Generate by Geneseng App

Tutorial #1

GENESENG

19 April, 2022

Table of Contents

Materials & Methods

Descriptive analysis

The tabular dataset was composed of **32** individuals and **7** variables including **6** biomarkers and a response variable called **Target** (**section Preliminary statistics - Global statistics**). Biomarkers were expressed using many settings such as median (and interquartile range) or mean (and standard deviation). Shapiro-Wilk's test was used to evaluate biomarker normality distribution (**section Preliminary statistics - Statistics by group**). Principal Component Analysis (PCA) was used to evaluate relationship between all continous variables (**section Variable Contribution (PCA)**). Subjects were divided into **2** groups according to **Target** and compared using non-exhaustive list of tests (**section Statistical tests**). Continuous relationship was assessed between all biomarkers using Spearman rank correlation coefficient. Biserial point correlation was performed to evaluate correlation between continuous and categorical variables. Spearman rank correlation coefficient was measures too. To categorical variables, relationship was measured using chi-squared test (**section Correlations**). Logistic regression was performed to evaluate biomarker individual performances. Youden or Closest topleft thresholds were used to maximize metrics such as AUC or Accuracy (**section Individual performances**).

A p-value < 0.05 was considered statistically significant.

Machine learning

For **Classification** problem, the tabular dataset was split in a training set (train) and a validation set (test) according to a **60:40** ratio. Rows containing missing valyes were removed from the training set. Any Biomarker combinations were computed and evaluated using simple machine learning algorithms such as **logistic regression**, **naive bayes**, **k-nearest neighbors**, **decision three**, **random forest** and **XGBoost**. Biomarker combinatory performances were indicated by **Accuracy** metric and compared to test set.

The current analysis report was fully generated using Geneseng app.

Preliminary statistics

Global statistics

Table 1: Summary statistics of variables

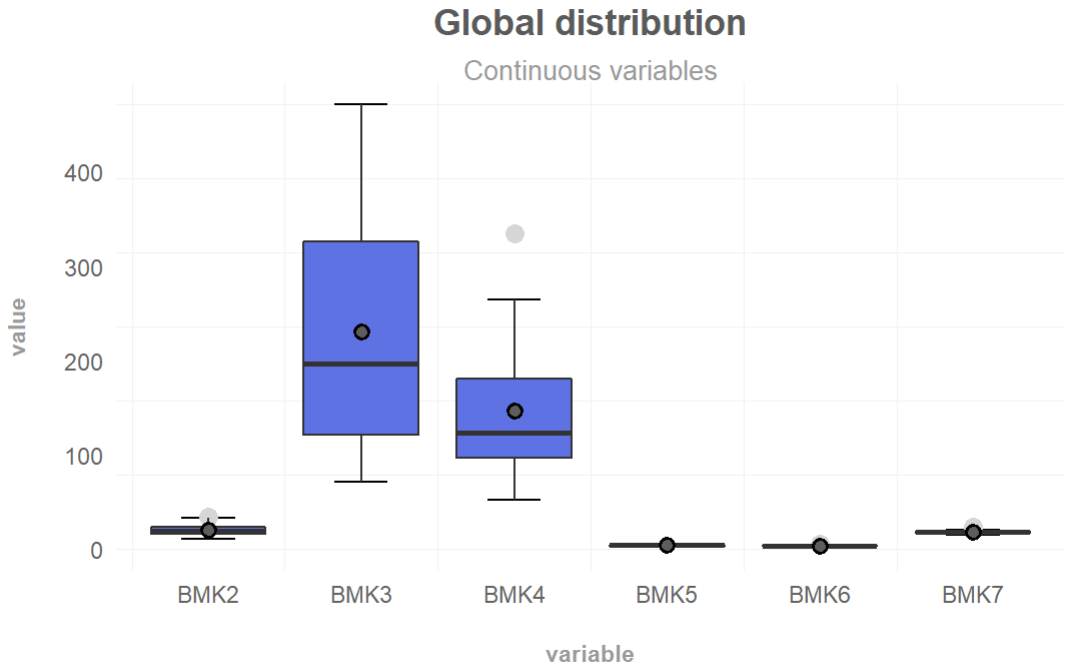
biomarker	n	n distinct	min	median	mean	sd	iqr	max	NA's[1]	Shapiro's test[2]	normality
BMK2	32	25	10.40	19.20	20.09	6.03	7.38	33.90	0	1.23e-01	no
BMK3	32	27	71.10	196.30	230.72	123.94	205.18	472.00	0	2.08e-02	no
BMK4	32	22	52.00	123.00	146.69	68.56	83.50	335.00	0	4.88e-02	no
BMK5	32	22	2.76	3.70	3.60	0.53	0.84	4.93	0	1.10e-01	no
BMK6	32	29	1.51	3.33	3.22	0.98	1.03	5.42	0	9.27e-02	no
BMK7	32	30	14.50	17.71	17.85	1.79	2.01	22.90	0	5.94e-01	yes

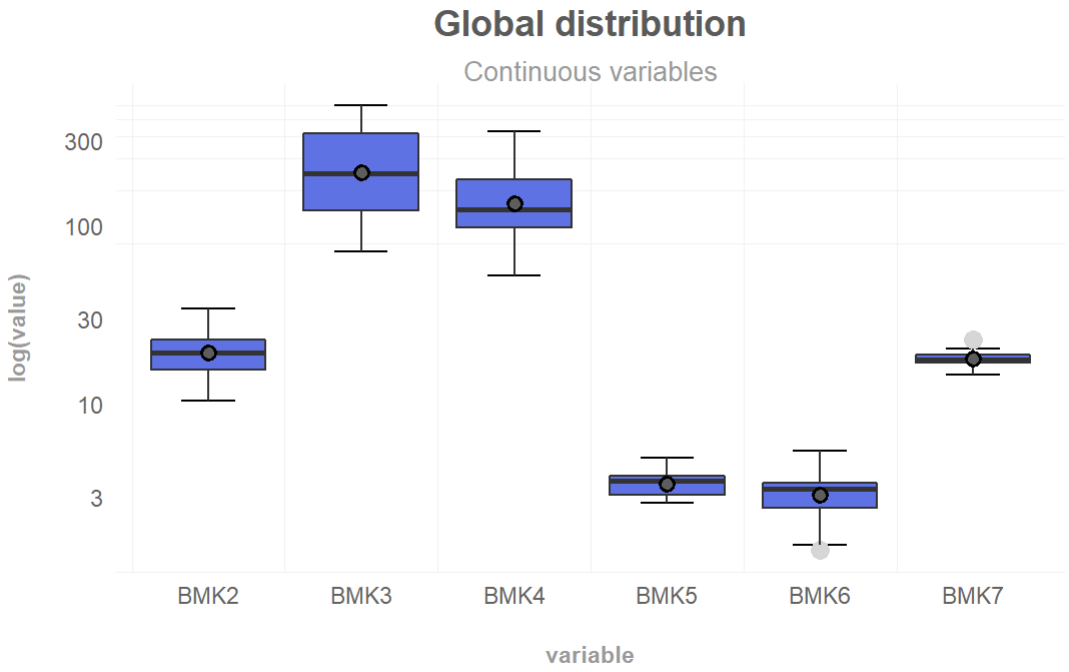
[1]Number of missing values

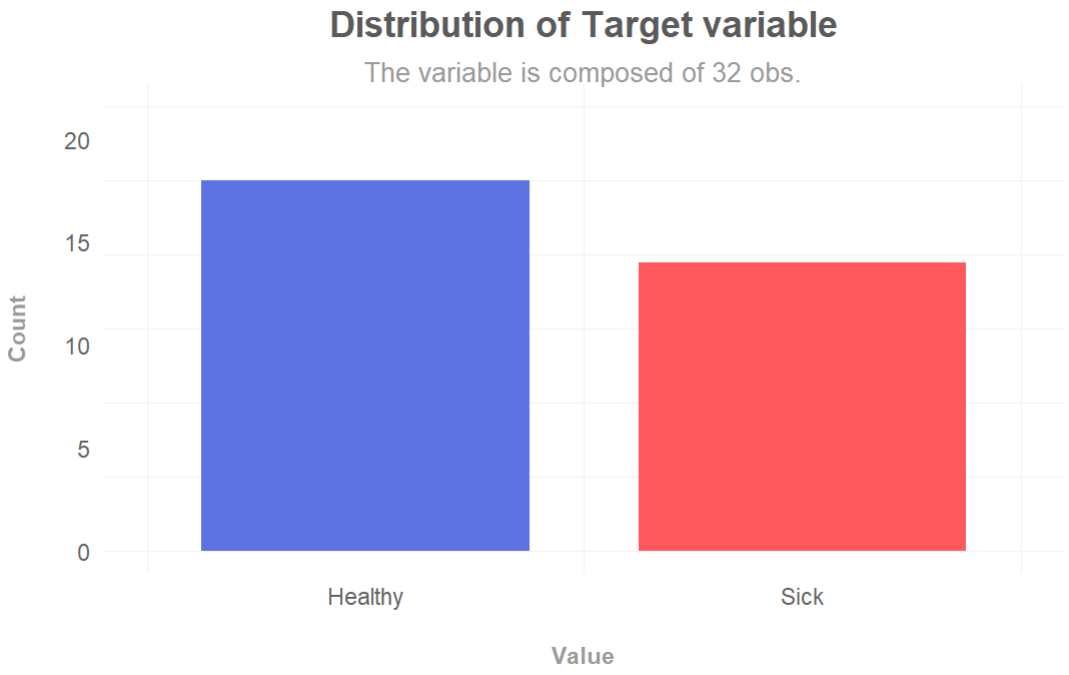
[2]Follow normal distribution if p-value > 0.05.

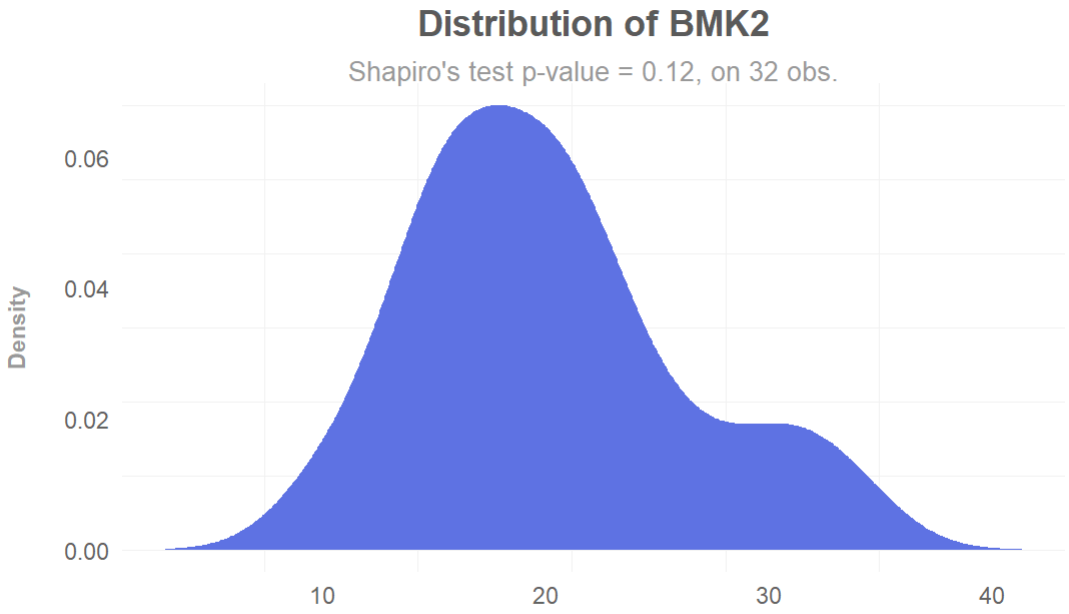
Table 2: Summary statistics of variables

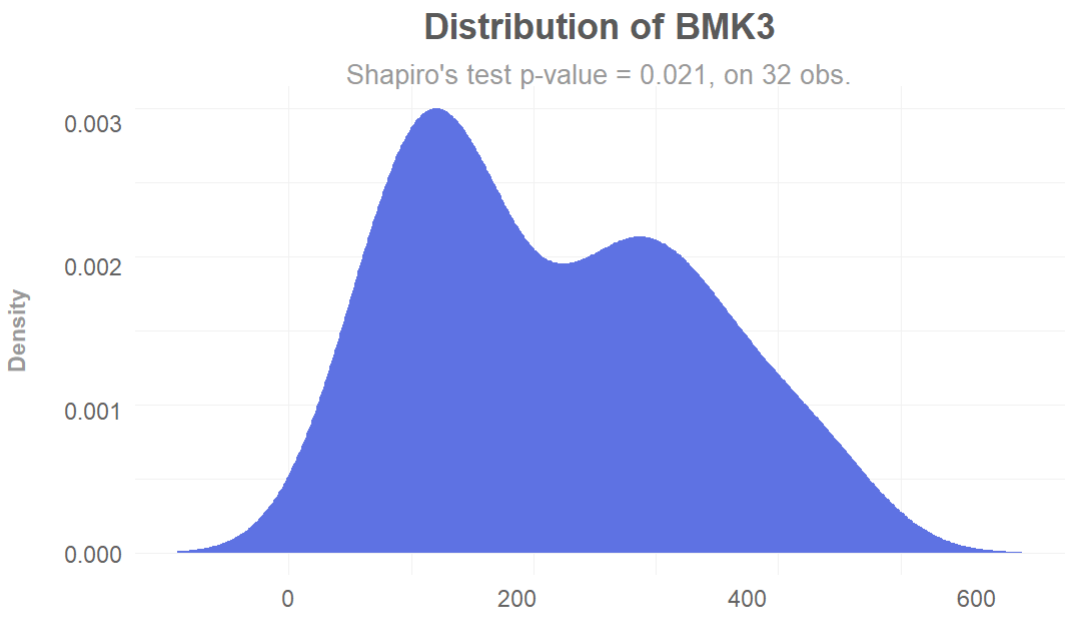
biomarker	value	n	percent
Target	Healthy	18	0.56
Target	Sick	14	0.44

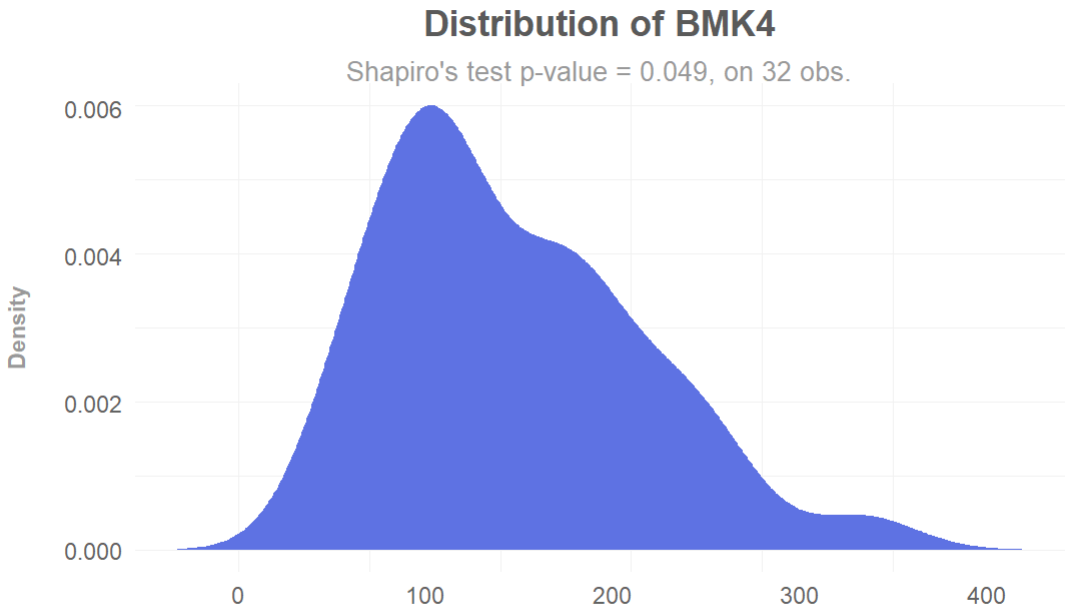

Global distribution
Continuous variables
value
400
300
200
100
0
BMK2
BMK3
BMK4
BMK5
BMK6
BMK7
variable

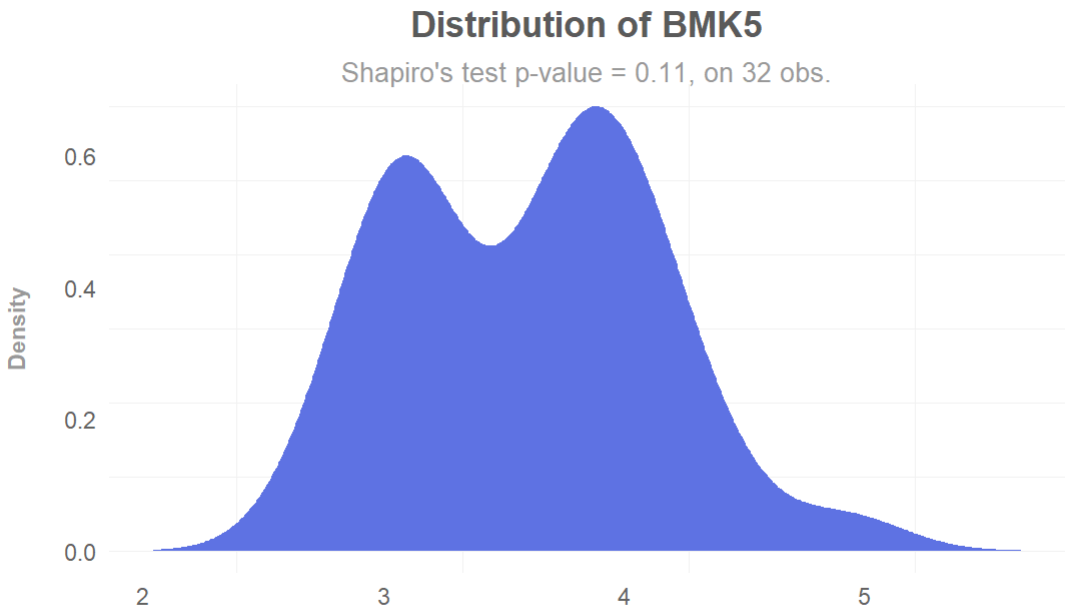

Global distribution
Continuous variables
log(value)
300
100
30
10
3
BMK2
BMK3
BMK4
BMK5
BMK6
BMK7
variable

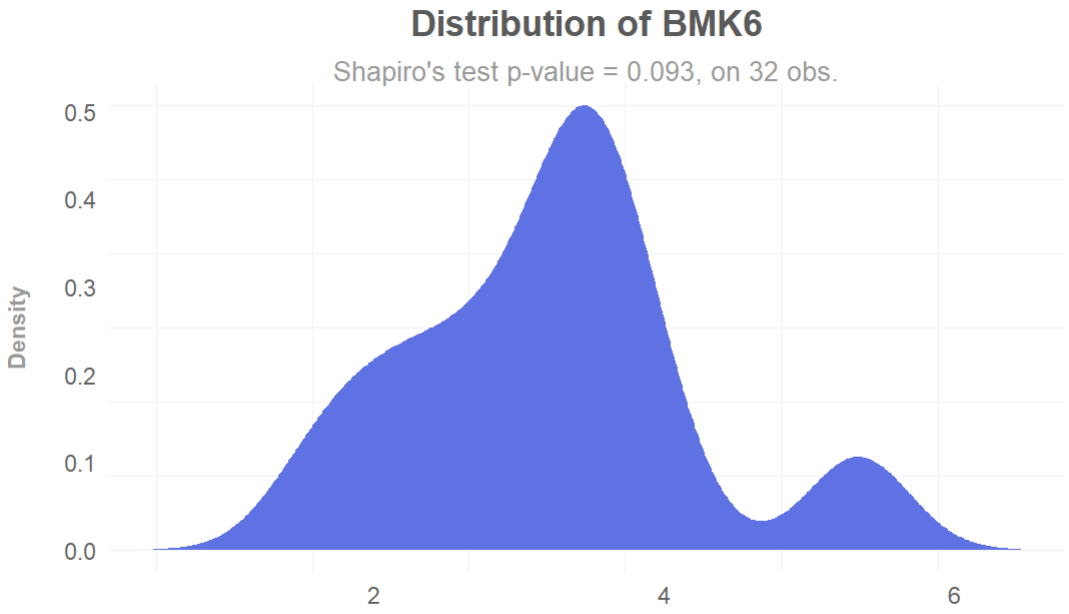

Distribution of Target variable
The variable is composed of 32 obs.
Count
20
15
10
5
0
Healthy
Sick
Value

