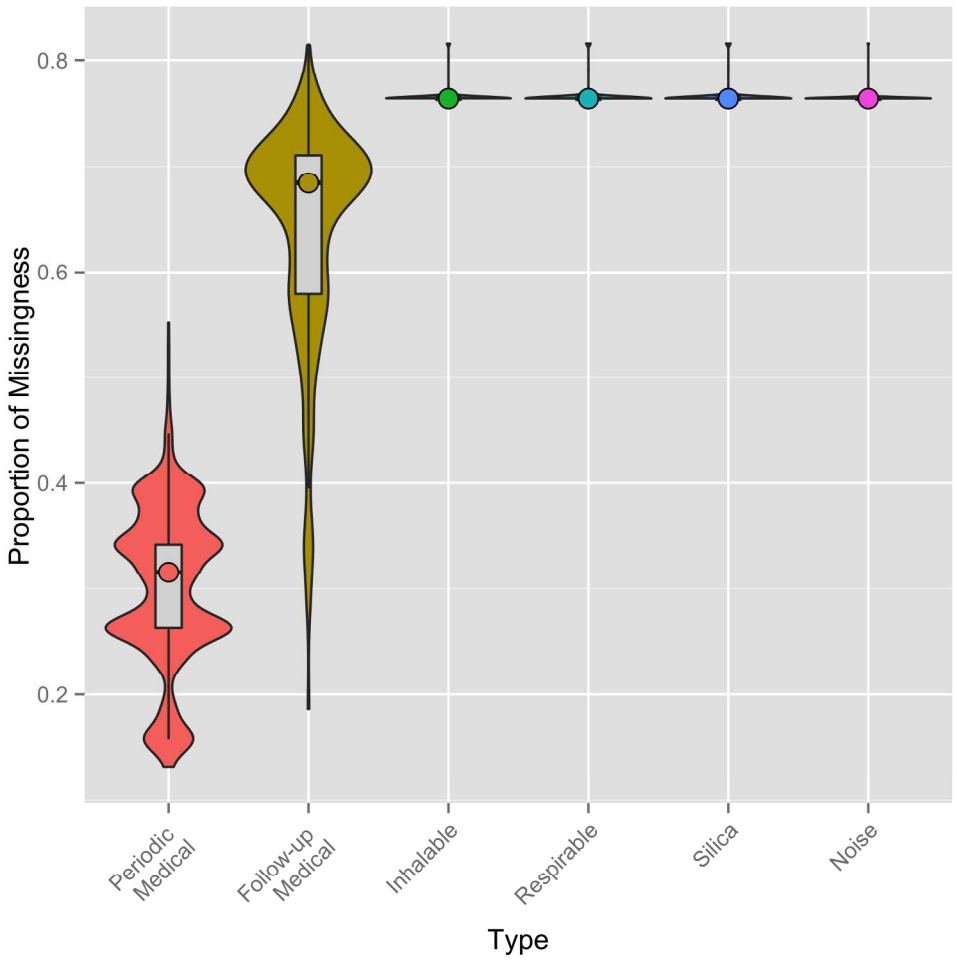
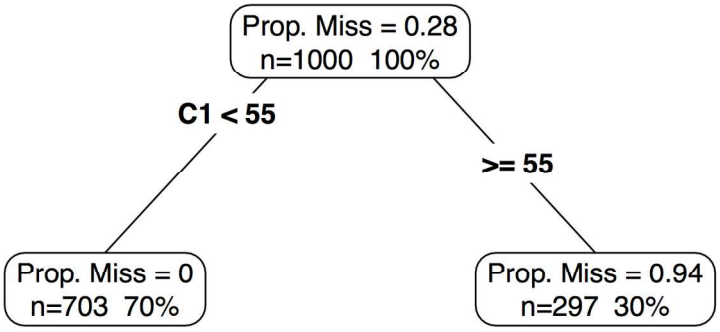
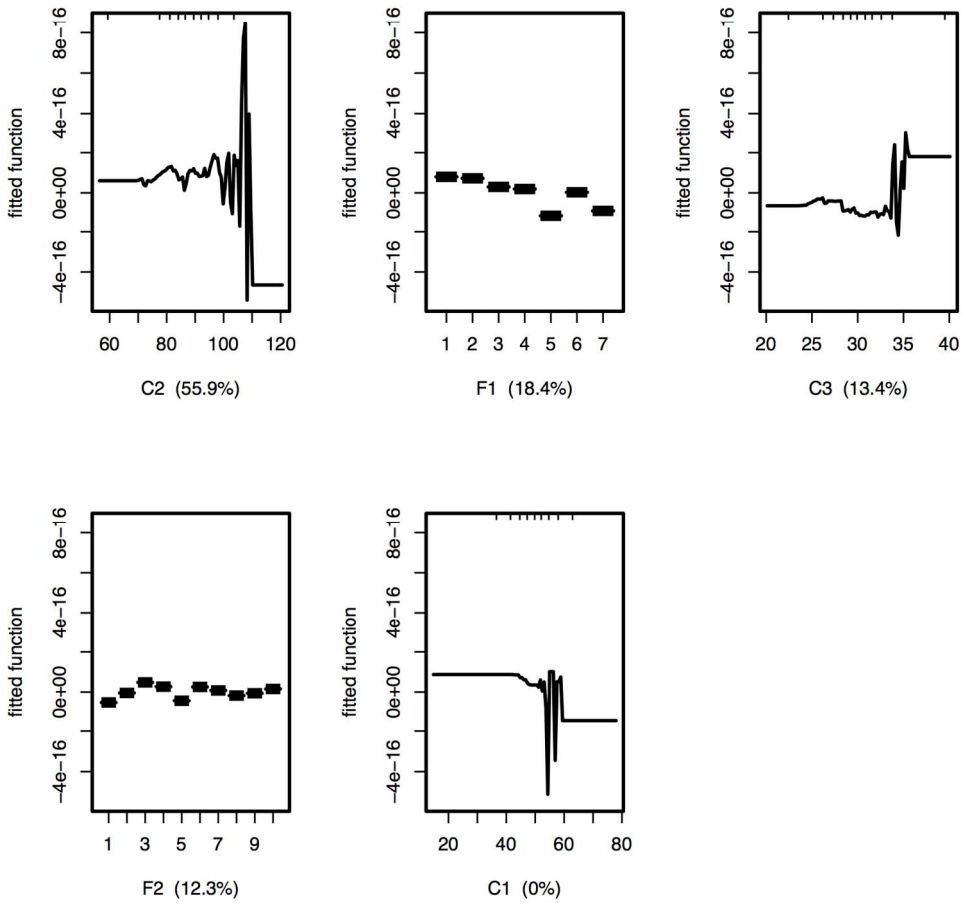


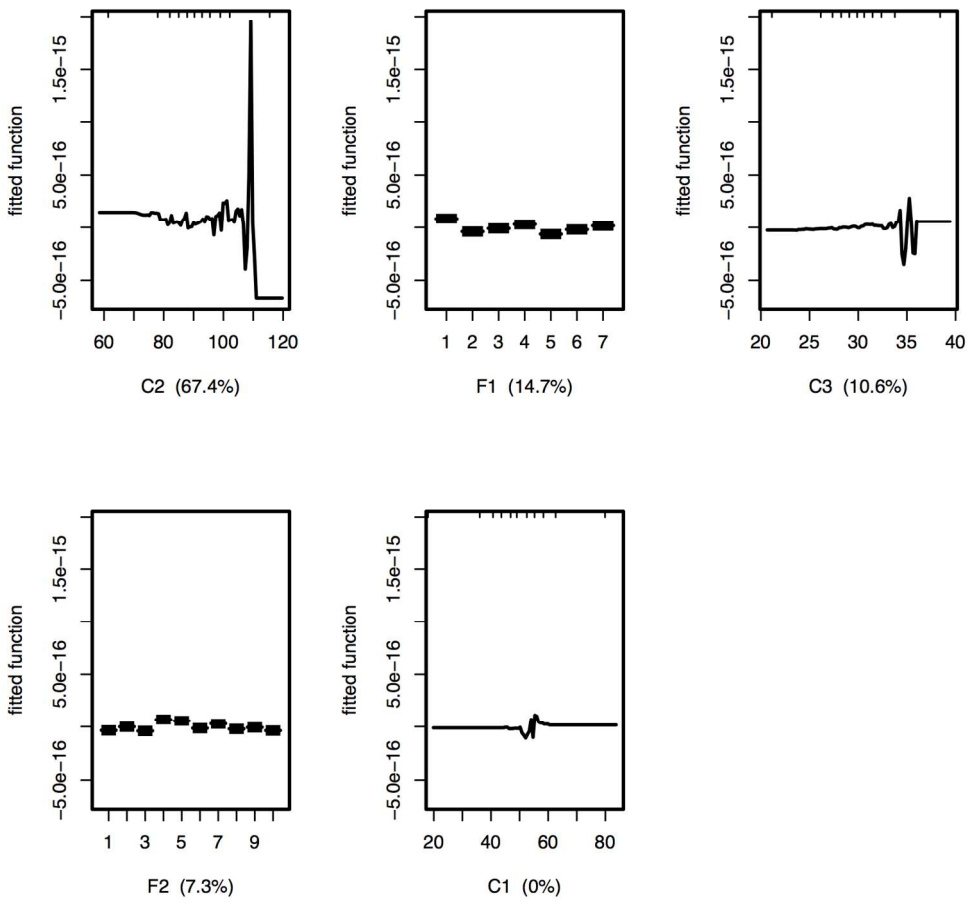
Supplementary Figure 1. Proportion of missing data per row based on the different data Types.



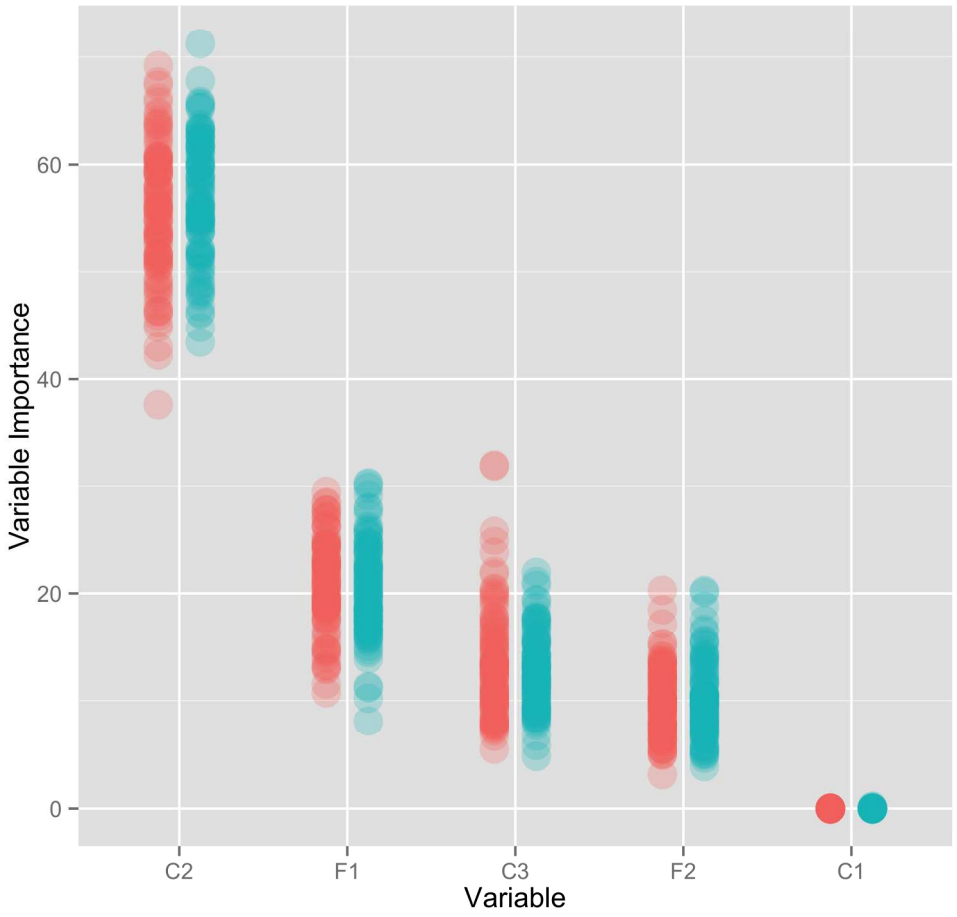
