MAM02 | Assignment 2

Parul Nagar & Pamela Sneekes

Table of contents

Descriptive statistics

Table 1. Descriptive statistics of the data.

Variable	No CVD (n = 1618)	CVD (n = 782)
Age (years): mean (SD)	46.4 (12.7)	48.4 (10.8)
Sex: n (%)		
Female	925 (57.2)	295 (37.7)
Male	693 (42.8)	487 (62.3)
Height: mean (SD)	172.5 (9.3)	172.3 (9.2)
Missing: n (%)	184 (11.4)	136 (17.4)
Weight: median (IQR)	73 (64-82)	75 (68-84)
Missing: n (%)	101 (6.2)	104 (13.3)
BMI: median (IQR)	24.5 (22.3-26.9)	25.5 (23.6-27.6)
Missing: n (%)	195 (12.1)	147 (18.8)
Alcohol use: n (%)		
Yes	1006 (62.2)	407 (52)
No	329 (20.3)	167 (21.4)
Unknown	0 (0)	0 (0)
Missing: n (%)	283 (17.5)	208 (26.6)
Smoking: n (%)		
Never	444 (27.4)	124 (15.9)
Ever	1003 (62)	595 (76.1)
Missing: n (%)	171 (10.6)	63 (8.1)
Systolic blood pressure: mean (SD)	133.4 (18.1)	138.2 (21.6)
Missing: n (%)	31 (1.9)	6 (0.8)
Diastolic blood pressure: mean (SD)	81.2 (10.2)	83.4 (11.0)
Missing: n (%)	31 (1.9)	6 (0.8)
Hypertension: n (%)		
No	1507 (93.1)	639 (81.7)
Yes	96 (5.9)	134 (17.1)
Missing: n (%)	15 (0.9)	9 (1.2)
Glucose: median (IQR)	4.9 (4.5-5.3)	5.1 (4.7-5.7)
Missing: n (%)	94 (5.8)	24 (3.1)
Hb1Ac (%): median (IQR)	5.5 (5-6)	5.8 (5.3-6.5)
Missing: n (%)	628 (38.8)	302 (38.6)
Diabetes: n (%)		
Yes	1567 (96.8)	695 (88.9)
No	51 (3.2)	87 (11.1)
HC in a 1e-degree family-member: n (%)		
Yes	378 (23.4)	125 (16.0)
No	1240 (76.6)	657 (84.0)
Serum total cholesterol (mmol/L): Mean (SD)	9.5 (1.9)	9.7 (2.2)
Missing: n (%)	143 (8.8)	105 (13.4)
Serum HDL cholesterol (mmol/L): median (IQR)	1.2 (1-1.4)	1.1 (0.9-1.3)
Missing: n (%)	234 (14.5)	206 (26.3)
Serum triglycerides (mmol/L): median (IQR)	1.5 (1-2.1)	1.80 (1.3-2.4)
Missing: n (%)	206 (12.7)	168 (21.5)
Serum Lpa (U/L): median (IQR)	150 (60-389.5)	230 (77-626.5)
Missing: n (%)	451 (27.9)	251 (32.1)
Serum homocysteine (µmol/L): median (IQR)	10.7 (8.8-13)	12 (10-15)
Missing: n (%)	906 (56)	400 (51.2)
Serum creatinine (µmol/L): mean (SD)	79.7 (14.5)	84.5 (16.9)
Missing: n (%)	62 (3.8)	9 (1.2)

Table 1 shows that there are several variables that have missing data. Some of the variables seem comparable in both groups, such as height, which have the same mean and standard deviation. Other variables seem less comparable in both groups, such as hypertension, but this difference might also be caused due to the size of the groups. The group with no CVD is twice as large as the group with CVD.

Univariable and multivariable regression analyses

Table 2. Univariable and multivariable regression analyses between the risk factors and cardiovascular disease (CVD).