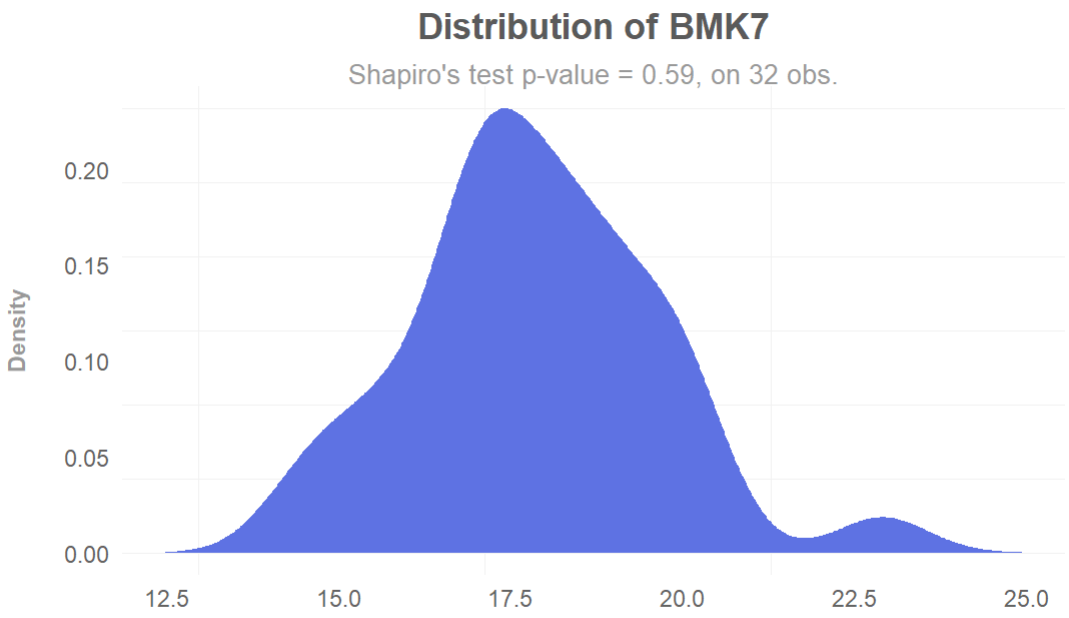

Distribution of BMK2
Shapiro's test p-value = 0.12, on 32 obs.
Density
0.06
0.04
0.02
0.00
10
20
30
40


Distribution of BMK3
Shapiro's test p-value = 0.021, on 32 obs.
Density
0.003
0.002
0.001
0.000
0
200
400
600


Distribution of BMK4
Shapiro's test p-value = 0.049, on 32 obs.
Density
0.006
0.004
0.002
0.000
0
100
200
300
400


Distribution of BMK5
Shapiro's test p-value = 0.11, on 32 obs.
Density
0.6
0.4
0.2
0.0
2
3
4
5


Distribution of BMK6
Shapiro's test p-value = 0.093, on 32 obs.
Density
0.5
0.4
0.3
0.2
0.1
0.0
2
4
6


Distribution of BMK7
Shapiro's test p-value = 0.59, on 32 obs.
Density
0.20
0.15
0.10
0.05
0.00
12.5
15.0
17.5
20.0
22.5
25.0

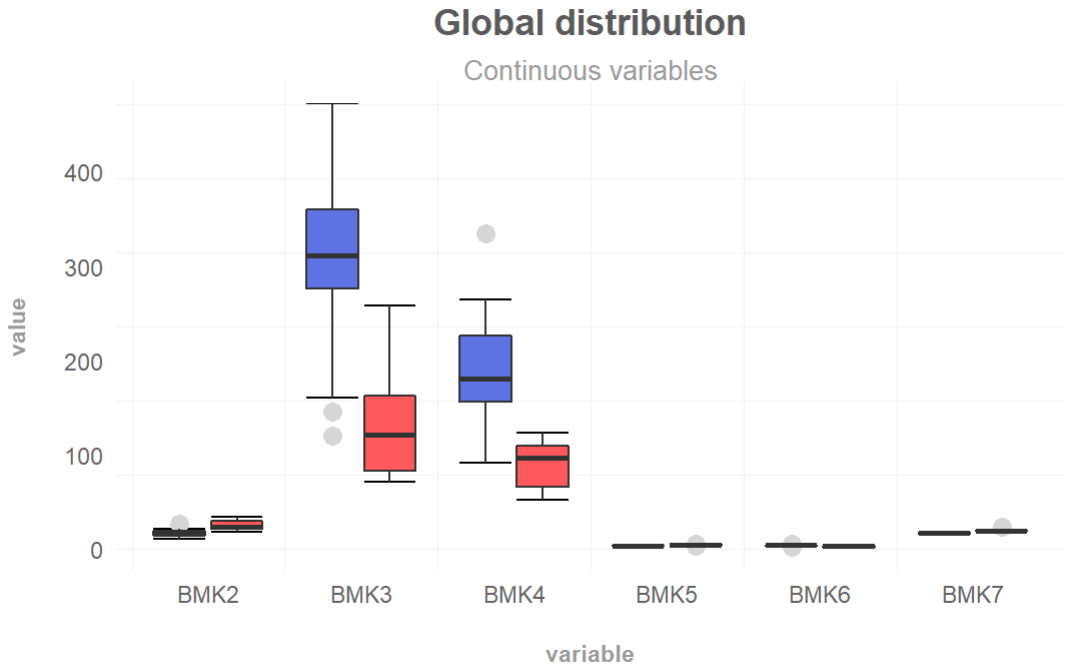
Statistics by group

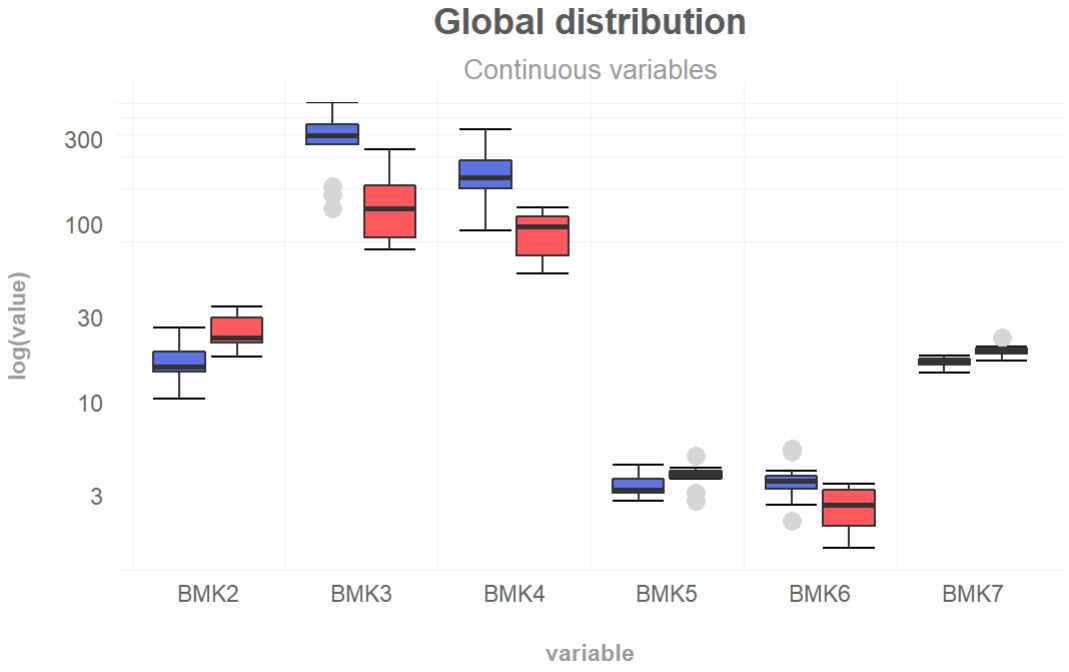
Table 3: Summary statistics by group

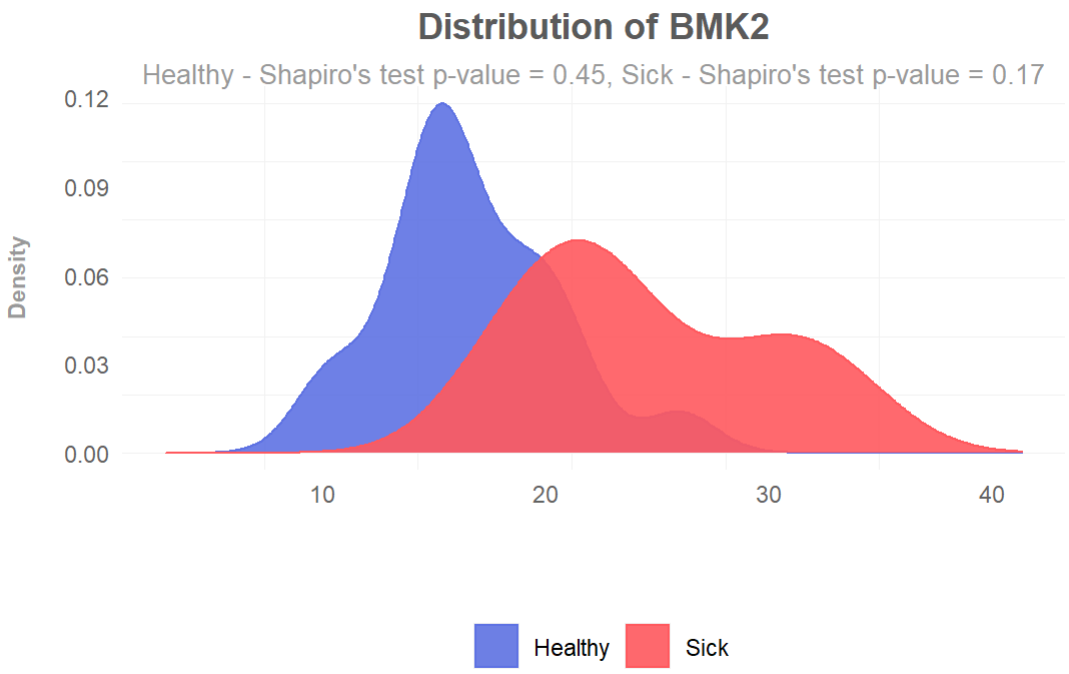
biomarker	group	n	n distinct	min	median	mean	sd	iqr	max	NA's[1]	Shapiro's test[2]	normality
BMK2	Healthy	18	15	10.40	15.65	16.62	3.86	4.30	26.00	0	4.49e-01	no
BMK2	Sick	14	11	17.80	22.80	24.56	5.38	8.22	33.90	0	1.67e-01	no
BMK3	Healthy	18	14	120.30	311.00	307.15	106.77	84.20	472.00	0	2.88e-01	no
BMK3	Sick	14	13	71.10	120.55	132.46	56.89	79.35	258.00	0	1.03e-01	no
BMK4	Healthy	18	11	91.00	180.00	189.72	60.28	70.00	335.00	0	5.60e-01	yes
BMK4	Sick	14	12	52.00	96.00	91.36	24.42	43.75	123.00	0	1.10e-01	no
BMK5	Healthy	18	14	2.76	3.18	3.39	0.47	0.63	4.43	0	4.52e-02	no
BMK5	Sick	14	11	2.76	3.92	3.86	0.51	0.36	4.93	0	1.25e-01	no
BMK6	Healthy	18	17	2.14	3.57	3.69	0.90	0.61	5.42	0	9.43e-02	no
BMK6	Sick	14	13	1.51	2.62	2.61	0.72	1.21	3.46	0	1.08e-01	no
BMK7	Healthy	18	17	14.50	17.02	16.69	1.09	1.42	18.00	0	5.93e-02	no
BMK7	Sick	14	13	16.90	19.17	19.33	1.35	1.37	22.90	0	9.63e-02	no

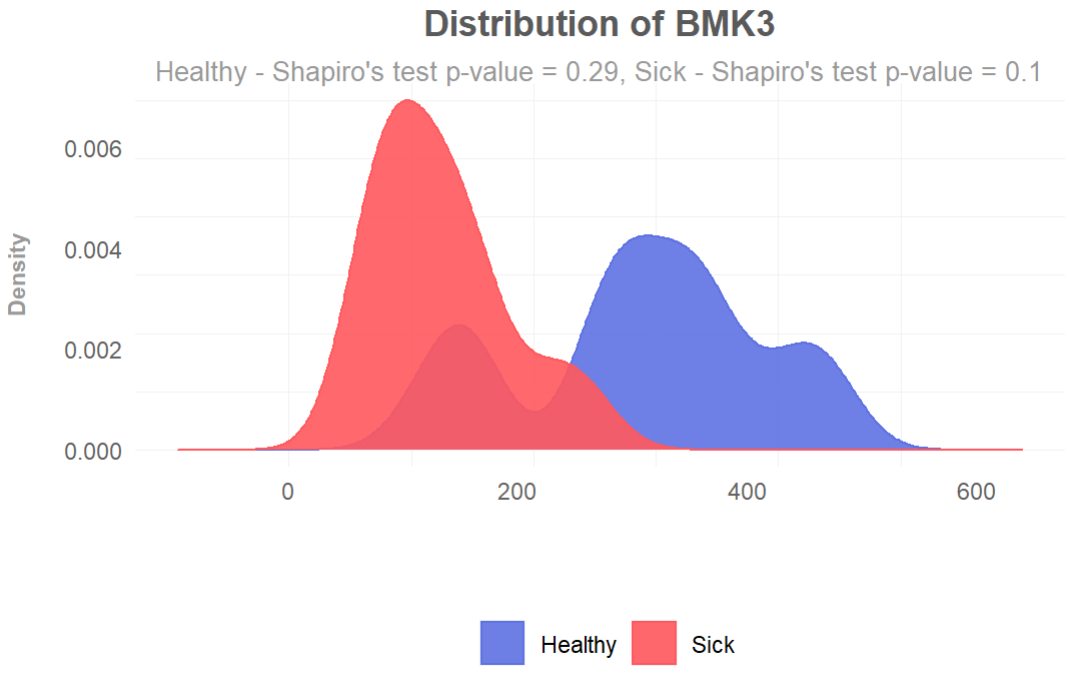
[1]Number of missing values

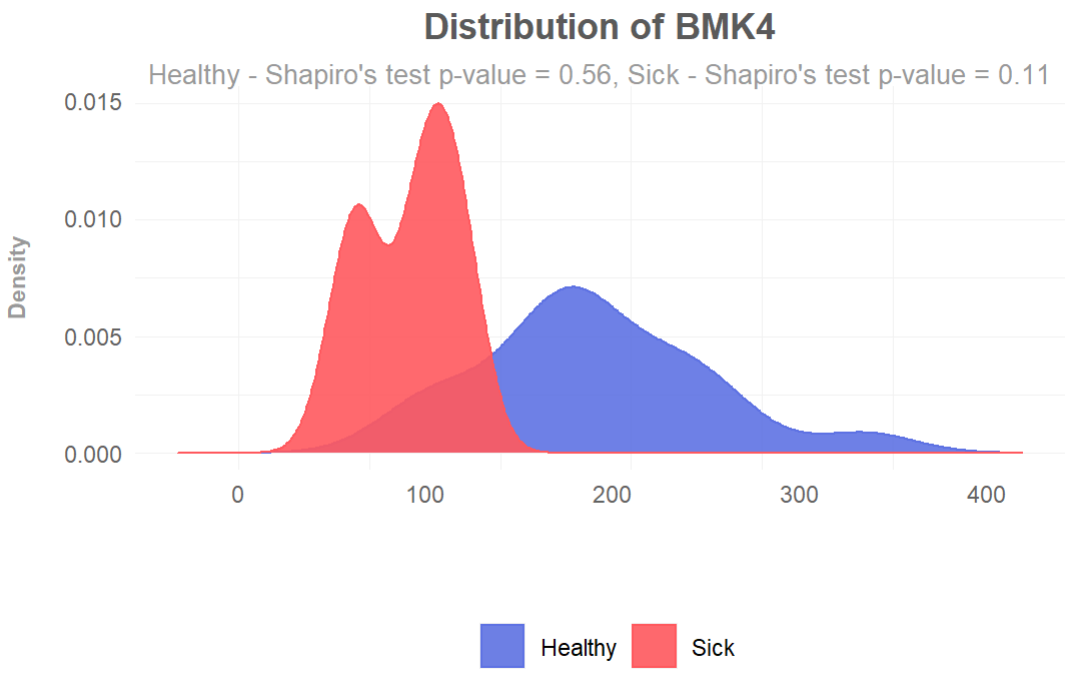
[2]Follow normal distribution if p-value > 0.05.

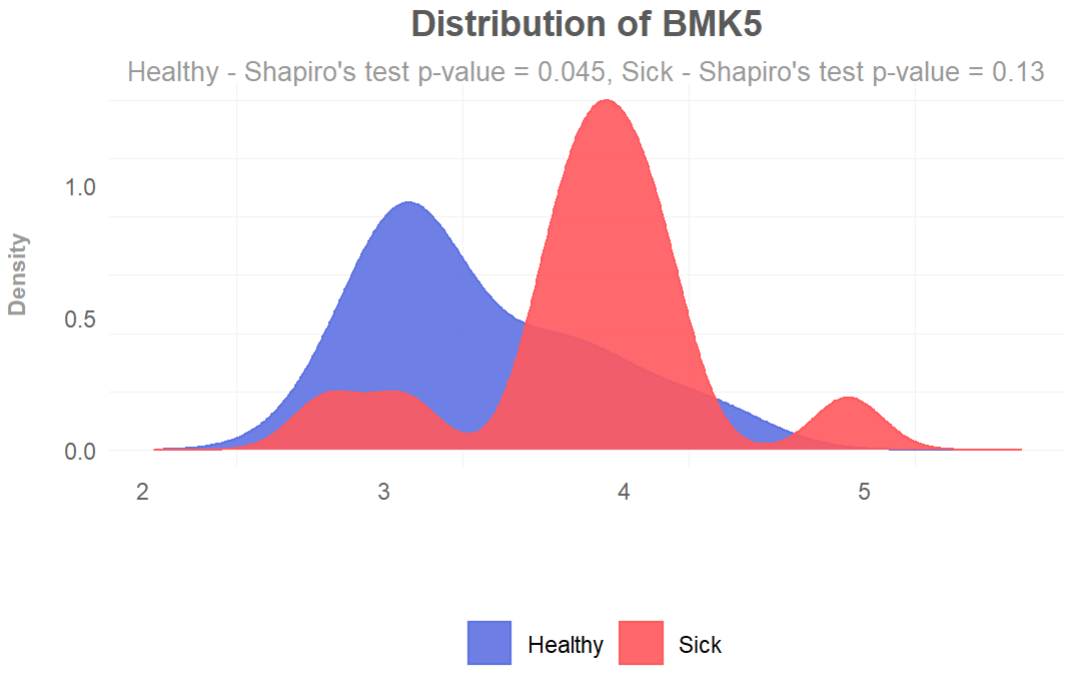

Global distribution
Continuous variables
value
400
300
200
100
0
BMK2
BMK3
BMK4
BMK5
BMK6
BMK7
variable