Supplementary Figure 2. Illustrative CART model based on the simulated data in the simulation study. The three numbers in each oval indicate the expected proportion of missing data (Prop. Miss) per row of data, the number of rows (n) and the percentage of total data (n%) in that node. All CART plots and summaries for conditions 1A and 1B can be extracted from the code provided in supplementary material.



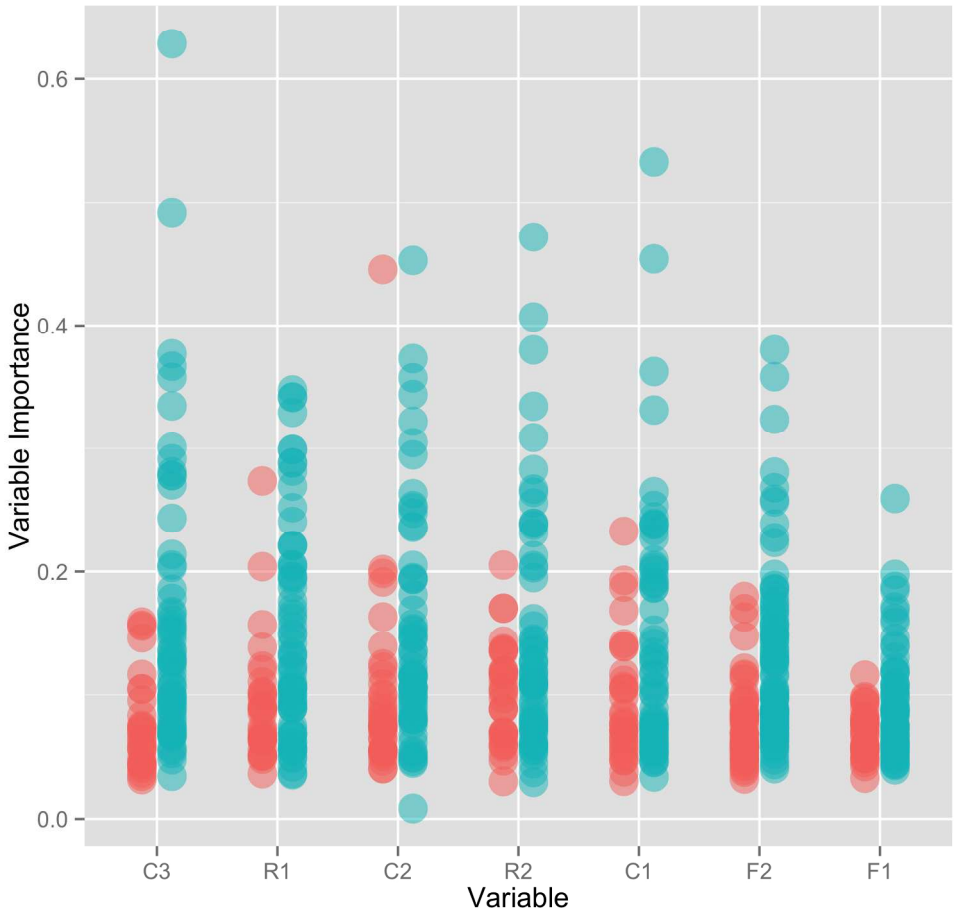
Supplementary Figure 3. Fitted function corresponding to the five variables considered in the simulation study, with the zero-point of the vertical axis indicating the model expected proportion of missingness, lines above 0.00 indicate more missingness than predicted, lines below indicate less missingness. In the top row (Part A) there is no missing data in variable C1, and in the bottom row (Part B) C1 is 50% MCAR.



