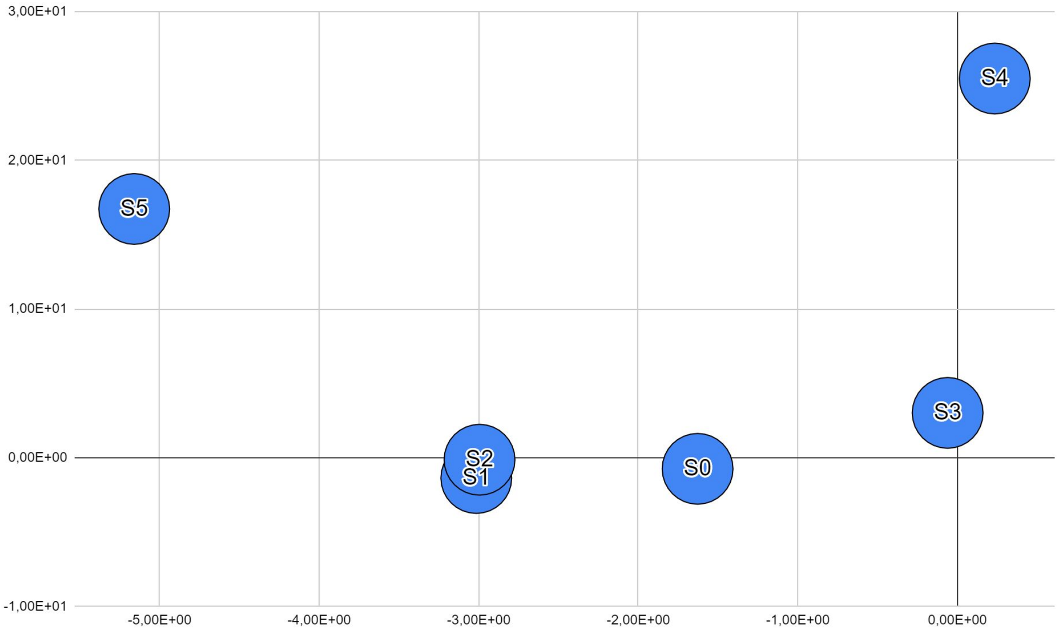
8.3 K-means clustering

- ML results reinforced the hypothesis mentioned earlier that different classes in our aggregate measure might be increasingly different as the number of vascular risk factors increases.
- To measure the distance between classes we ran the K-means clustering algorithm and computed the centroids for each class.



- S5, patients with an aggregate score of five, and S0, patients with an aggregate score of zero, are the most distant ones, reinforcing our hypothesis that these two classes differ the most.

- S1, S2 and S3 are the closest aggregate measures to S0, showing why these comparisons obtained the lowest performance metrics and why they are the most similar classes from our target variable.



* Summary so far...

1. Combination of brain and heart structure’ datasets did not improve when trying to predict vascular health.

2. We may reject our initial hypothesis, since these datasets may NOT provide unique information, and somehow they are related.

3. Combined effects of risk factors were better predicted by heart CMR radiomics when using FA as the dimensionality reduction technique.

4. Combined effects of risk factors were better detected by brain MRI indices when using SelectKBest as the dimensionality reduction technique.

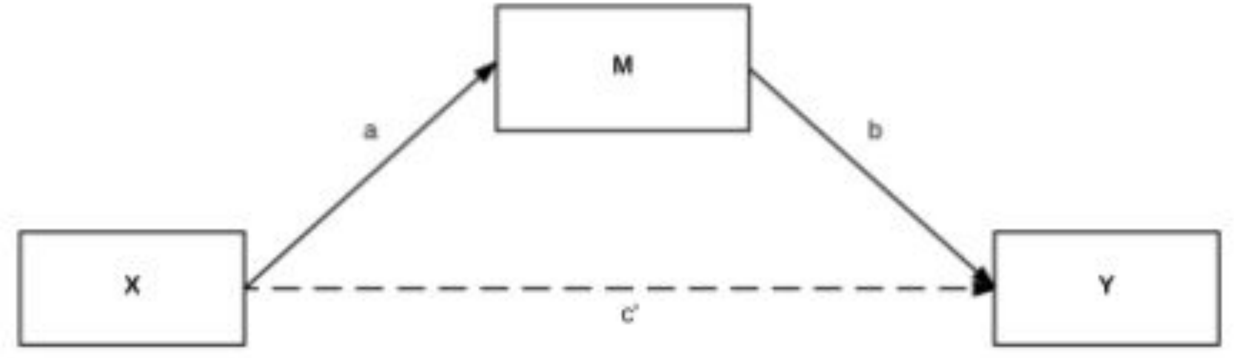
5. Direct link between VRFs and brain structure?