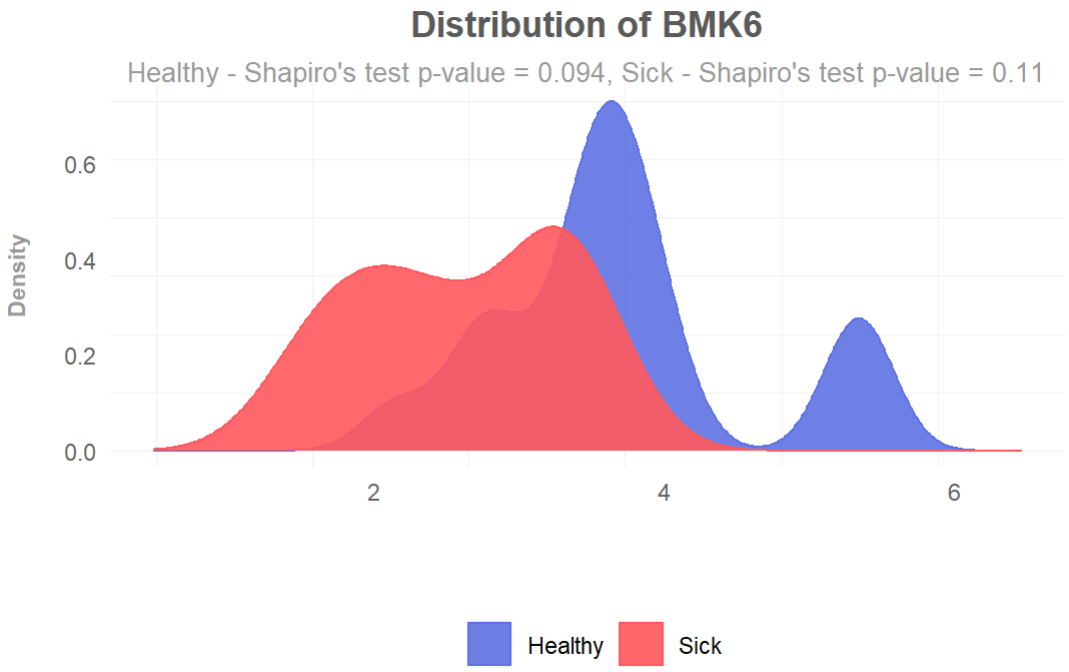

Global distribution
Continuous variables
log(value)
300
100
30
10
3
BMK2
BMK3
BMK4
BMK5
BMK6
BMK7
variable

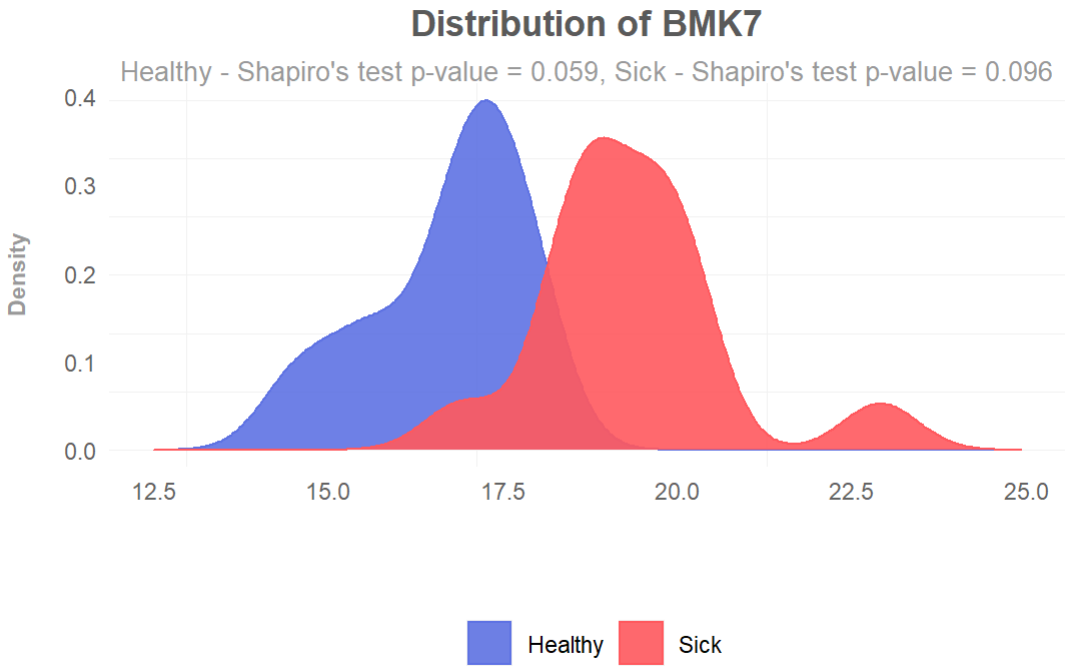

Distribution of BMK2
Healthy - Shapiro's test p-value = 0.45, Sick - Shapiro's test p-value = 0.17
Density
0.12
0.09
0.06
0.03
0.00
10
20
30
40
Healthy
Sick


Distribution of BMK3
Healthy - Shapiro's test p-value = 0.29, Sick - Shapiro's test p-value = 0.1
Density
0.006
0.004
0.002
0.000
0
200
400
600
Healthy
Sick


Distribution of BMK4
Healthy - Shapiro's test p-value = 0.56, Sick - Shapiro's test p-value = 0.11
Density
0.015
0.010
0.005
0.000
0
100
200
300
400
Healthy
Sick


Distribution of BMK5
Healthy - Shapiro's test p-value = 0.045, Sick - Shapiro's test p-value = 0.13
Density
1.0
0.5
0.0
2
3
4
5
Healthy
Sick

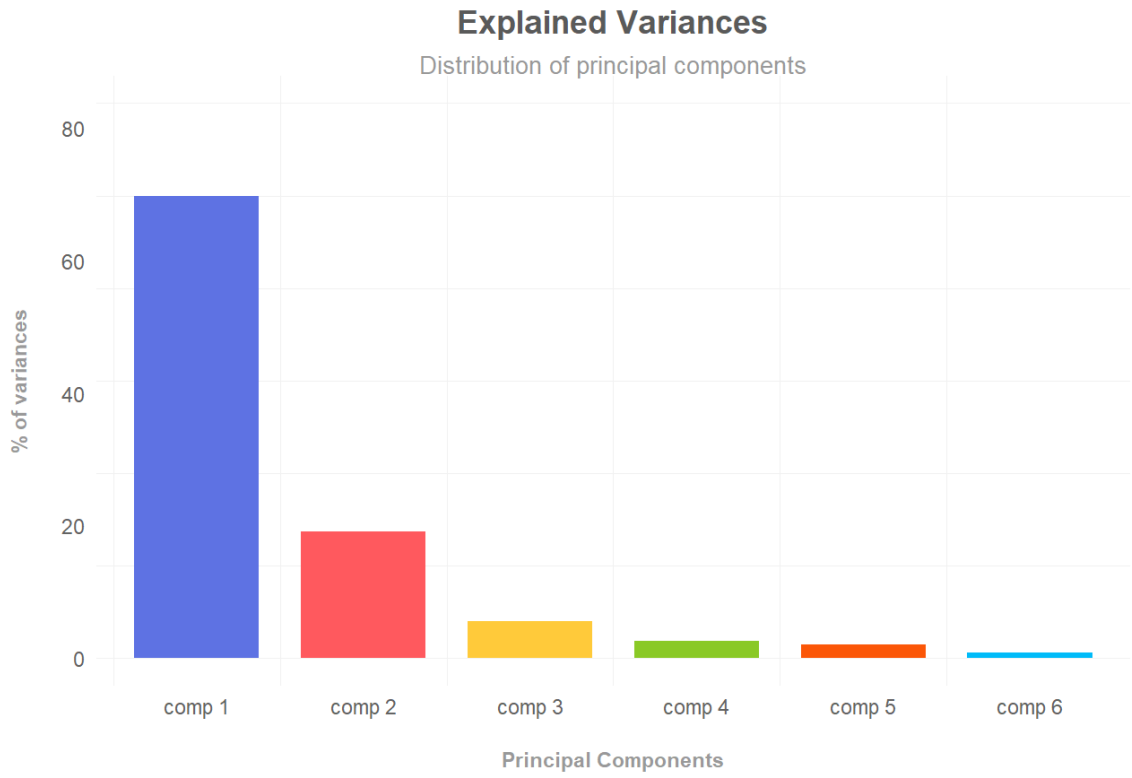

Distribution of BMK6
Healthy - Shapiro's test p-value = 0.094, Sick - Shapiro's test p-value = 0.11
Density
0.6
0.4
0.2
0.0
2
4
6
Healthy
Sick

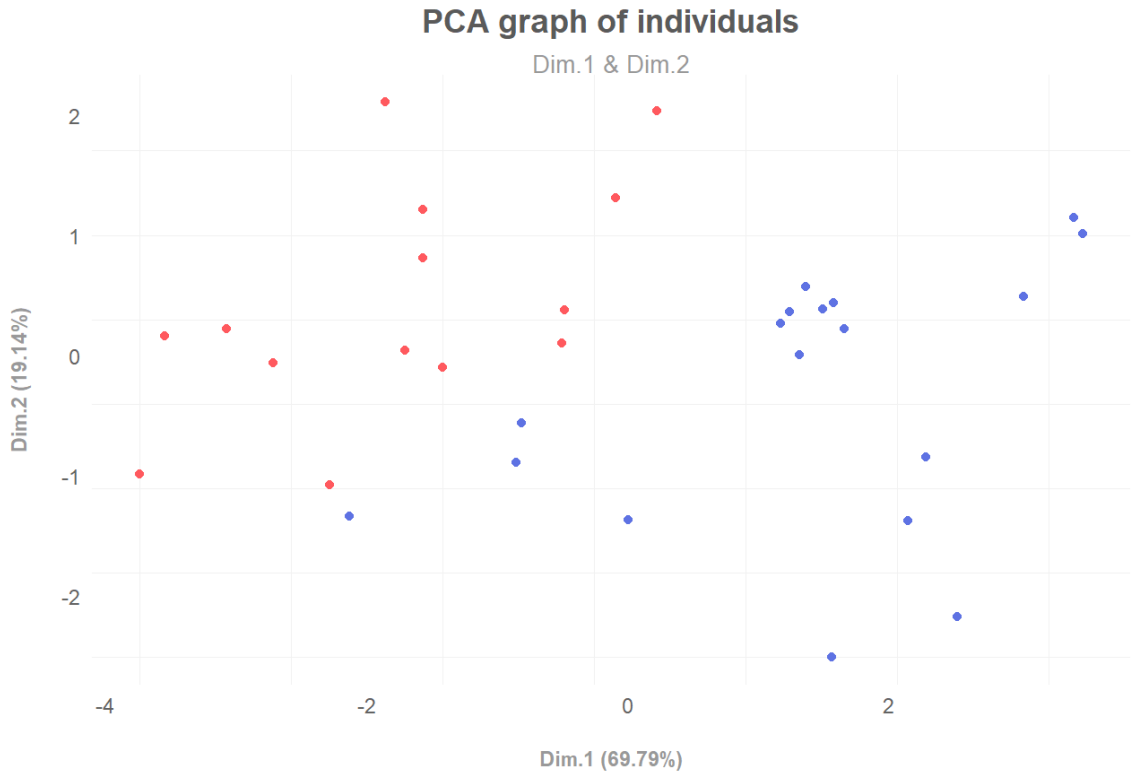

Distribution of BMK7
Healthy - Shapiro's test p-value = 0.059, Sick - Shapiro's test p-value = 0.096
Density
0.4
0.3
0.2
0.1
0.0
12.5
15.0
17.5
20.0
22.5
25.0
Healthy
Sick

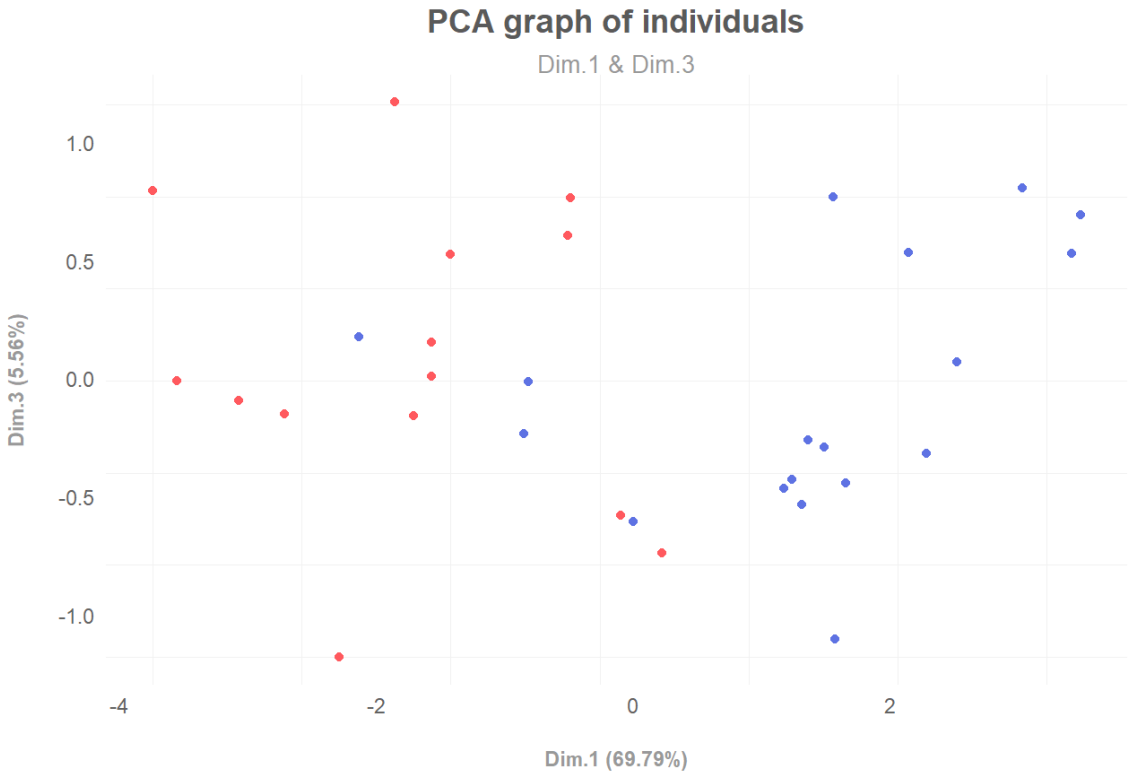
Variable Contribution (PCA)

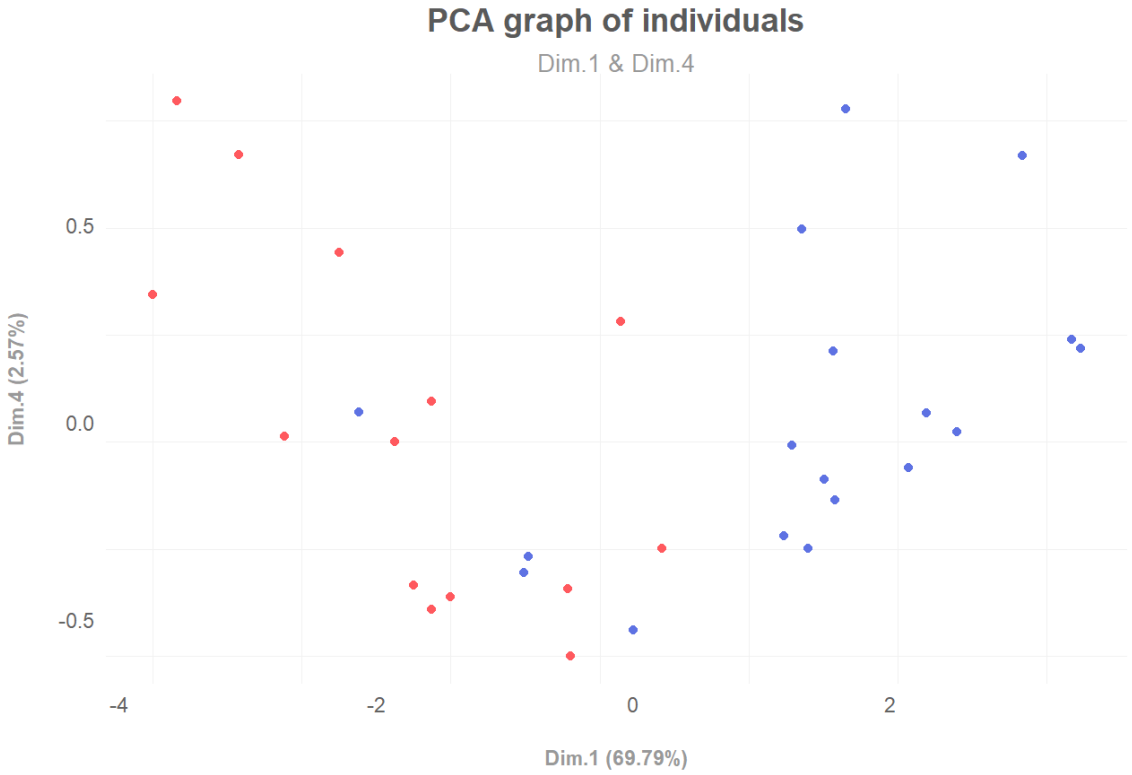
Table 4: Principal Component Analysis - Variable contribution (%)

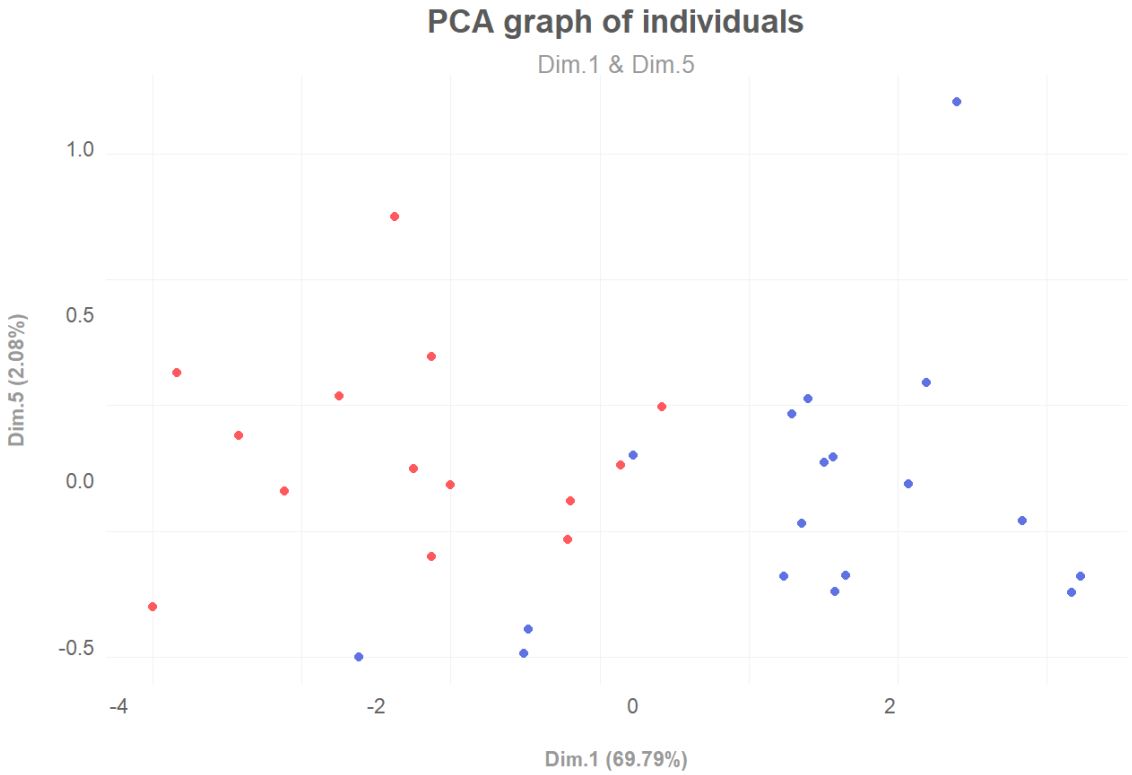
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
BMK2	21.04	0.34	3.79	61.16	1.24
BMK3	21.72	0.37	0.94	36.00	8.68
BMK4	18.14	13.07	2.14	1.51	64.92
BMK5	13.48	19.06	64.08	0.05	2.07
BMK6	19.24	8.97	17.45	1.09	5.30
BMK7	6.39	58.19	11.60	0.18	17.80