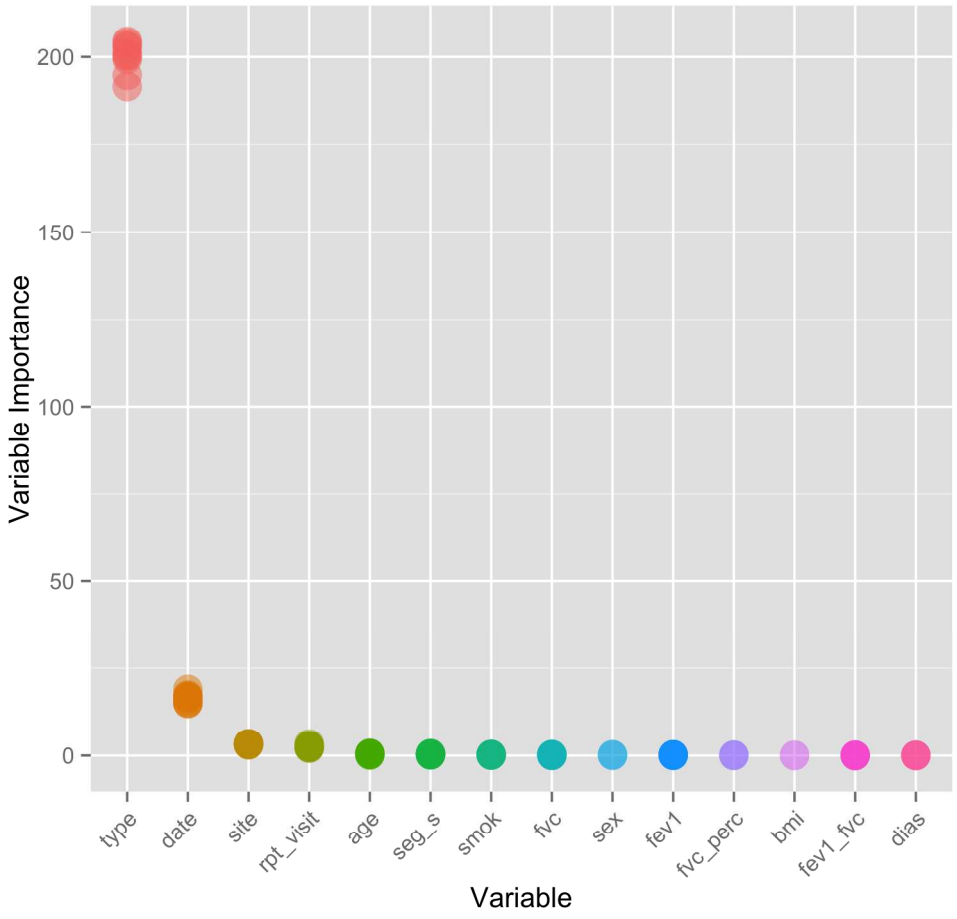
Supplementary Figure 3. Fitted function corresponding to the five variables considered in the simulation study, with the zero-point of the vertical axis indicating the model expected proportion of missingness, lines above 0.00 indicate more missingness than predicted, lines below indicate less missingness. In the top row (Part A) there is no missing data in variable C1, and in the bottom row (Part B) C1 is 50% MCAR.