Variable	Univariable OR (95% CI)	Multivariable OR (95% CI)
Age	1.01 (1.01 - 1.02)	1.00 (0.99 - 1.01)
Sex	2.20 (1.85 - 2.63)	3.15 (2.28 - 4.36)
Height	0.99 (0.99 - 1.01)	0.99 (0.92 - 1.06)
Weight	1.01 (1.01 - 1.02)	0.98 (0.90 - 1.06)
BMI	1.07 (1.05 - 1.10)	1.08 (0.85 - 1.37)
Alcohol use	1.25 (1.01 - 1.56)	1.52 (1.17 - 1.98)
Smoking	2.12 (1.70 - 2.66)	1.77 (1.37 - 2.28)
Systolic blood pressure	1.01 (1.01 - 1.02)	1.01 (0.99 - 1.01)
Diastolic blood pressure	1.02 (1.01 - 1.03)	0.99 (0.99 - 1.01)
Hypertension	3.29 (2.50 - 4.36)	2.27 (1.63 - 3.17)
Glucose	1.48 (1.34 - 1.64)	1.15 (1.01 - 1.30)
Hb1Ac	1.34 (1.22 - 1.47)	1.17 (1.05 - 1.30)
Diabetes	3.85 (2.70 - 5.53)	1.79 (1.09 - 2.93)
HC in a 1e-degree family-member	1.60 (1.28 - 2.01)	1.37 (1.07 - 1.76)
Serum total cholesterol	1.06 (1.01 - 1.11)	1.05 (0.99 - 1.11)
Serum HDL cholesterol	0.43 (0.32 - 0.59)	0.68 (0.47 - 0.99)
Serum triglycerides	1.39 (1.27 - 1.53)	1.08 (0.97 - 1.21)
Serum Lpa	1.00 (1.00 - 1.00)	1.00 (1.00 - 1.00)
Serum homocysteine	1.05 (1.02 - 1.07)	1.02 (1.00 - 1.04)
Serum creatinine	1.02 (1.01 - 1.03)	1.00 (0.99 - 1.01)

The variable height is not significant in both the uni- and multivariable regression analyses. Some of the variables significant in the univariable regression analyses are not significant in the multivariable regression analyses, such as age, weight and diastolic blood pressure. This suggests that certain associations with the outcome variable are controlled for the effects by the other variables in the model. Some variables have a high odds ratio in both the univariable and multivariable regression analyses, such as sex and hypertension.

Backward and forward selection

Table 3. Backward and forward models with selected variables and AIC for each imputed dataset.