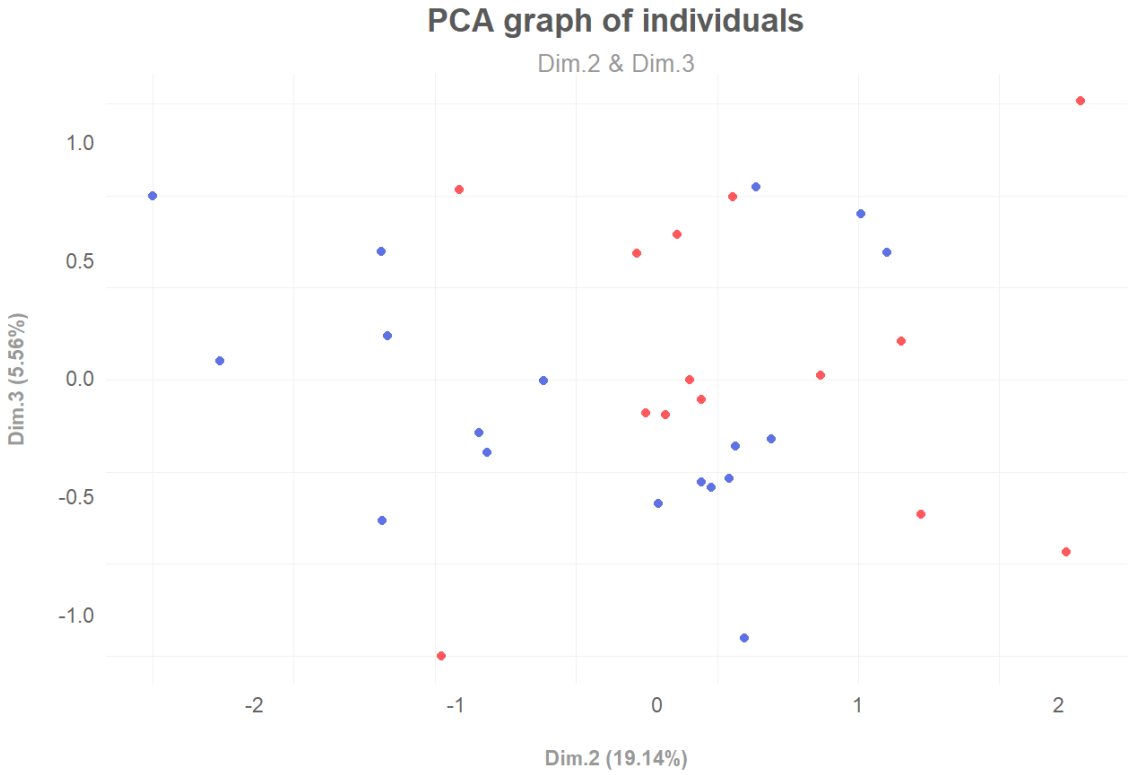

Explained Variances
Distribution of principal components
80
60
40
20
0
% of variances
comp 1
comp 2
comp 3
comp 4
comp 5
comp 6
Principal Components

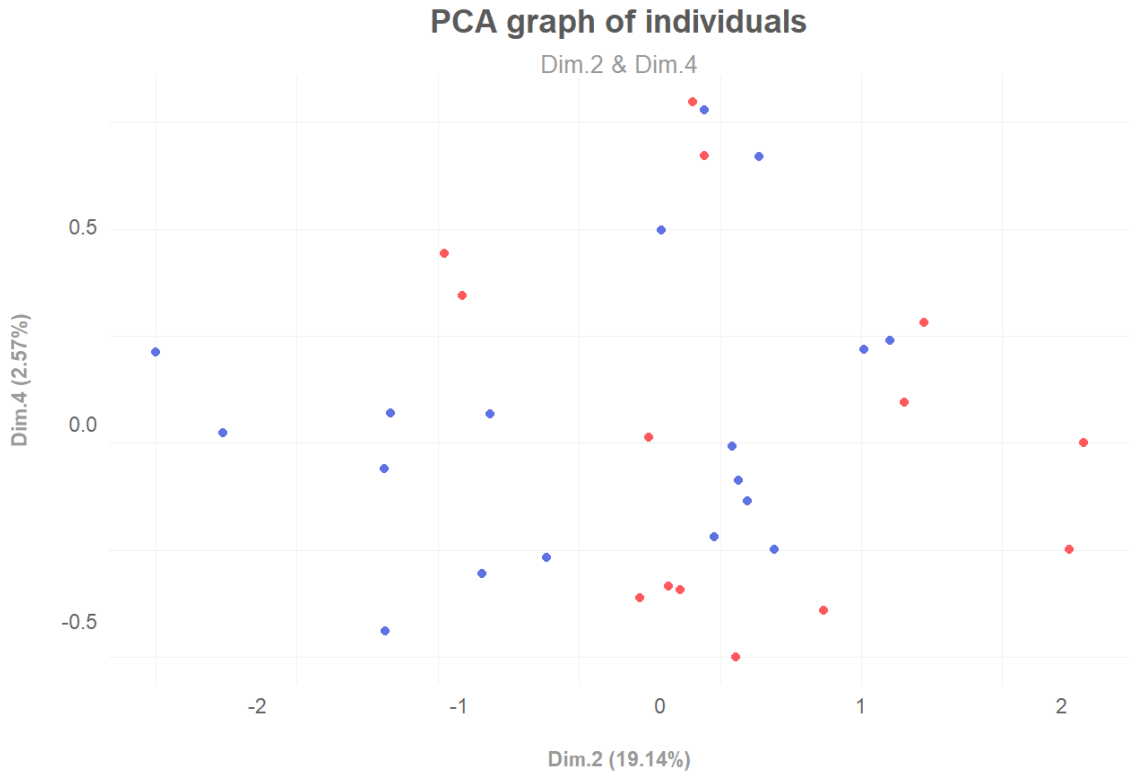

PCA graph of individuals
Dim.1 & Dim.2
2
1
0
-1
-2
Dim.2 (19.14%)
-4
-2
0
2
Dim.1 (69.79%)

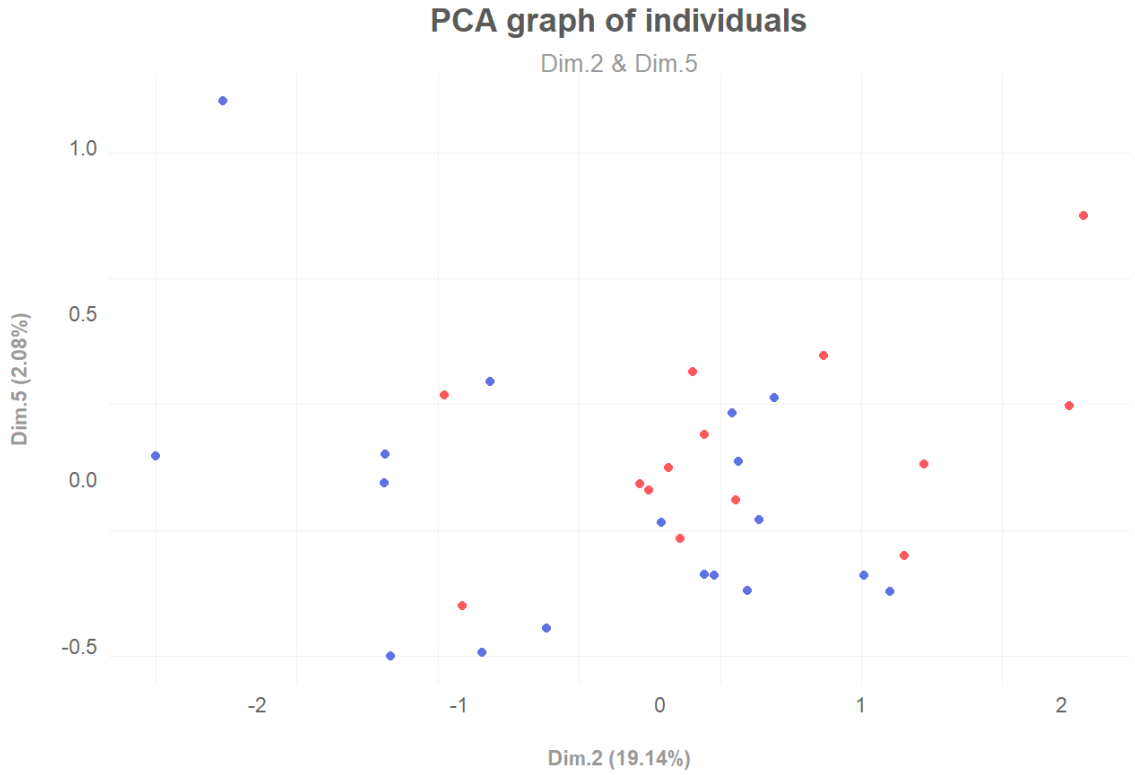

PCA graph of individuals
Dim.1 & Dim.3
1.0
0.5
0.0
-0.5
-1.0
Dim.3 (5.56%)
-4
-2
0
2
Dim.1 (69.79%)

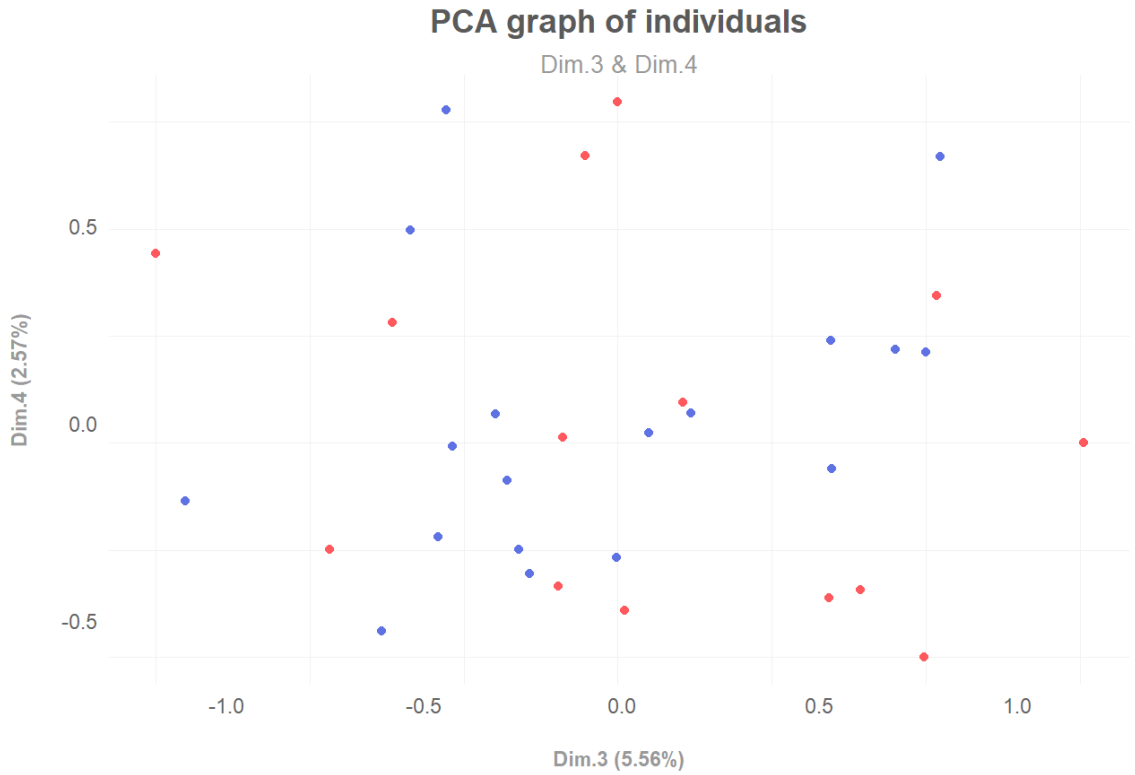

PCA graph of individuals
Dim.1 & Dim.4
0.5
0.0
-0.5
Dim.4 (2.57%)
-4
-2
0
2
Dim.1 (69.79%)

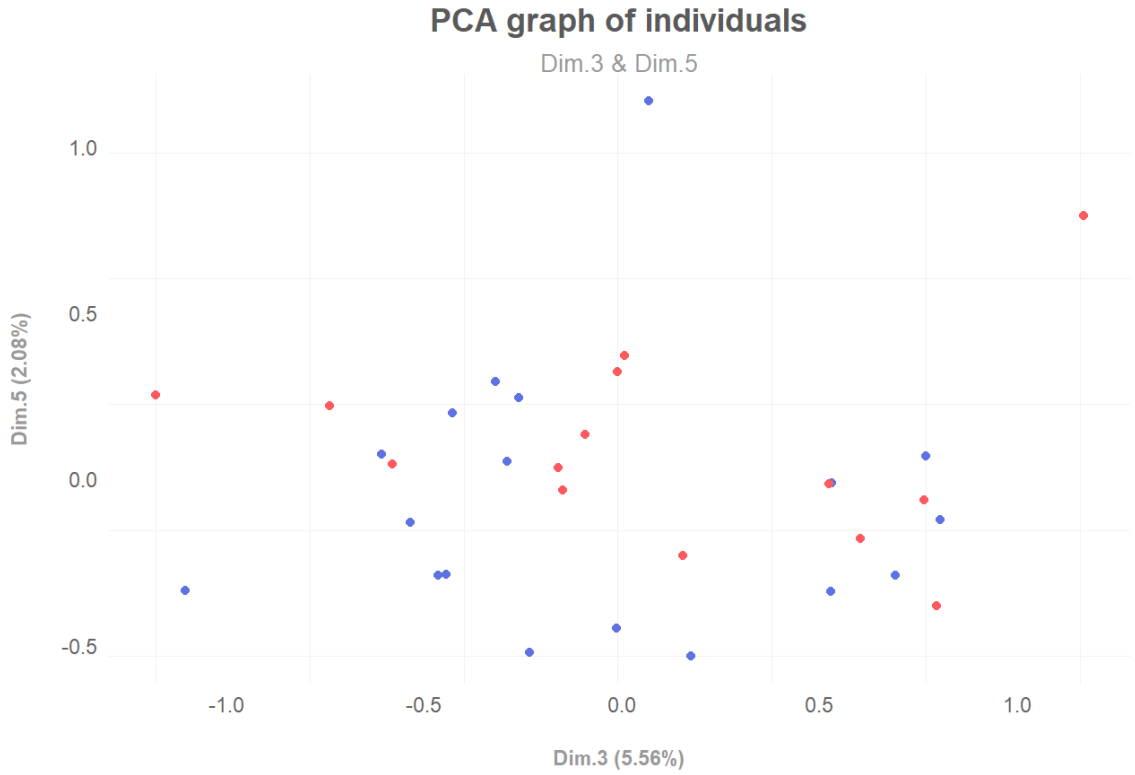

PCA graph of individuals
Dim.1 & Dim.5
1.0
0.5
0.0
-0.5
Dim.5 (2.08%)
-4
-2
0
2
Dim.1 (69.79%)

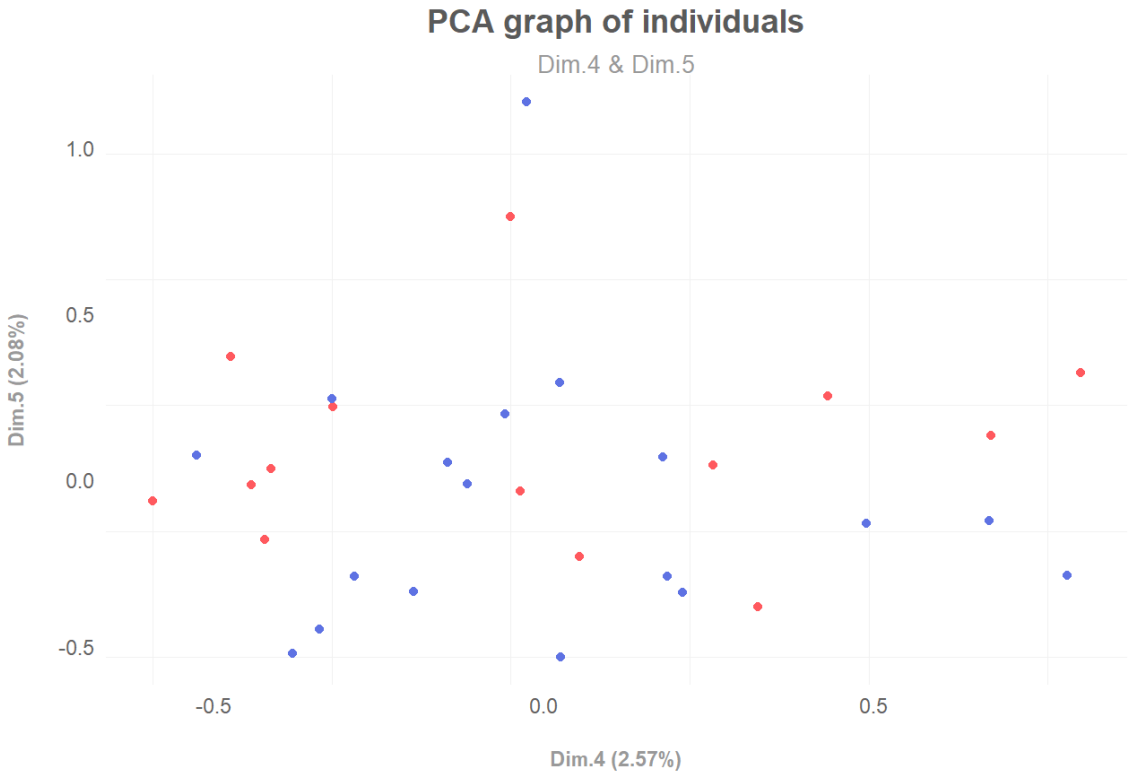

PCA graph of individuals
Dim.2 & Dim.3
1.0
0.5
0.0
-0.5
-1.0
Dim.3 (5.56%)
-2
-1
0
1
2
Dim.2 (19.14%)


PCA graph of individuals
Dim.2 & Dim.4
0.5
0.0
-0.5
Dim.4 (2.57%)
-2
-1
0
1
2
Dim.2 (19.14%)


PCA graph of individuals
Dim.2 & Dim.5
1.0
0.5
0.0
-0.5
Dim.5 (2.08%)
-2
-1
0
1
2
Dim.2 (19.14%)


PCA graph of individuals
Dim.3 & Dim.4
0.5
0.0
-0.5
Dim.4 (2.57%)
-1.0
-0.5
0.0
0.5
1.0
Dim.3 (5.56%)


PCA graph of individuals
Dim.3 & Dim.5
1.0
0.5
0.0
-0.5
Dim.5 (2.08%)
-1.0
-0.5
0.0
0.5
1.0
Dim.3 (5.56%)

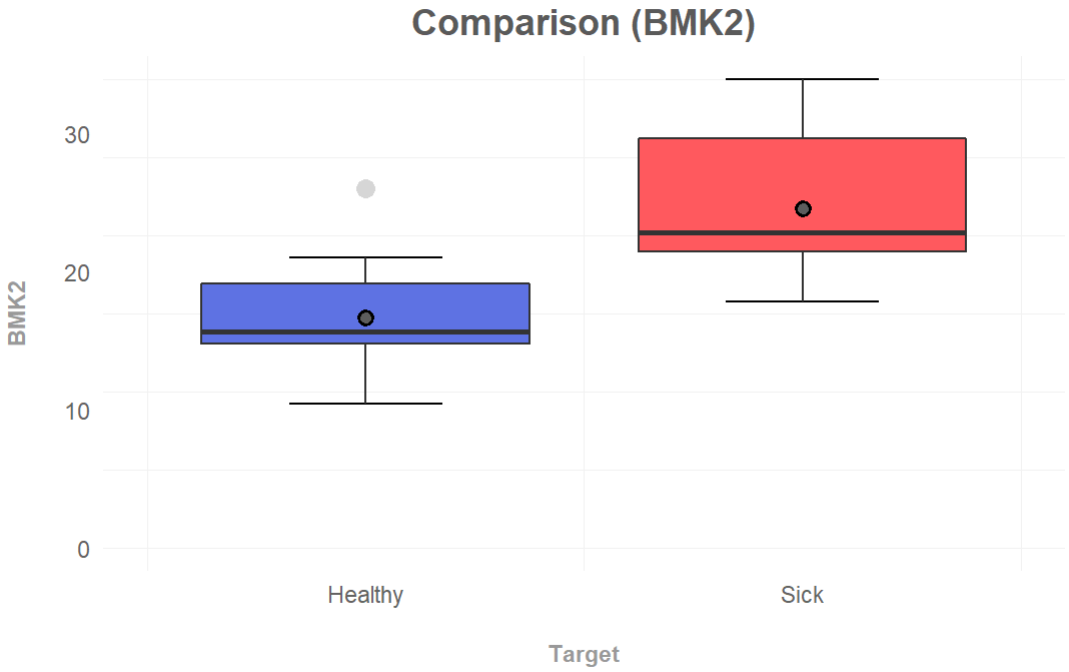

PCA graph of individuals
Dim.4 & Dim.5
1.0
0.5
0.0
-0.5
Dim.5 (2.08%)
-0.5
0.0
0.5
Dim.4 (2.57%)