6. Direct link between VRFs and heart structure?

7. An intermediate factor is playing a mediating role between these associations?

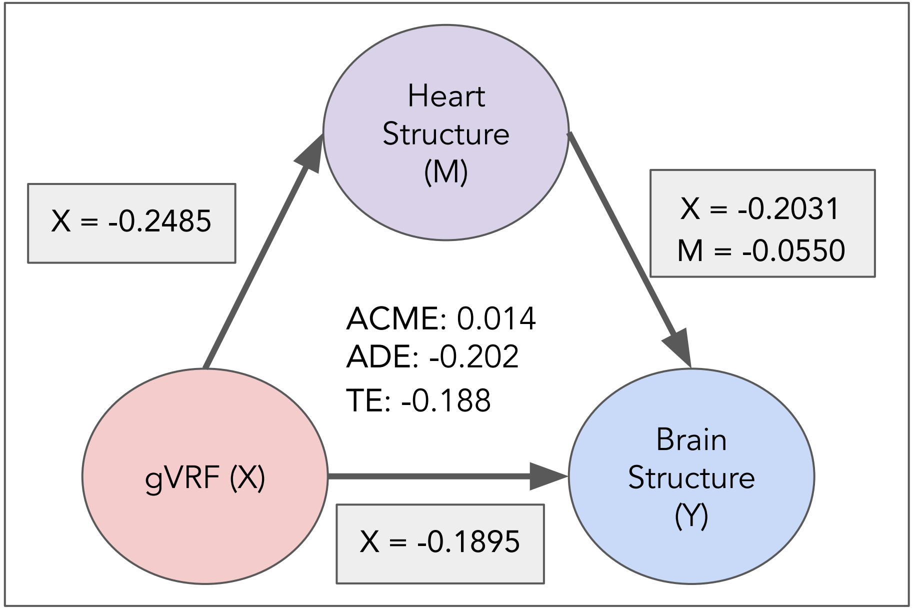
9. Causal Inference techniques (Mediation Analysis)

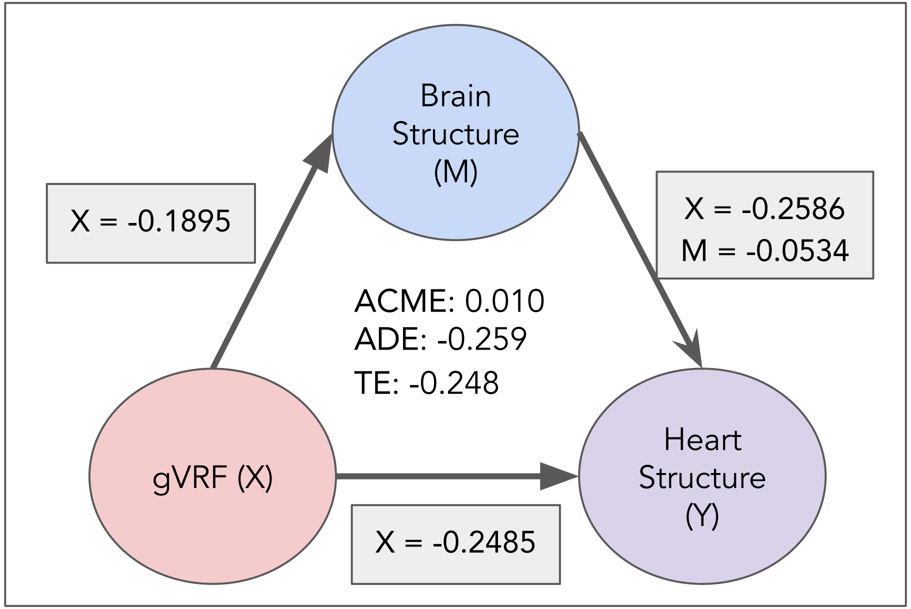
A mediation analysis is comprised of three sets of regressions: X → Y, X → M, and X + M → Y



- ACME (average causal mediation effect): Total effect minus the direct effect (TE - ADE).
- ADE (average direct effect) : A direct effect of X on Y after taking into account a mediation indirect effect of M (X + M → Y).
- TE (total effect) (indirect + direct effect): A total effect of X on Y (without M) (X → Y).

How much of the connection between cardiovascular risk and heart/brain structure can be explained by changes in brain/heart structure?





- VRF → Heart: strongly negatively correlated
- VRF → Brain: strongly negatively correlated
- Heart → Brain: weakly negatively correlated

What these factors and negative correlations mean?