#	Backward		Forward	
	AIC	**Selected variables in final model**	**AIC**	**Selected variables in final model**
1	2613.7	sex, height, alcoholuse, smoking, systbp, hypertension, Glucose, Hba1c, diabetes, familiarHC, Tc, HDL, Tg, Lpa, homocysteine, creatinine	2613.7	sex, hypertension, Lpa, Hba1c, Tg, smoking, height, Glucose, HDL, alcoholuse, homocysteine, Tc, familiarHC, systbp, diabetes, creatinine
2	2645	sex, weight, bmi, alcoholuse, smoking, systbp, hypertension, Glucose, Hba1c, diabetes, familiarHC, Tc, HDL, Tg, Lpa, homocysteine	2644.6	sex, hypertension, Lpa, diabetes, height, smoking, alcoholuse, Hba1c, homocysteine, Tg, familiarHC, systbp, Tc, Glucose, HDL
3	2643.7	sex, weight, bmi, alcoholuse, smoking, systbp, hypertension, Glucose, Hba1c, diabetes, familiarHC, Tc, HDL, Lpa, homocysteine	2643	sex, hypertension, Lpa, diabetes, height, homocysteine, smoking, Hba1c, alcoholuse, Tc, familiarHC, systbp, HDL, Glucose
4	2629.9	sex, weight, bmi, alcoholuse, smoking, systbp, hypertension, Glucose, Hba1c, diabetes, familiarHC, Tc, HDL, Lpa, homocysteine	2630.7	sex, hypertension, Lpa, diabetes, height, smoking, homocysteine, alcoholuse, Hba1c, Tg, systbp, HDL, Glucose, familiarHC, Tc
5	2590.3	sex, height, alcoholuse, smoking, systbp, hypertension, Glucose, Hba1c, diabetes, familiarHC, Tc, HDL, Tg, Lpa, homocysteine	2590.3	sex, hypertension, Lpa, Hba1c, height, homocysteine, Tg, smoking, alcoholuse, diabetes, systbp, HDL, familiarHC, Tc, Glucose
6	2638.4	sex, weight, bmi, alcoholuse, smoking, systbp, hypertension, Glucose, Hba1c, diabetes, familiarHC, Tc, HDL, Tg, Lpa, homocysteine	2638.5	sex, hypertension, Lpa, height, diabetes, smoking, alcoholuse, Tg, homocysteine, systbp, Hba1c, familiarHC, Glucose, HDL, Tc
7	2622.5	sex, weight, bmi, alcoholuse, smoking, systbp, hypertension, Glucose, Hba1c, diabetes, familiarHC, Tc, Tg, Lpa, homocysteine	2621.4	sex, hypertension, Lpa, Hba1c, height, Glucose, smoking, alcoholuse, homocysteine, Tc, familiarHC, diabetes, Tg, systbp
8	2592	sex, weight, bmi, alcoholuse, smoking, systbp, hypertension, Glucose, Hba1c, diabetes, familiarHC, HDL, Tg, Lpa, homocysteine	2591	Lpa, sex, Hba1c, hypertension, height, homocysteine, Tg, smoking, diabetes, HDL, familiarHC, alcoholuse, Glucose, systbp
9	2604.6	sex, height, alcoholuse, smoking, systbp, hypertension, Glucose, Hba1c, diabetes, familiarHC, Tc, HDL, Lpa, homocysteine	2604.6	sex, hypertension, height, Lpa, Hba1c, smoking, Glucose, homocysteine, alcoholuse, Tc, HDL, systbp, familiarHC, diabetes
10	2646.1	sex, height, alcoholuse, smoking, systbp, hypertension, Glucose, Hba1c, diabetes, familiarHC, Tc, HDL, Tg, Lpa, homocysteine	2646.1	sex, hypertension, Lpa, diabetes, smoking, height, alcoholuse, Hba1c, Tg, homocysteine, familiarHC, systbp, HDL, Glucose, Tc

Table 4. Frequency of selected variables and average of AIC in back- and forward selection.

Frequency of variable	**Backward**	**Forward**
Age	0	0
Sex	10	10
Height	4	10
Weight	6	0
BMI	6	0
Alcohol use	10	10
Smoking	10	10
Systolic blood pressure	10	10
Diastolic blood pressure	0	0
Hypertension	10	10
Glucose	10	10
HbA1c	10	10
Diabetes	10	10
HC in a 1e-degree family-member	10	10
Serum total cholesterol	9	9
Serum HDL cholesterol	9	9
Serum triglycerides	7	8
Serum Lpa	10	10
Serum homocysteine	10	10
Serum creatinine	1	1
Average AIC*	*2622.6*	*2622.4*

* Average AIC of 10 imputed datasets as represented in Table 3.

The final models of both the backward and forward selection seem to select the same variables resulting in a similar AIC. Only 11 variables are selected in all models. The average AIC of 10 imputed datasets for both the backward (2622.6) and forward selection (2622.4) are similar. The fifth dataset resulted in both cases the lowest AIC (2590.3) and also includes the same variables.

Bootstrapping with backward selection

Table 5. Frequency table of selected variables in 1000 bootstrap samples using backward selection.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Age	44	33	84	51	29	58	46	55	86	77
Sex	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
Height	540	409	352	640	593	246	627	227	594	377
Weight	460	598	650	385	427	795	407	800	501	639
BMI	459	604	609	369	426	803	405	793	531	620
Smoking	966	980	987	995	971	995	999	990	996	999
Alcohol use	766	843	527	853	955	878	936	981	975	727
Systolic blood pressure	505	399	544	464	443	386	405	357	372	499
Diastolic blood pressure	106	92	90	95	81	64	123	84	70	96
Hypertension	999	996	995	999	991	996	995	999	998	1000
Glucose	563	522	709	564	779	594	515	609	562	580
Hba1c	950	963	957	707	997	752	754	938	926	859
diabetes	563	524	495	663	335	664	736	564	534	663
familiar HC	601	578	690	493	512	743	564	673	398	663
Tc	549	603	260	498	578	498	527	477	547	456
HDL	847	228	844	540	424	413	583	593	532	592
Tg	366	708	469	443	210	227	133	223	152	622
Lpa	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
Homocysteine	950	916	623	995	923	221	859	857	1000	708
Creatinine	82	132	169	173	63	309	106	116	50	179