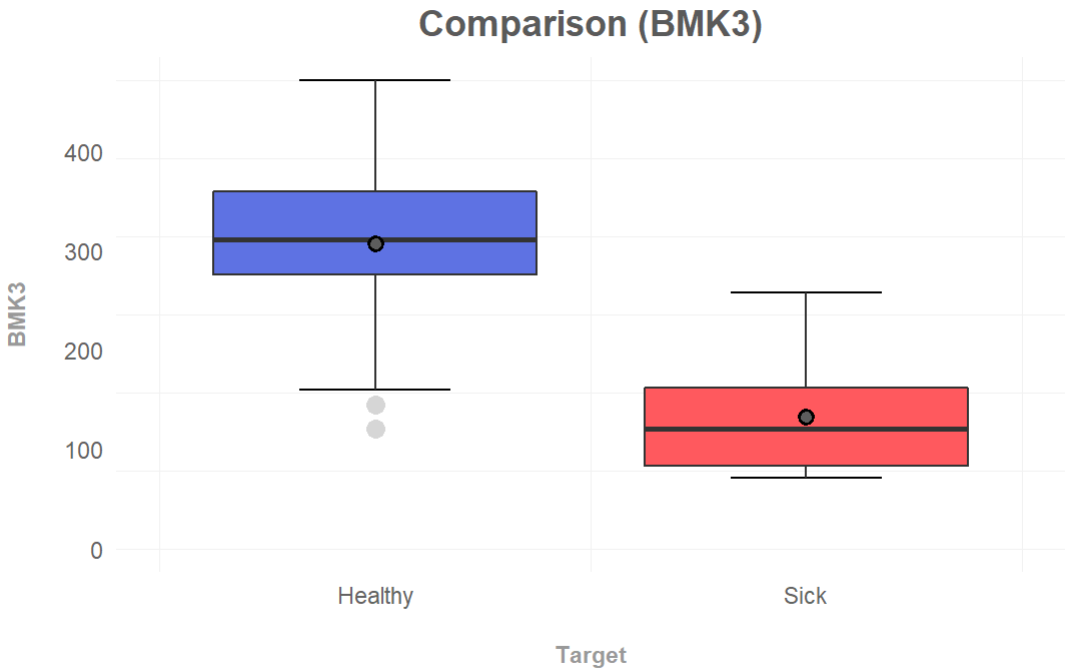
Statistical tests

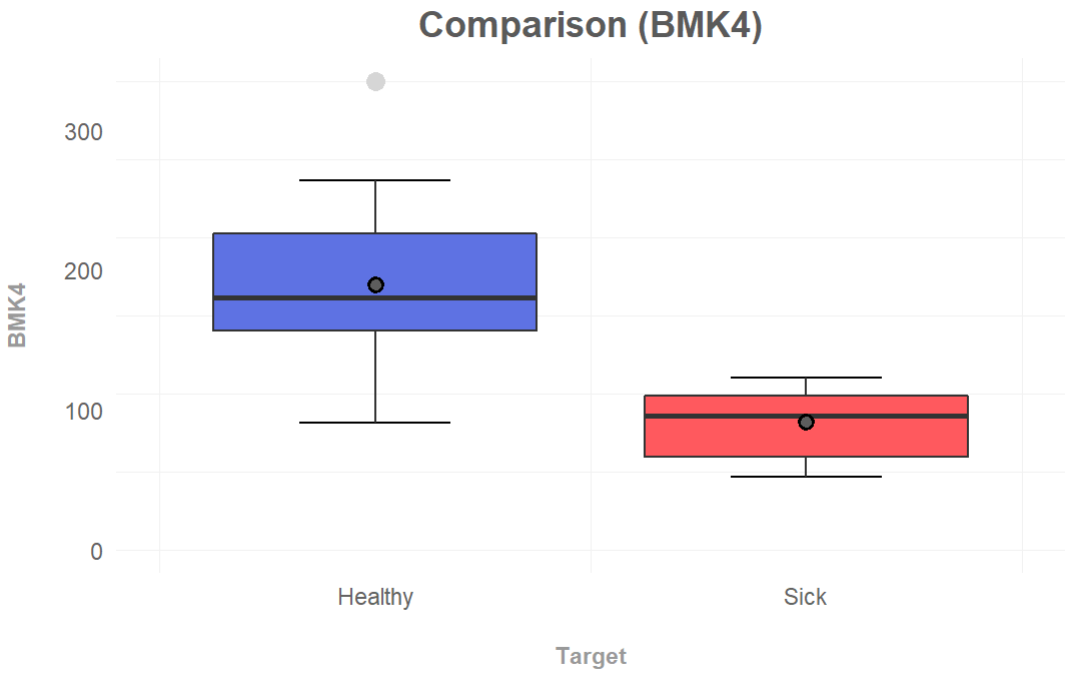
Continuous variables

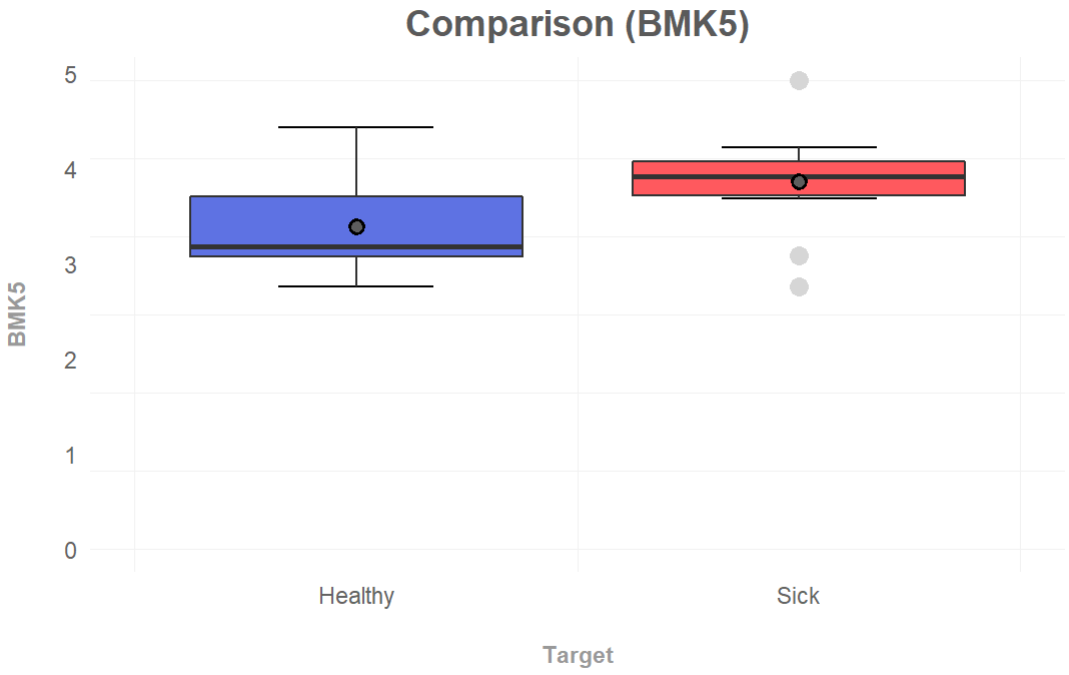
Table 5: Overview of statistical tests for continuous variables

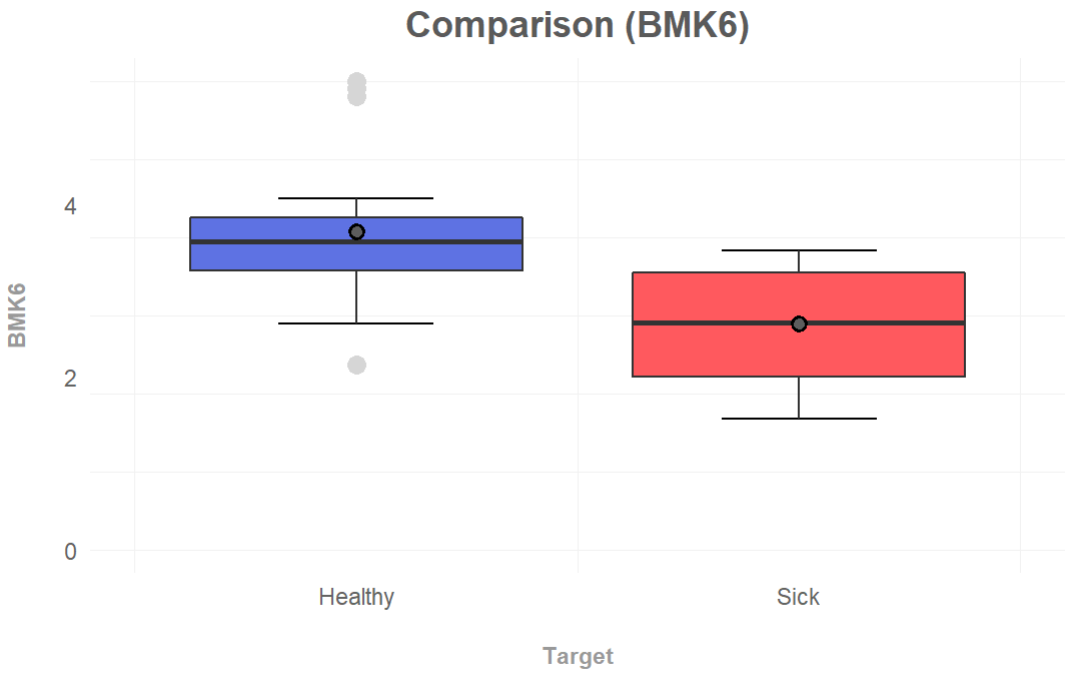
biomarker	group	Mann-Whitney's test	F-test	T-test	Welch's t-test	Paired t-test
BMK2	Target	8.35e-05	2.00e-01	3.42e-05	1.10e-04	3.42e-05
BMK3	Target	5.60e-05	2.61e-02	5.24e-06	2.48e-06	5.24e-06
BMK4	Target	2.86e-05	2.00e-03	2.94e-06	1.82e-06	2.94e-06
BMK5	Target	1.27e-02	7.88e-01	1.17e-02	1.29e-02	1.17e-02
BMK6	Target	1.08e-03	3.96e-01	9.80e-04	7.28e-04	9.80e-04
BMK7	Target	1.05e-05	4.00e-01	1.03e-06	3.52e-06	1.03e-06

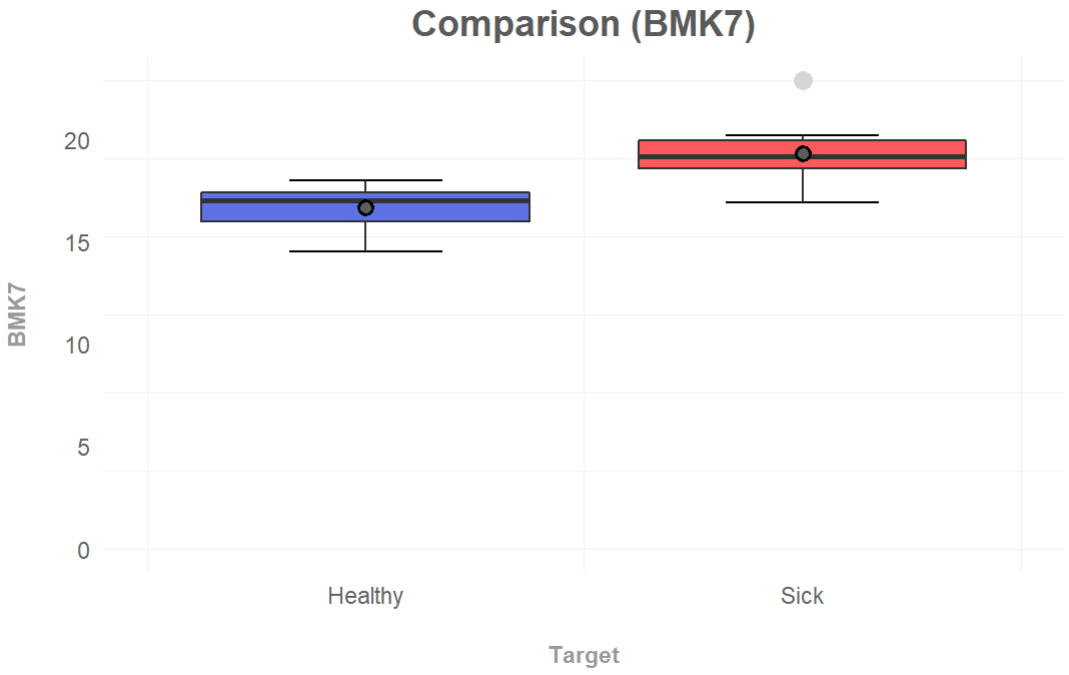

Comparison (BMK2)
BMK2
30
20
10
0
Healthy
Sick
Target


Comparison (BMK3)
BMK3
400
300
200
100
0
Healthy
Sick
Target


Comparison (BMK4)
300
200
100
0
BMK4
Healthy
Sick
Target


Comparison (BMK5)
5
4
3
2
1
0
BMK5
Healthy
Sick
Target


Comparison (BMK6)
BMK6
4
2
0
Healthy
Sick
Target


Comparison (BMK7)
BMK7
20
15
10
5
0
Healthy
Sick
Target

Categorical variables

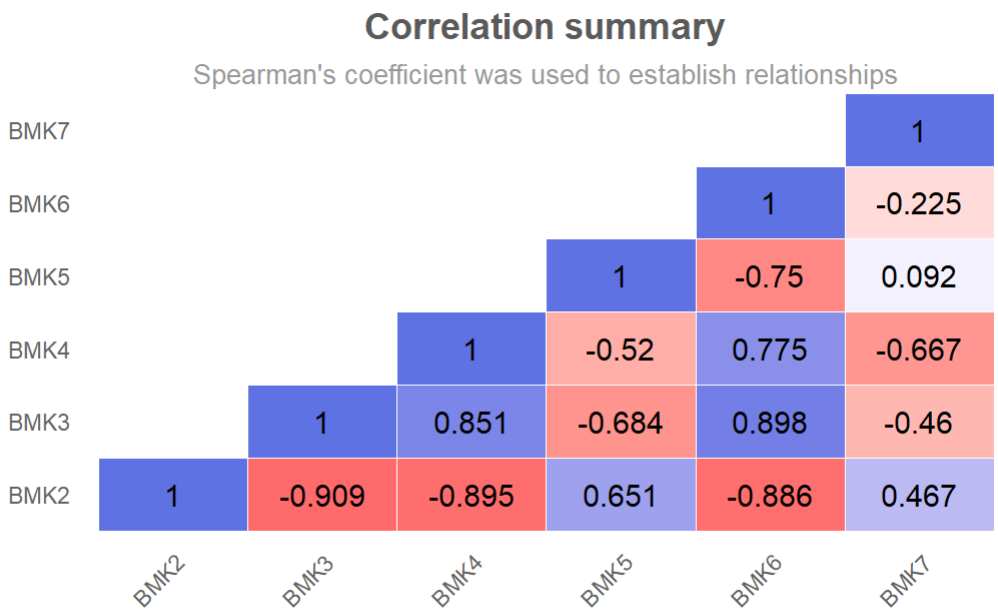
```
## [1] "No data available."
```

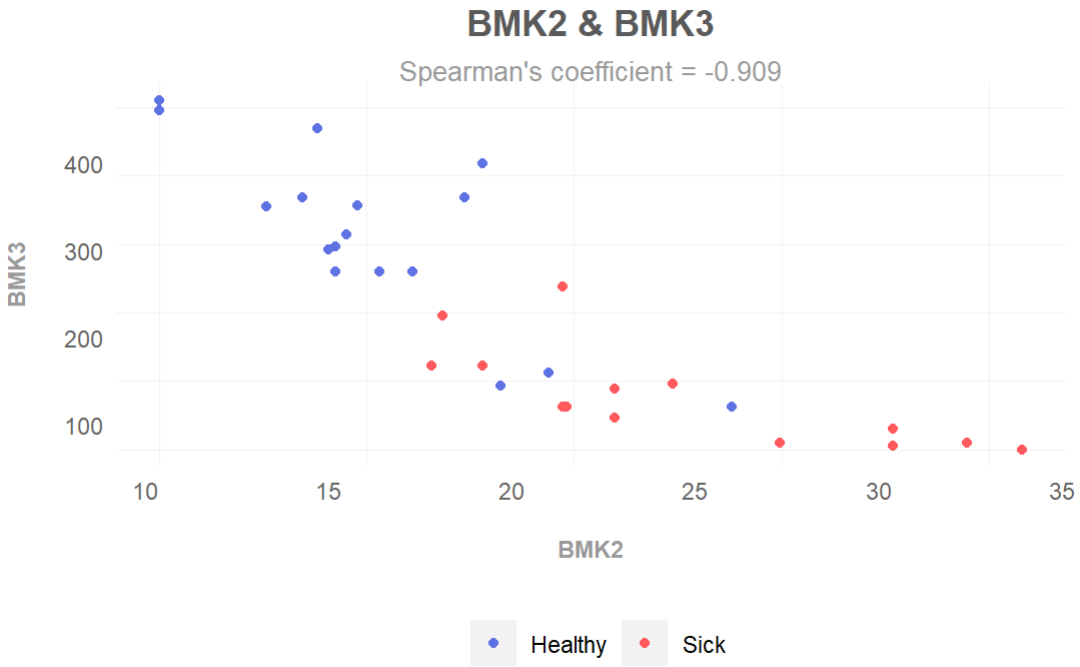
Correlations

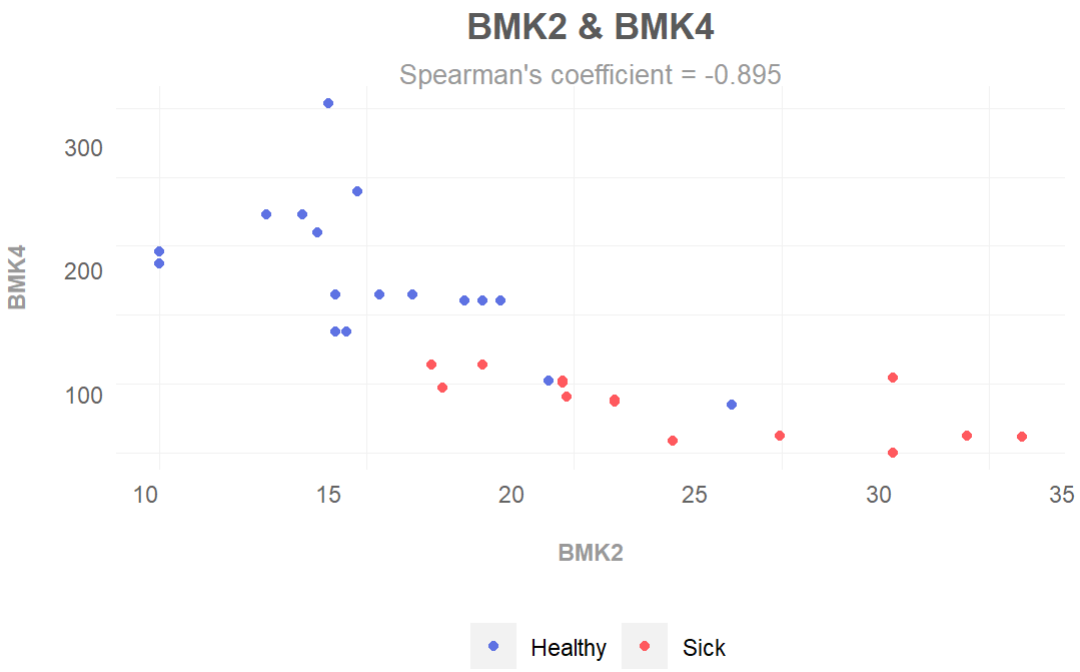
Continuous relationships

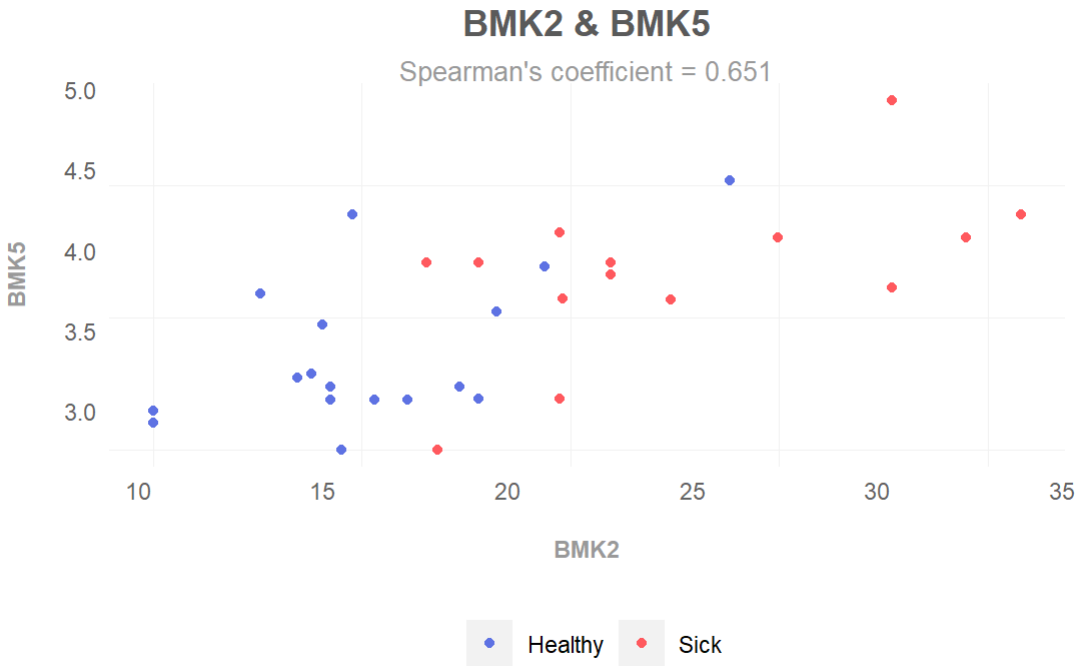
Table 6: Summary of continuous correlations