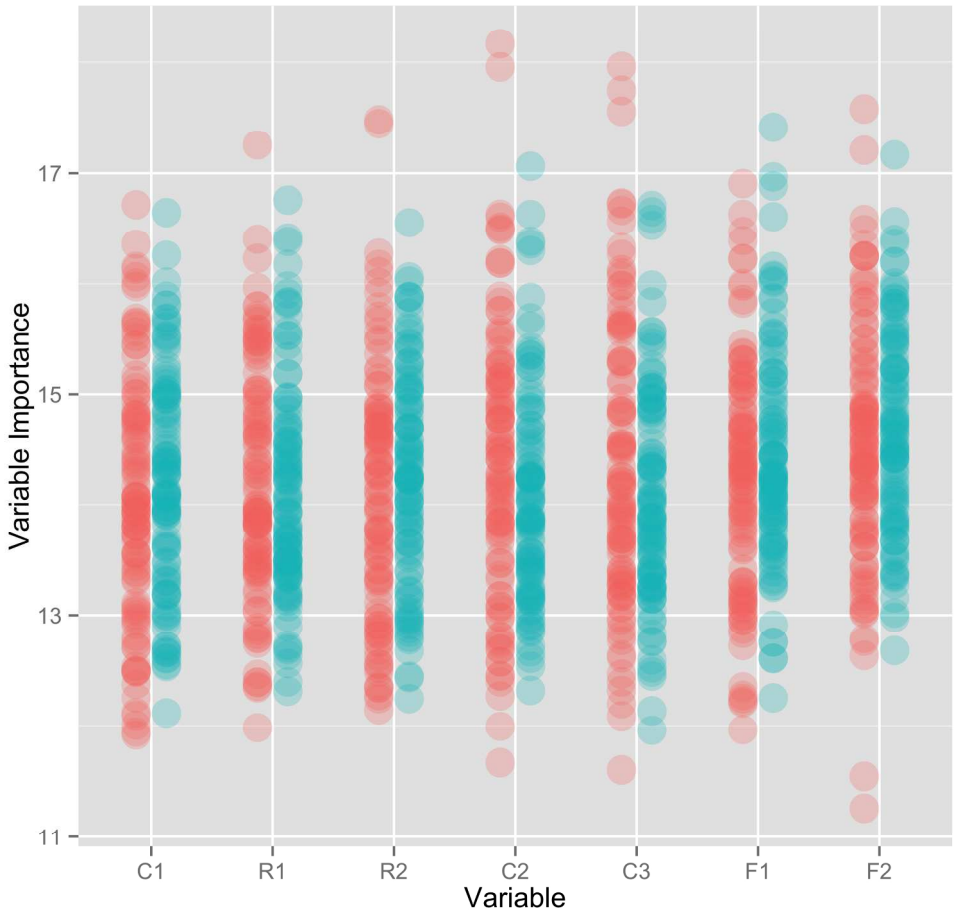
Supplementary Figure 4., Depicting the BRT model variable importance over all simulated datasets, where the red dots indicate when C1 is present, and the teal indicates when C1 is 50% MCAR.