When a variable appears at least in 50% of the bootstrap samples in an imputed dataset, this is highlighted in red. Only 7 variables appeared at least in 50% of the bootstrap samples for each imputed datasets.

Multicollinearity

Table 6. Multicollinearity of the risk factors using VIF scores.

Risk factor	VIF score
Sex=male	2.67
Height	37.86
Weight	92.75
BMI	62.91
Smoking=ever	1.05
Systolic BP	1.88
Diastolic BP	1.81
Hypertension	1.23
Glucose	1.54
HbA1c	1.18
Diabetes=never	1.47
Familiar HC=no	1.03
Tc	1.07
HDL	1.26
Tg	1.29
Lpa	1.08
Homocysteine	1.05
Creatinine	1.62
Age	1.38

After this, multivariable analysis was performed by eliminating combinations of height, weight, and BMI, and the AICs were noted.

Table 7. Elimination of variables using VIF with AIC.

Variables eliminated	AIC
Height and weight	651.37
Height and BMI	669.24
Weight and BMI	646.74
Height	649.28
Weight	651.37
BMI	648.26
Height, weight, and BMI	671.24

The lowest AIC was seen by eliminating weight and BMI. However, we decided to use the other methods for variable selection since they were more robust, and this method eliminated only two variables.

Using the backward and forward selection in combination with the bootstrapping, we ended up with three prediction models. The first model consists of variables which resulted in the lowest AIC from the fifth imputed dataset using both backward and forward selection. The second prediction model consists of variables that appeared in all models from all 10 imputed datasets using both backward and forward selection. The third model selected the variables that appeared at least in 50% of the bootstrap samples in all imputed datasets. The prediction models are described in Table 8.

Table 8. Final models for prediction

Final model	Variables selected
1	sex, height, alcoholuse, smoking, systbp, hypertension, Glucose, Hba1c, diabetes, familiarHC, Tc, HDL, Tg, Lpa, homocysteine
2	sex, alcoholuse, smoking, systbpa, hypertension, glucose, Hba1c, diabetes, familiarhc, lpa, homocysteine
3	sex, alcoholuse, smoking, alcoholuse, hypertension, Glucose, Hba1c, Lpa