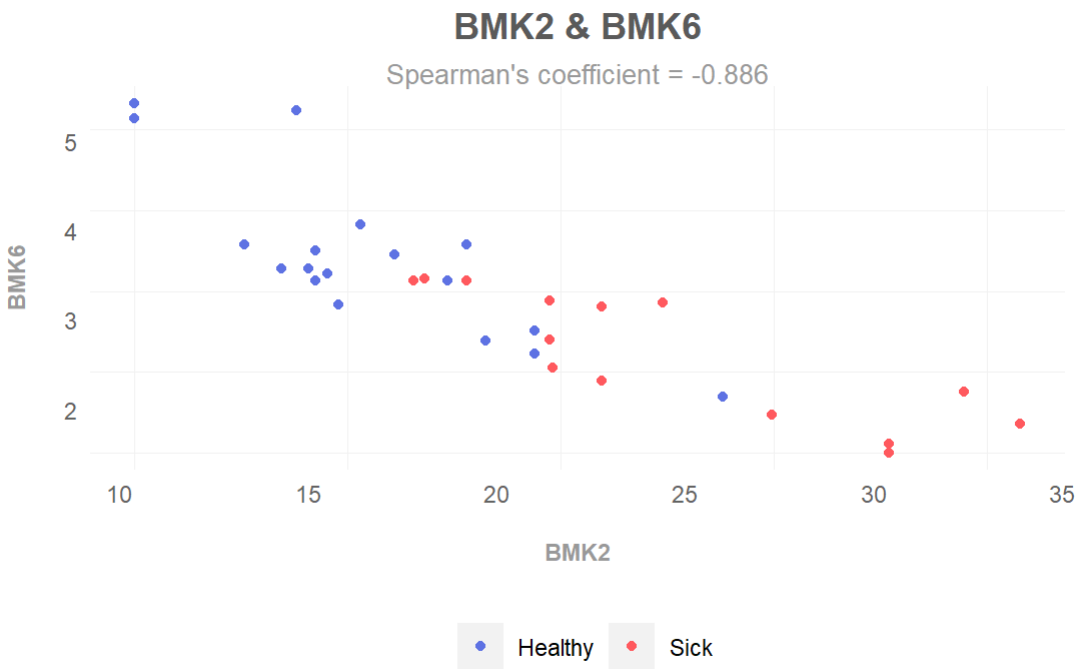
	BMK2	BMK3	BMK4	BMK5	BMK6	BMK7
BMK2	1.000	-0.909	-0.895	0.651	-0.886	0.467
BMK3	-0.909	1.000	0.851	-0.684	0.898	-0.460
BMK4	-0.895	0.851	1.000	-0.520	0.775	-0.667
BMK5	0.651	-0.684	-0.520	1.000	-0.750	0.092
BMK6	-0.886	0.898	0.775	-0.750	1.000	-0.225
BMK7	0.467	-0.460	-0.667	0.092	-0.225	1.000

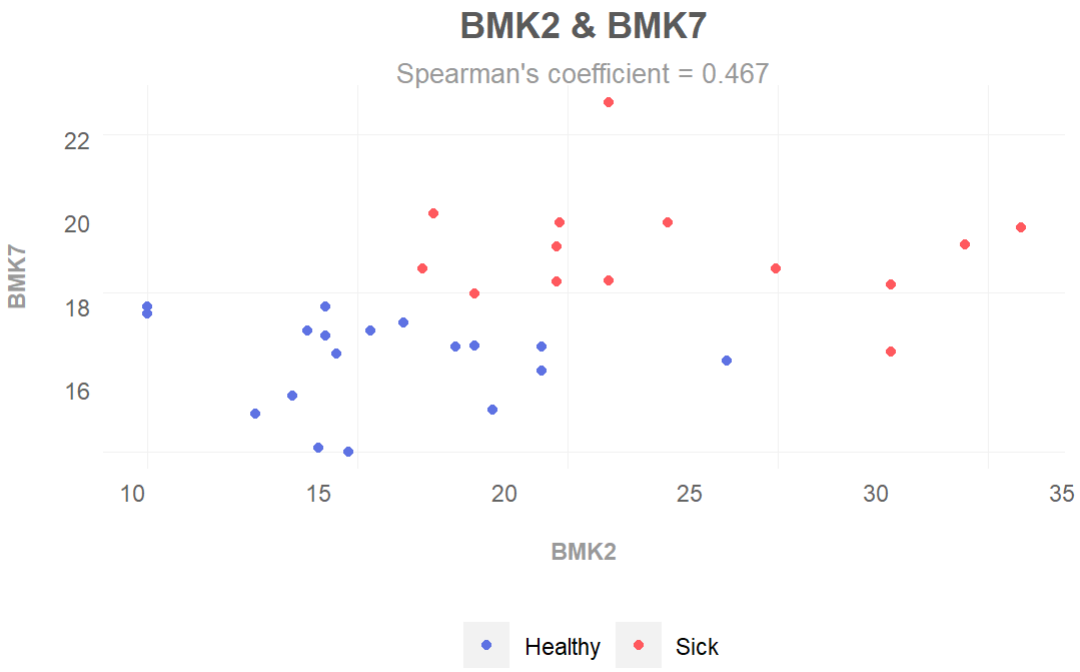

Correlation summary
Spearman's coefficient was used to establish relationships
BMK7
BMK6
BMK5
BMK4
BMK3
BMK2
1
1 -0.225
1 -0.75 0.092
1 -0.52 0.775 -0.667
1 0.851 -0.684 0.898 -0.46
1 -0.909 -0.895 0.651 -0.886 0.467
BMK2
BMK3
BMK4
BMK5
BMK6
BMK7

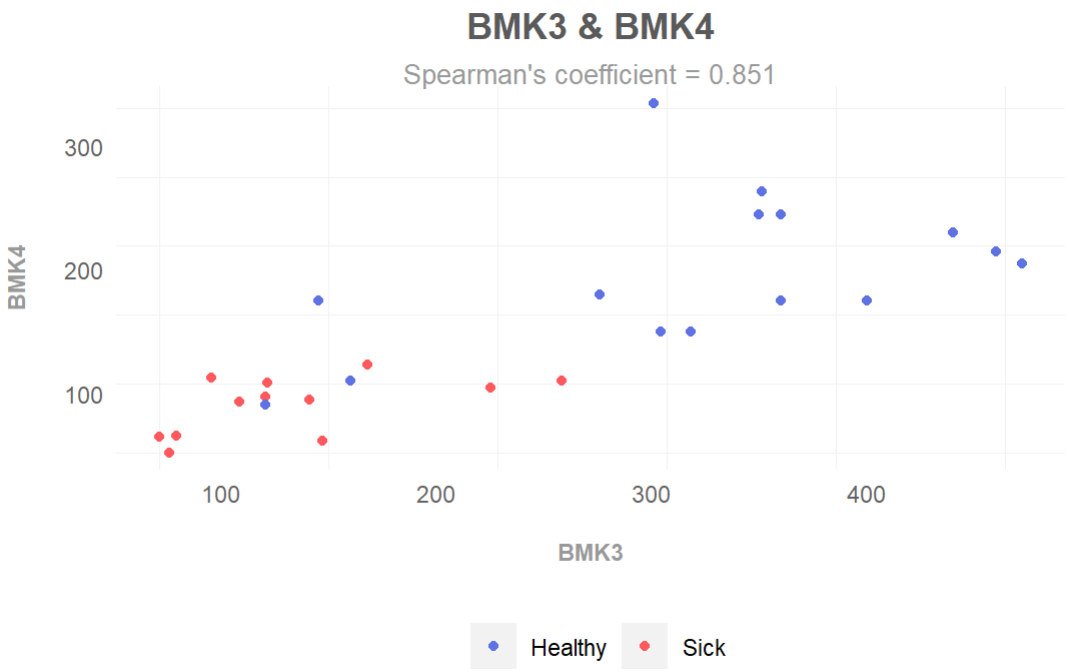

BMK2 & BMK3
Spearman's coefficient = -0.909
BMK3
400
300
200
100
10
15
20
25
30
35
BMK2
Healthy
Sick

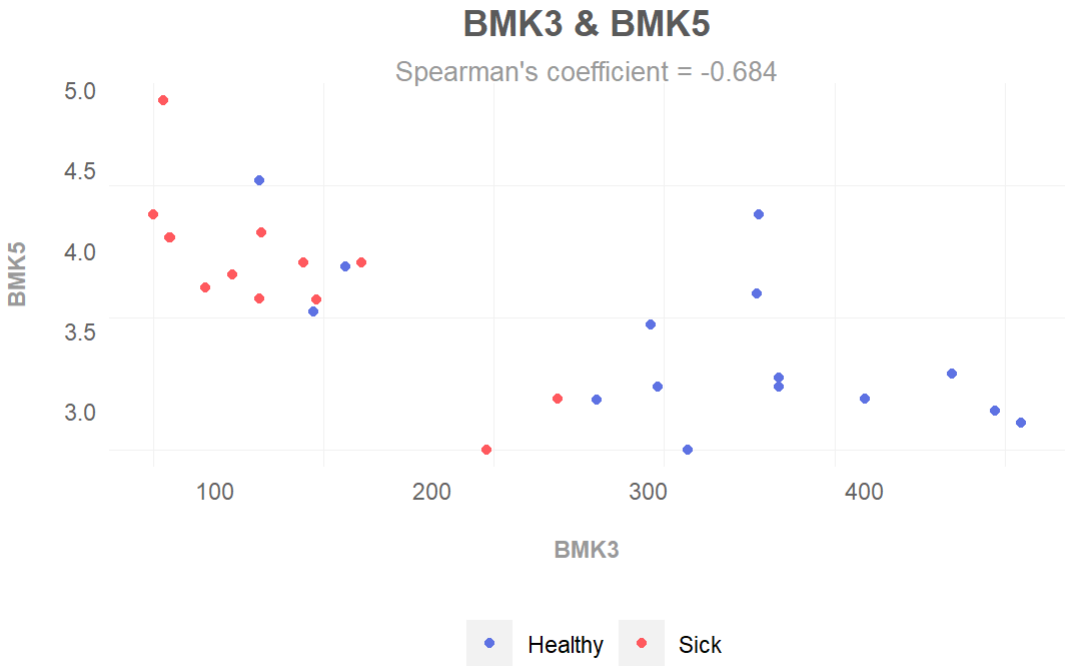

BMK2 & BMK4
Spearman's coefficient = -0.895
BMK4
300
200
100
10
15
20
25
30
35
BMK2
Healthy
Sick

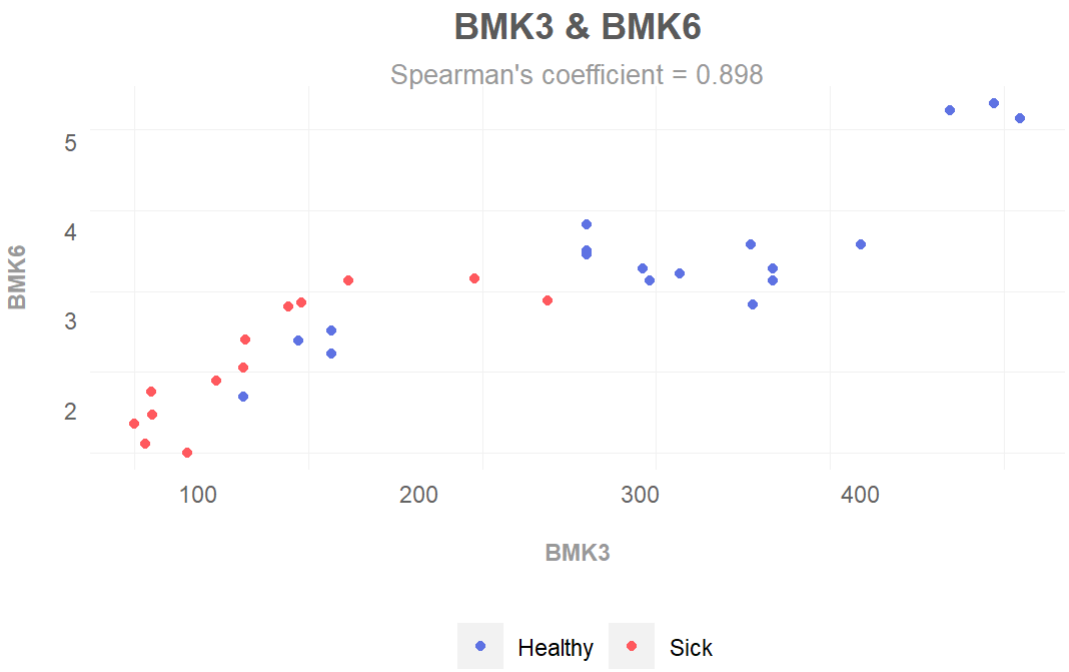

BMK2 & BMK5
Spearman's coefficient = 0.651
BMK5
5.0
4.5
4.0
3.5
3.0
10
15
20
25
30
35
BMK2
Healthy
Sick

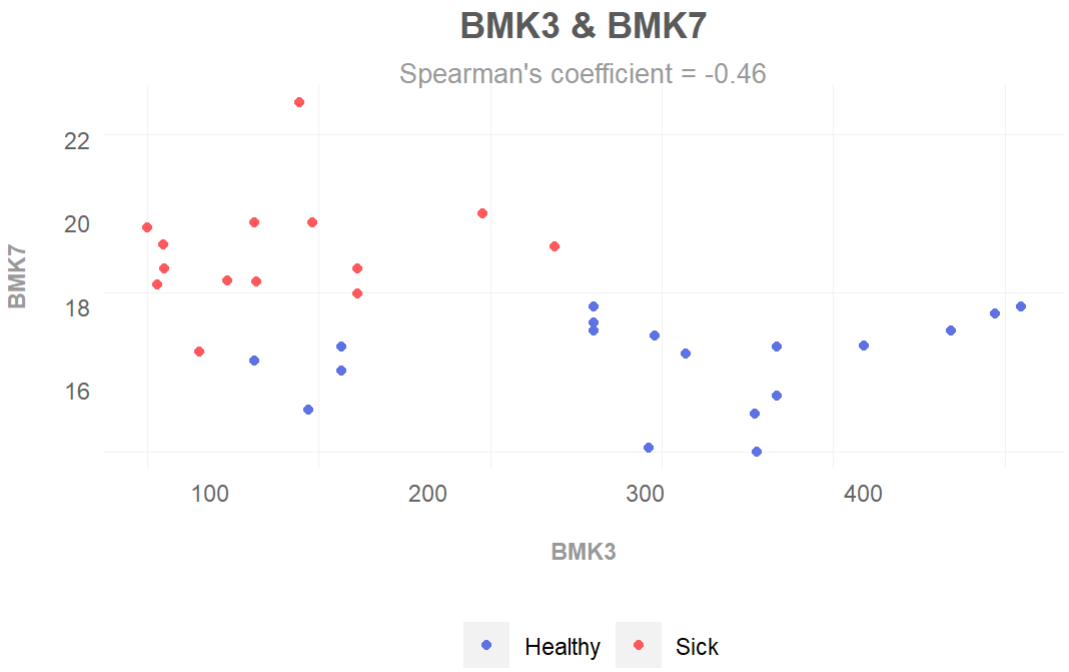

BMK2 & BMK6
Spearman's coefficient = -0.886
BMK6
5
4
3
2
10
15
20
25
30
35
BMK2
Healthy
Sick

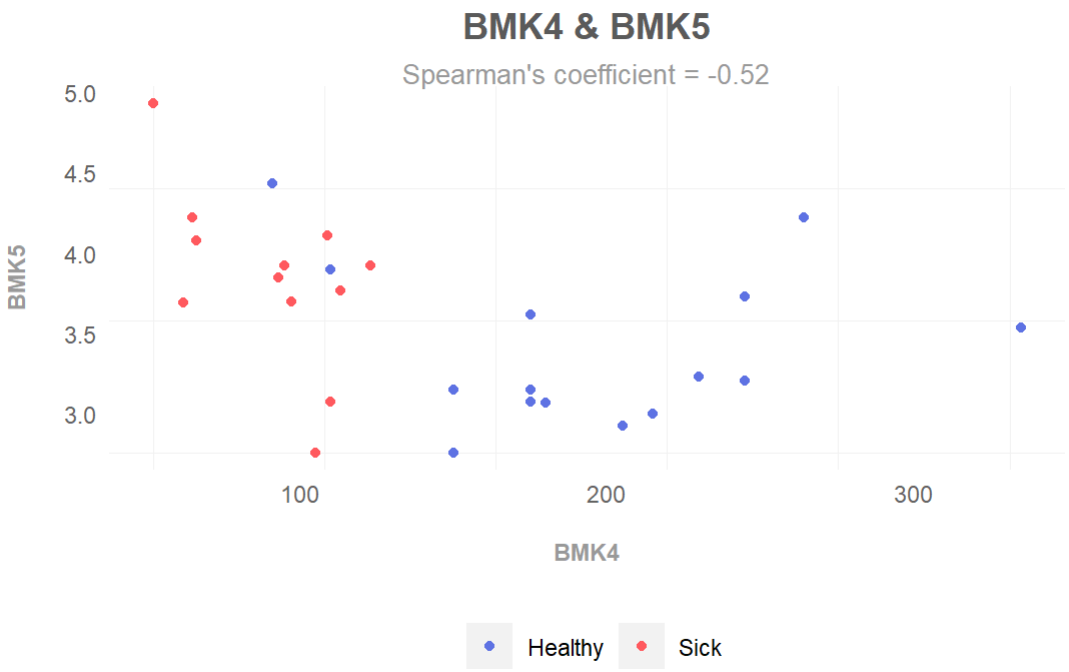

BMK2 & BMK7
Spearman's coefficient = 0.467
BMK7
22
20
18
16
10
15
20
25
30
35
BMK2
Healthy
Sick