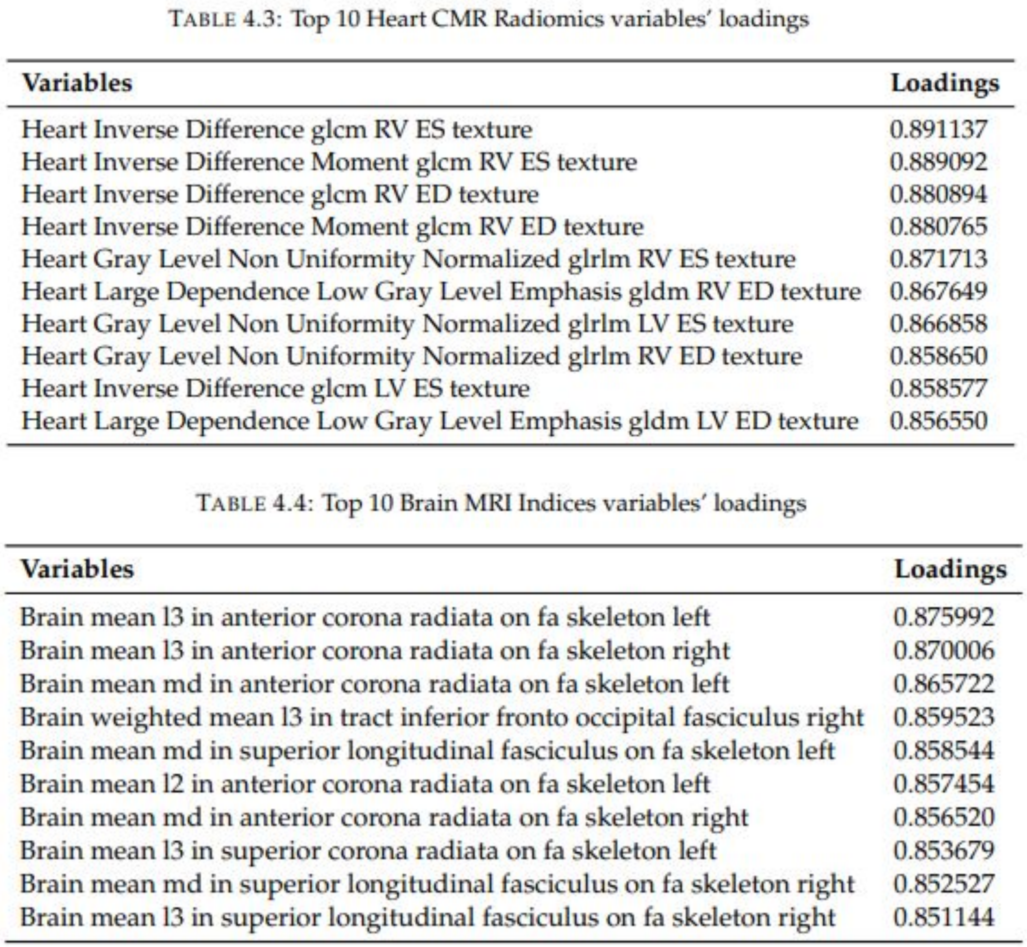


TABLE 4.3: Top 10 Heart CMR Radiomics variables' loadings

Variables	Loadings
Heart Inverse Difference glcm RV ES texture	0.891137
Heart Inverse Difference Moment glcm RV ES texture	0.889092
Heart Inverse Difference glcm RV ED texture	0.880894
Heart Inverse Difference Moment glcm RV ED texture	0.880765
Heart Gray Level Non Uniformity Normalized glrlm RV ES texture	0.871713
Heart Large Dependence Low Gray Level Emphasis gldm RV ED texture	0.867649
Heart Gray Level Non Uniformity Normalized glrlm LV ES texture	0.866858
Heart Gray Level Non Uniformity Normalized glrlm RV ED texture	0.858650
Heart Inverse Difference glcm LV ES texture	0.858577
Heart Large Dependence Low Gray Level Emphasis gldm LV ED texture	0.856550

TABLE 4.4: Top 10 Brain MRI Indices variables' loadings

Variables	Loadings
Brain mean l3 in anterior corona radiata on fa skeleton left	0.875992
Brain mean l3 in anterior corona radiata on fa skeleton right	0.870006
Brain mean md in anterior corona radiata on fa skeleton left	0.865722
Brain weighted mean l3 in tract inferior fronto occipital fasciculus right	0.859523
Brain mean md in superior longitudinal fasciculus on fa skeleton left	0.858544
Brain mean l2 in anterior corona radiata on fa skeleton left	0.857454
Brain mean md in anterior corona radiata on fa skeleton right	0.856520
Brain mean l3 in superior corona radiata on fa skeleton left	0.853679
Brain mean md in superior longitudinal fasciculus on fa skeleton right	0.852527
Brain mean l3 in superior longitudinal fasciculus on fa skeleton right	0.851144

10. Discussion and Future Work

1. Combination of features did not improve performance, but we only used a subset of features from each dataset (5 SelectKBest features and 5 latent factors) for comparison purposes.

2. Correlation was not very high between these subsets of features (maybe with more features involved correlation increases).

3. Classes in aggregate measure too similar between them. (Better way to measure vascular health?)

4. VRF → Heart: strongly negatively correlated.

5. VRF → Brain: strongly negatively correlated.

6. Heart ⟷ Brain: weakly negatively correlated.

7. Small mediation role of the heart and the brain (which might increase with more features involved).