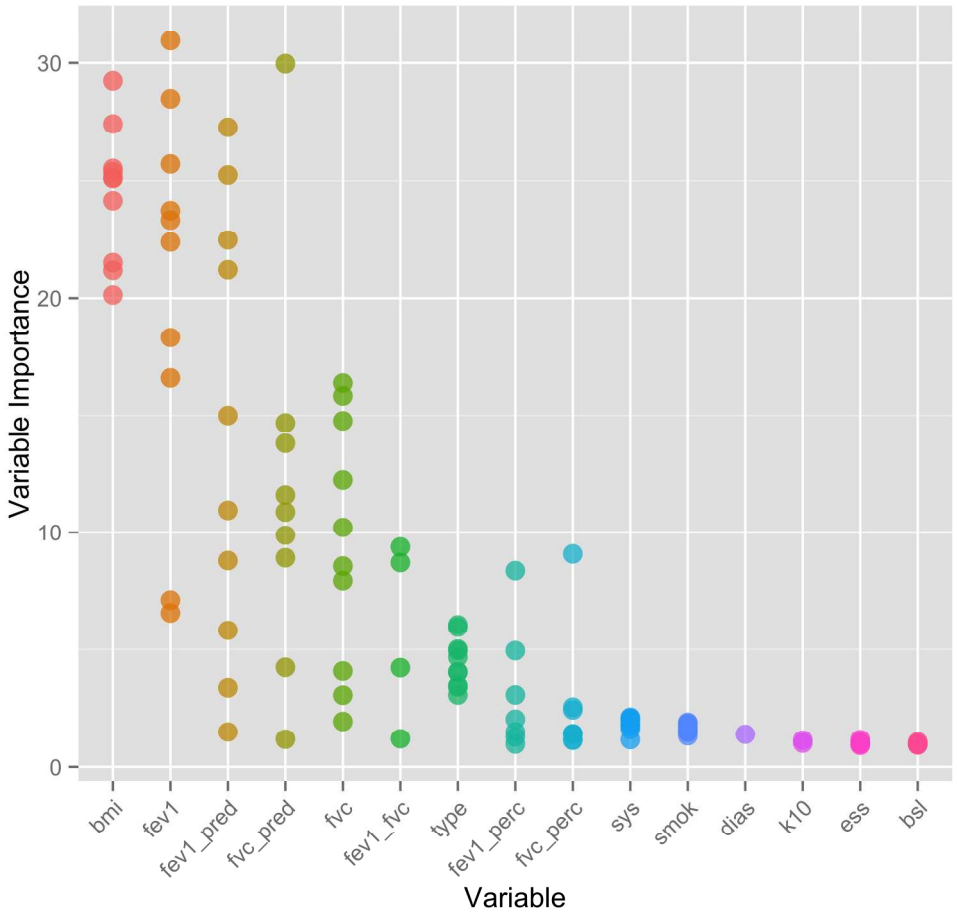
Supplementary Figure 5. Variable importance for the CART model in experiment 2, where the data is MCAR 20% (red) and MCAR 50% (blue).



Supplementary Figure 6. Variable importance for the CART model when data is 80% resampled (with replacement), points have 50% transparency to help display duplicate values.



Supplementary Figure 7. Variable Importance for the BRT model, for MCAR 20% and MCAR 50%, where the red points indicate MCAR 20%, and blue indicates MCAR 50% in experiment two.



Supplementary Figure 8. Variable Importance for the BRT model, when the data is resampled 80% (with replacement).

Variable	Detail
site	Site of work
uin	unique identifying number
sex	gender
type	type of data
date	date of examination
FVC	Forced Vital Capacity
FVC%	FVC Percent Predicted
FEV1	Forced Expiratory Volume in 1 second
FEV1%	FEV1 percent predicted
FEV1/FVC	FEV1 / FVC ratio
seg_p	Primary Similar Exposure Group
seg_s	Secondary Exposure Group
Age	Age at time of medical examination
BMI	Body Mass Index
Code	Medical Code
sys	Systolic blood pressure
dias	Diastolic blood pressure
hdl	High Density Lipoprotein Cholesterol
chol	Total Cholesterol
CRS	Cardiac Risk Score
smok	Smoking Status
ess	Epworth Sleeping Scale
k10	K10 Depression Score
etoh	Alcohol Audit Score
BHL	Binaural Hearing Loss
rep_vis	Number of medical Attendances
ex_per_week	Number of exercise sessions a week
weight	Weight
height	Height
waist	Waist Circumference
bsl	Blood Sugar Level
pulse	Pulse Rate (bpm)
conc	concentration of dust
laeq	Noise

Supplementary Table 1., A list of the variables used in the decision tree analysis, and their details.