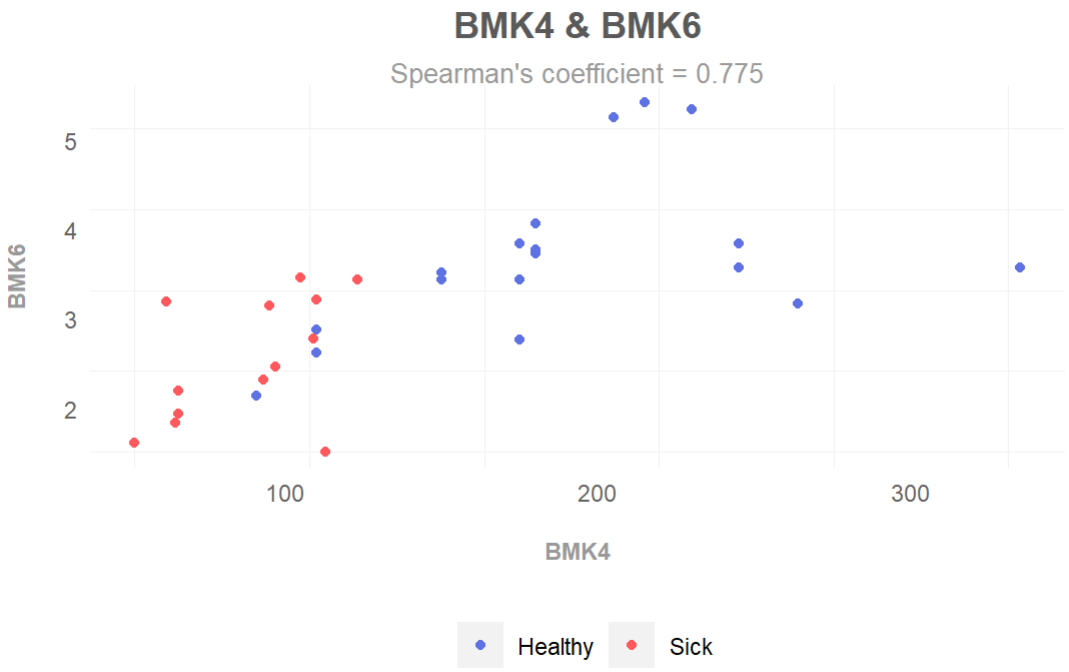

BMK3 & BMK4
Spearman's coefficient = 0.851
BMK4
300
200
100
100
200
300
400
BMK3
Healthy
Sick

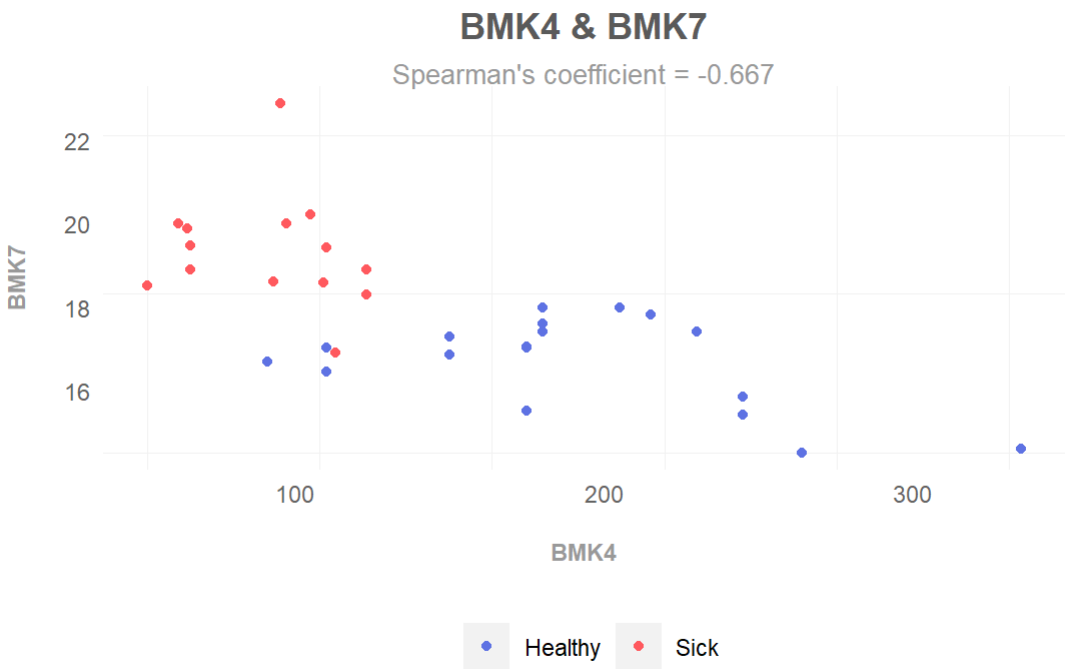

BMK3 & BMK5
Spearman's coefficient = -0.684
BMK5
5.0
4.5
4.0
3.5
3.0
100
200
300
400
BMK3
Healthy
Sick

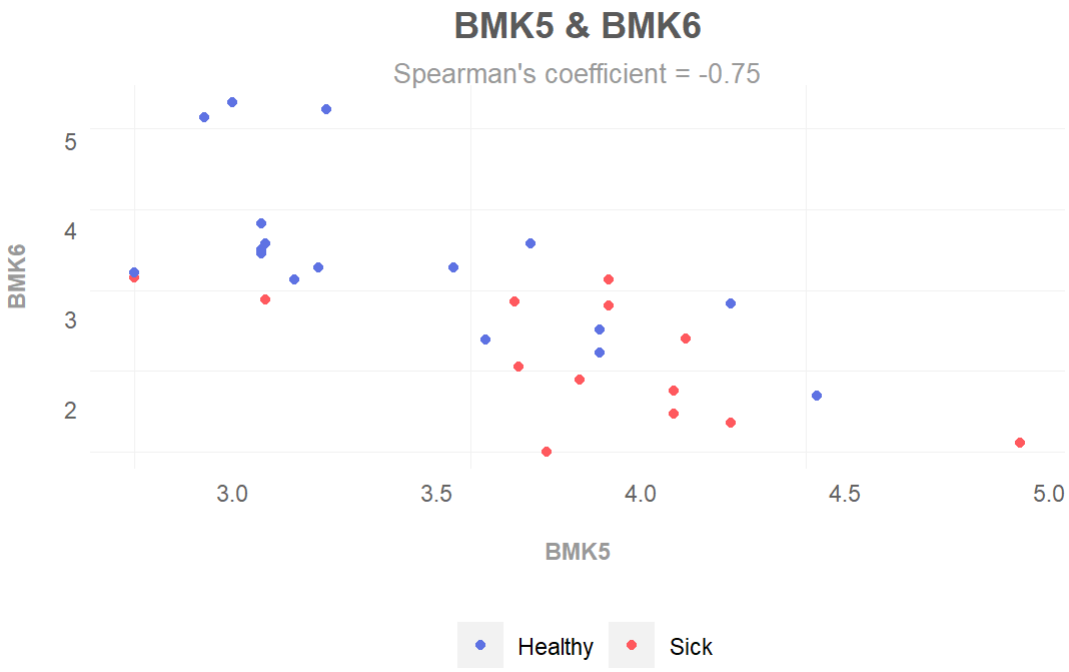

BMK3 & BMK6
Spearman's coefficient = 0.898
BMK6
5
4
3
2
100
200
300
400
BMK3
Healthy
Sick

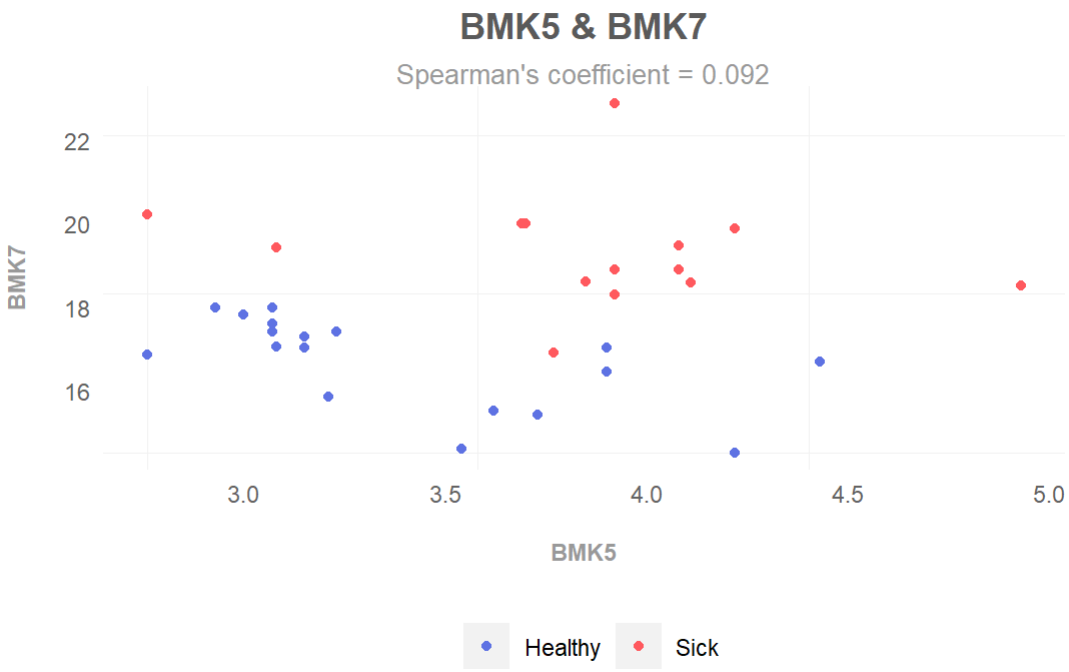

BMK3 & BMK7
Spearman's coefficient = -0.46
BMK7
22
20
18
16
100
200
300
400
BMK3
Healthy
Sick


BMK4 & BMK5
Spearman's coefficient = -0.52
5.0
4.5
4.0
3.5
3.0
BMK5
100
200
300
BMK4
Healthy
Sick


BMK4 & BMK6
Spearman's coefficient = 0.775
5
4
3
2
BMK6
100
200
300
BMK4
Healthy
Sick


BMK4 & BMK7
Spearman's coefficient = -0.667
BMK7
22
20
18
16
100
200
300
BMK4
Healthy
Sick

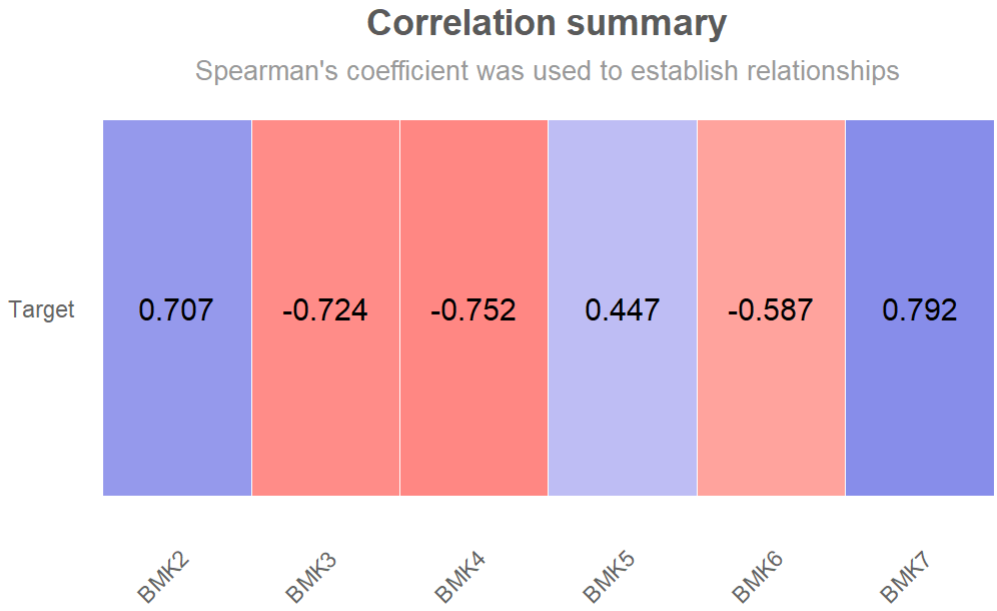

BMK5 & BMK6
Spearman's coefficient = -0.75
BMK6
5
4
3
2
3.0
3.5
4.0
4.5
5.0
BMK5
Healthy
Sick

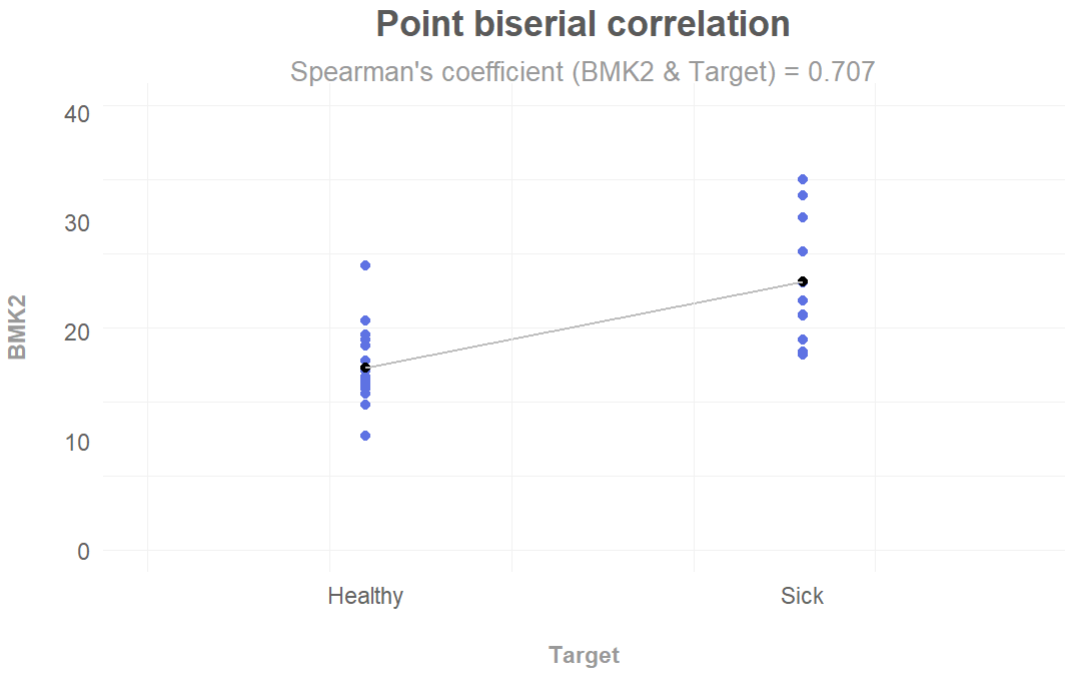

BMK5 & BMK7
Spearman's coefficient = 0.092
BMK7
22
20
18
16
3.0
3.5
4.0
4.5
5.0
BMK5
Healthy
Sick

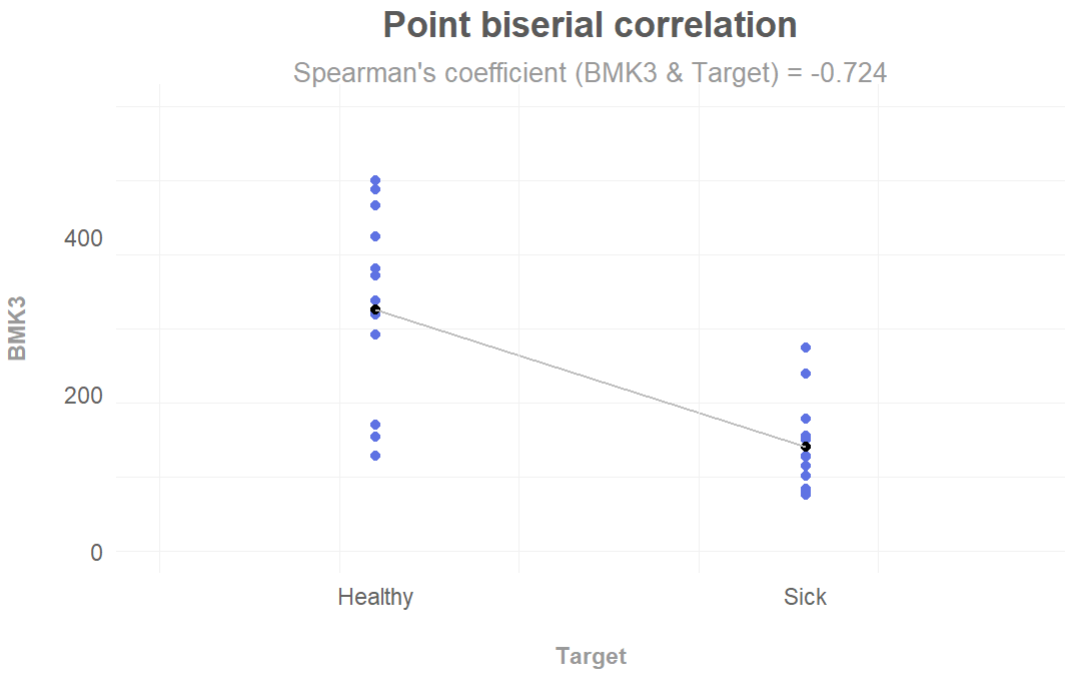
Continuous & Categorical relationships

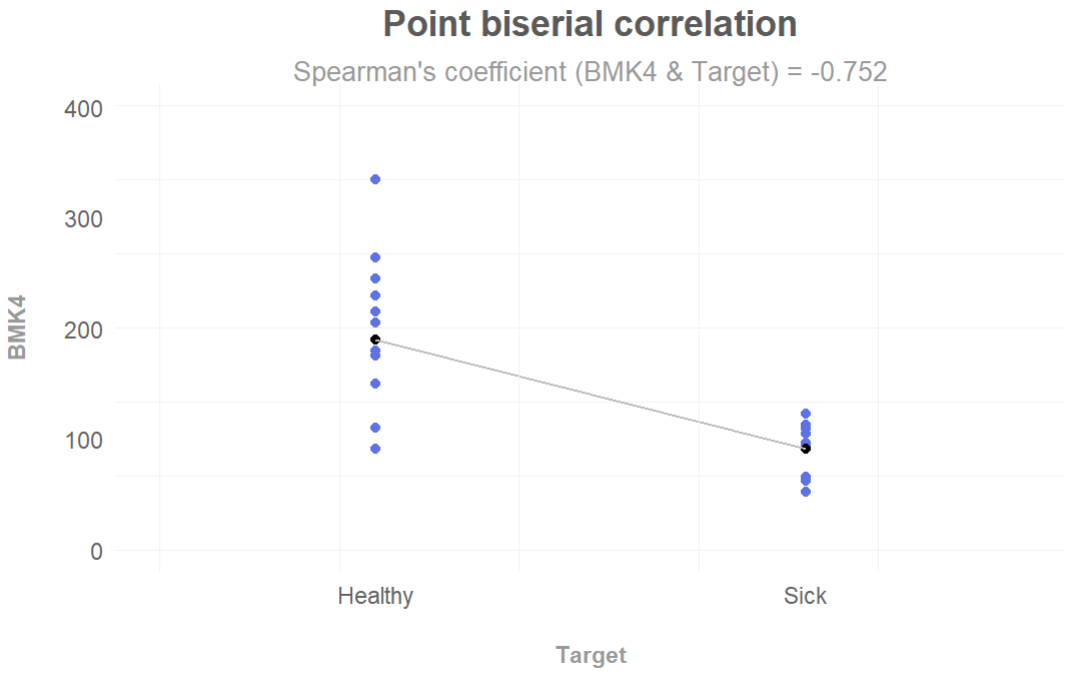
Table 7: Summary of continuous and categorical correlations

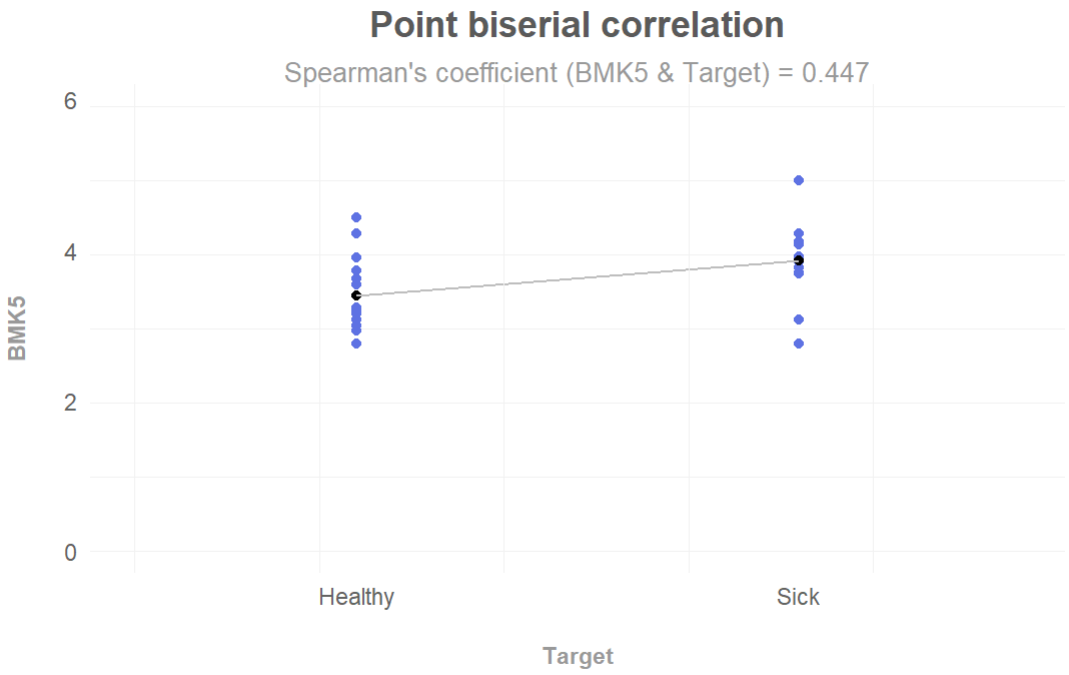
	BMK2	BMK3	BMK4	BMK5	BMK6	BMK7
Target	0.707	-0.724	-0.752	0.447	-0.587	0.792

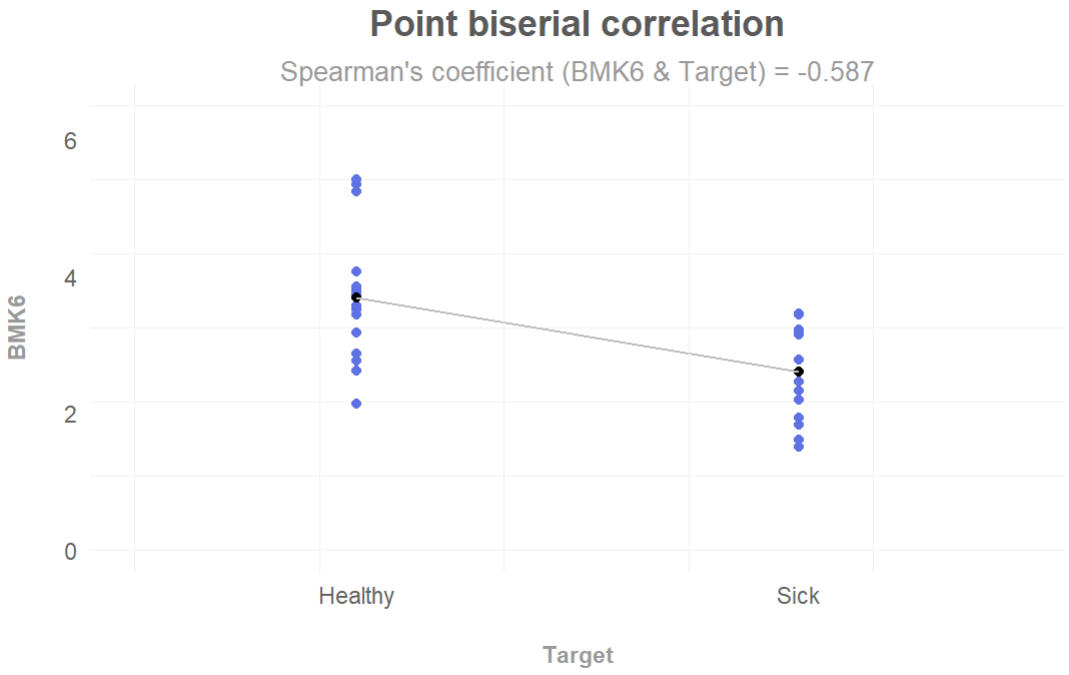

Correlation summary
Spearman's coefficient was used to establish relationships
Target
0.707
-0.724
-0.752
0.447
-0.587
0.792
BMK2
BMK3
BMK4
BMK5
BMK6
BMK7


Point biserial correlation
Spearman's coefficient (BMK2 & Target) = 0.707
BMK2
40
30
20
10
0
Healthy
Sick
Target


Point biserial correlation
Spearman's coefficient (BMK3 & Target) = -0.724
BMK3
400
200
0
Healthy
Sick
Target


Point biserial correlation
Spearman's coefficient (BMK4 & Target) = -0.752
BMK4
400
300
200
100
0
Healthy
Sick
Target


Point biserial correlation
Spearman's coefficient (BMK5 & Target) = 0.447
BMK5
6
4
2
0
Healthy
Sick
Target


Point biserial correlation
Spearman's coefficient (BMK6 & Target) = -0.587
BMK6
6
4
2
0
Healthy
Sick
Target

Categorical relationships

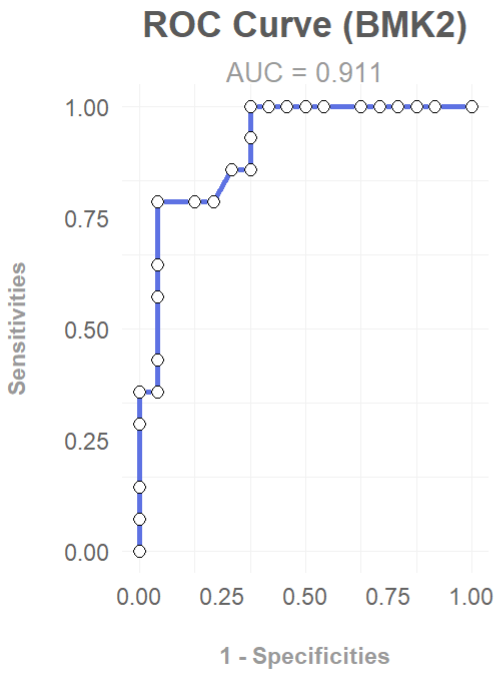
```
## [1] "No data available."
```

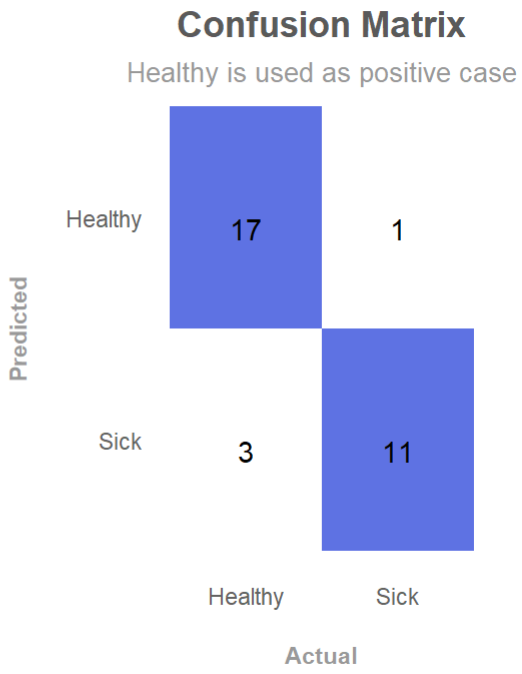
Individual performances

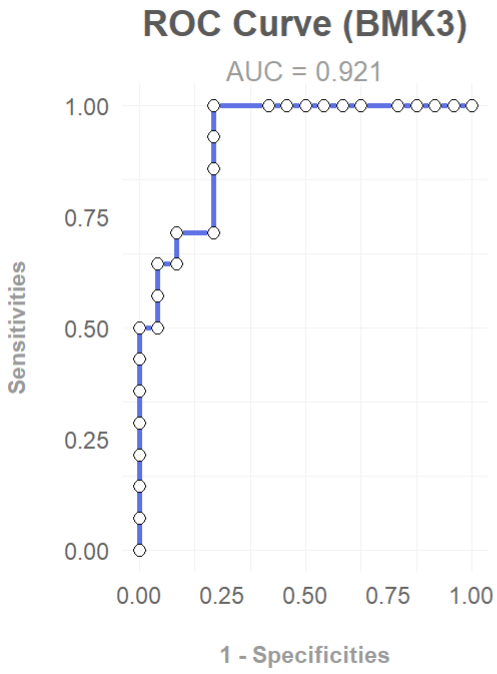
Classification

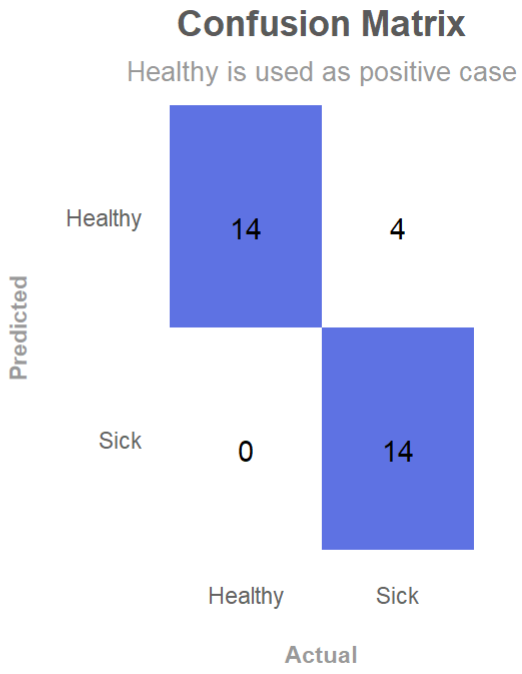
Table 8: Summary of individual performances

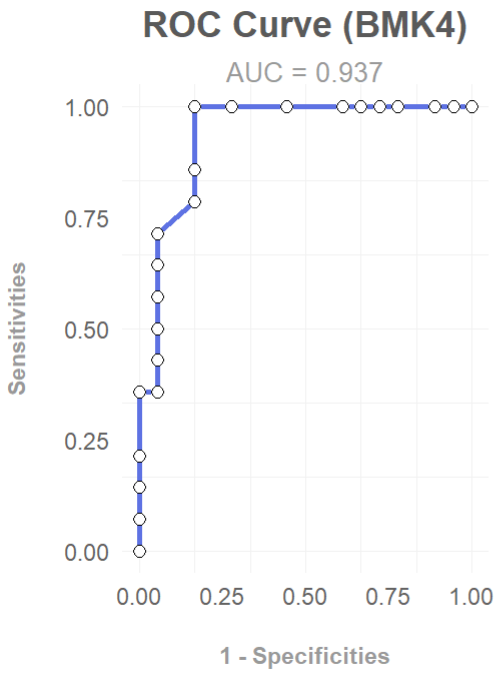
biomarker	class1	class2	logLoss	best.method	threshold	auc	sens	spe	PPV	NPV	accuracy	no.info.rate	balanced.accuracy	precision	f1	TP	FP	TN	FN
BMK2	Healthy	Sick	0.3989584	youden	21.200	0.911	0.850	0.917	0.944	0.786	0.875	0.625	0.883	0.944	0.895	17	1	11	3
BMK3	Healthy	Sick	0.3546224	youden	266.900	0.921	1.000	0.778	0.778	1.000	0.875	0.562	0.889	0.778	0.875	14	4	14	0
BMK4	Healthy	Sick	0.2630872	youden	136.500	0.937	1.000	0.824	0.833	1.000	0.906	0.531	0.912	0.833	0.909	15	3	14	0
BMK5	Healthy	Sick	0.5806063	youden	3.655	0.760	0.867	0.706	0.722	0.857	0.781	0.531	0.786	0.722	0.788	13	5	12	2
BMK6	Healthy	Sick	0.4901052	youden	3.490	0.841	1.000	0.667	0.611	1.000	0.781	0.656	0.833	0.611	0.759	11	7	14	0
BMK7	Healthy	Sick	0.2199398	youden	18.150	0.960	0.947	1.000	1.000	0.929	0.969	0.594	0.974	1.000	0.973	18	0	13	1

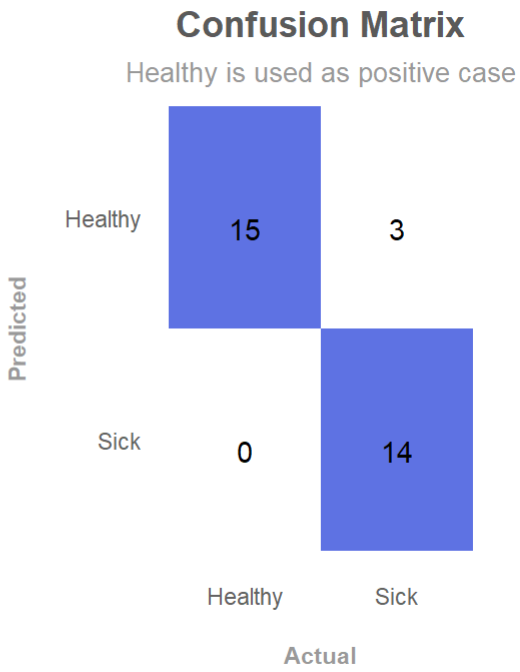

ROC Curve (BMK2)
AUC = 0.911
Sensitivities
1 - Specificities
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00

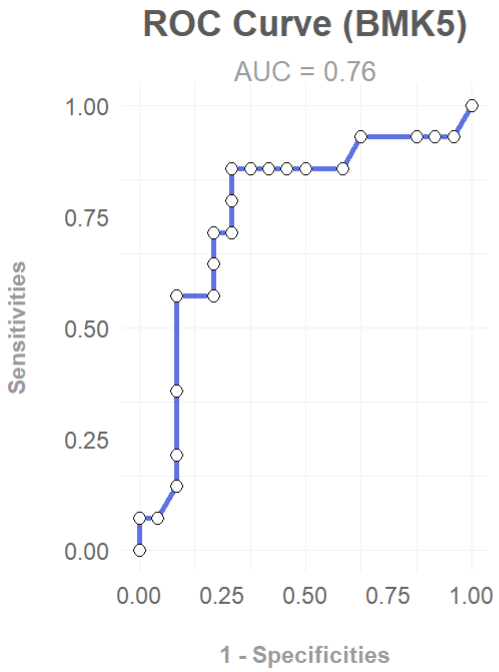

Confusion Matrix
Healthy is used as positive case
Predicted
Actual
Healthy
Sick
17
1
3
11
Healthy
Sick

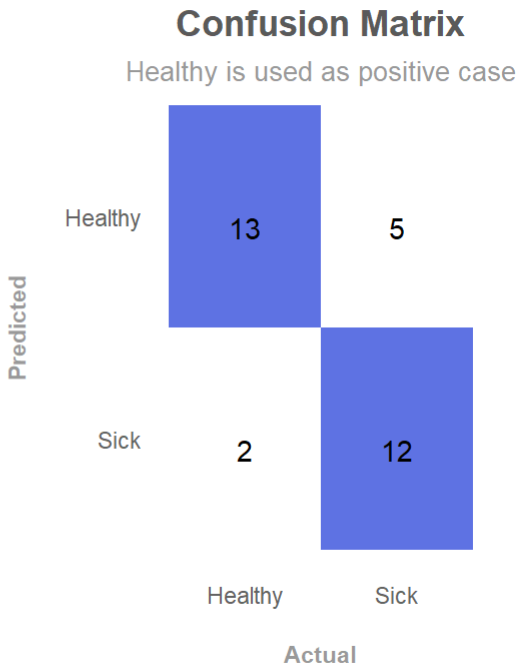

ROC Curve (BMK3)
AUC = 0.921
Sensitivities
1 - Specificities
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00

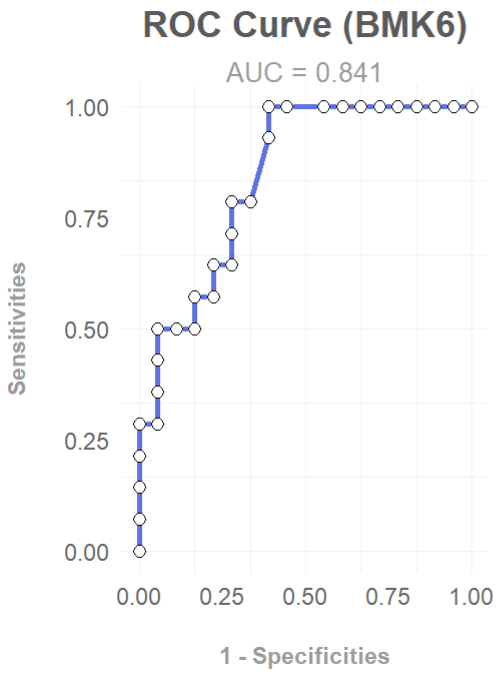

Confusion Matrix
Healthy is used as positive case
Predicted
Actual
Healthy
Sick
14
4
0
14
Healthy
Sick

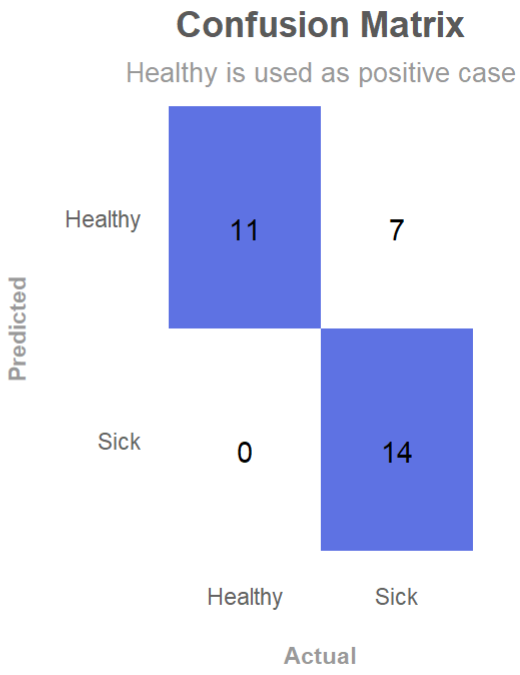

ROC Curve (BMK4)
AUC = 0.937
Sensitivities
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificities

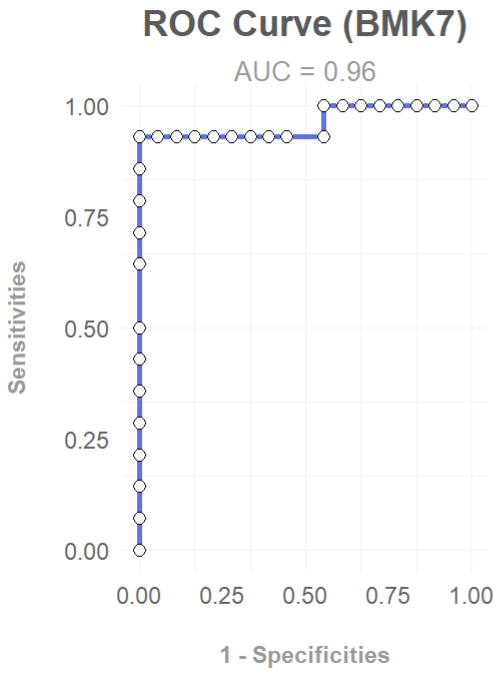

Confusion Matrix
Healthy is used as positive case
Predicted
Healthy
Sick
15
3
0
14
Healthy
Sick
Actual

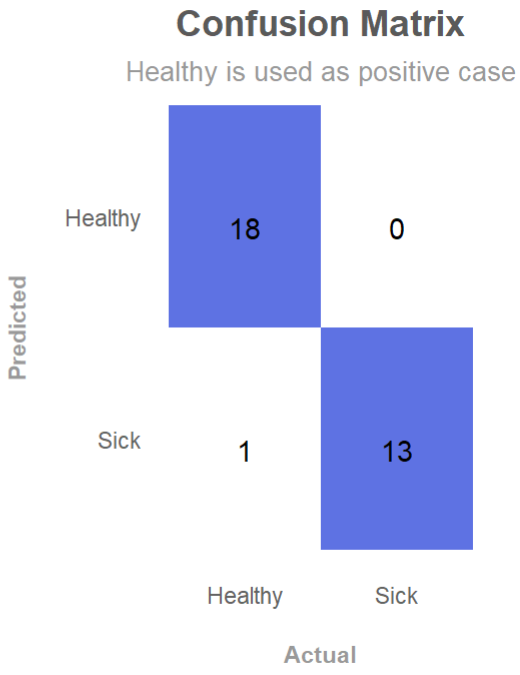

ROC Curve (BMK5)
AUC = 0.76
Sensitivities
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificities


Confusion Matrix
Healthy is used as positive case
Predicted
Healthy
Sick
13
5
2
12
Healthy
Sick
Actual


ROC Curve (BMK6)
AUC = 0.841
Sensitivities
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificities


Confusion Matrix
Healthy is used as positive case
Predicted
Healthy
Sick
11
7
0
14
Healthy
Sick
Actual


ROC Curve (BMK7)
AUC = 0.96
Sensitivities
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificities


Confusion Matrix
Healthy is used as positive case
Predicted
Healthy
Sick
18
0
1
13
Healthy
Sick
Actual

Predictive models

```
## [1] "No yet available."
```

GENESENG