Calibration plots

Figure 1. Calibration using the first prediction model with 10 imputed datasets

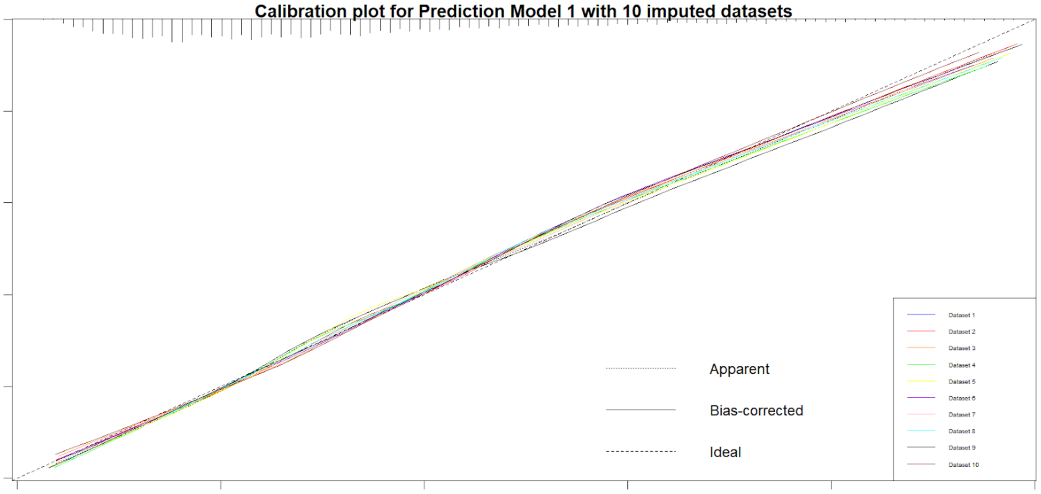

Figure 2. Calibration using the second prediction model with 10 imputed datasets

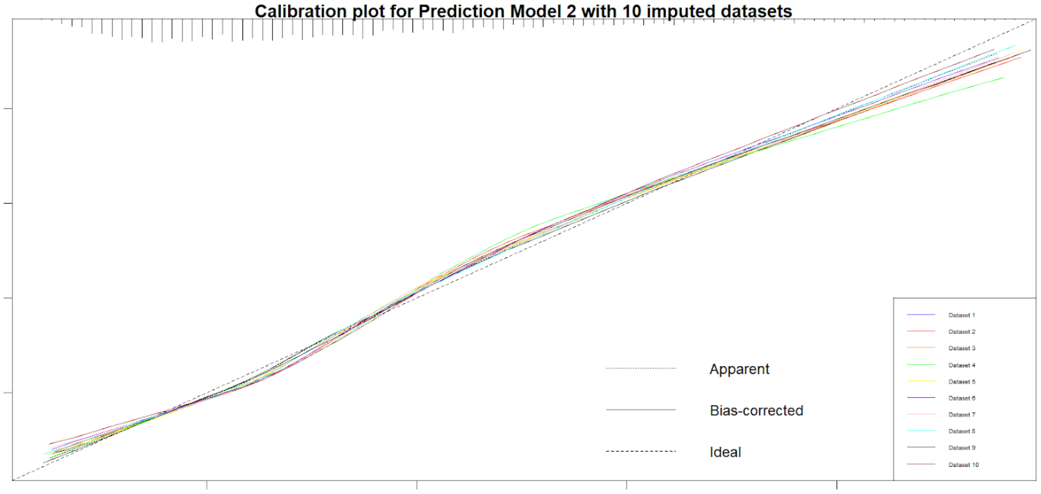

Figure 3. Calibration using the third prediction model with 10 imputed datasets.

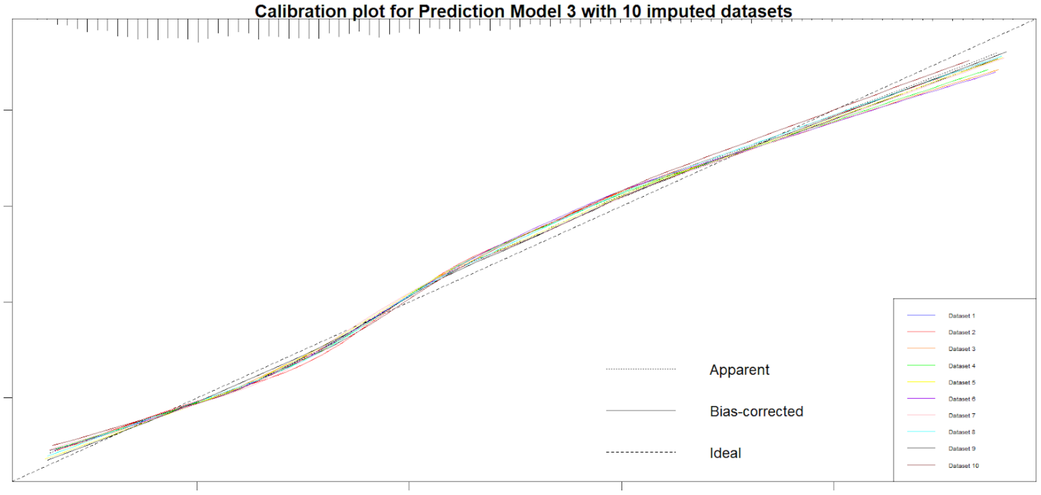

Figure 1, 2, and 3 illustrate the three calibration plots with all 10 imputed datasets. From these plots, it is difficult to determine the final model as in all of them the apparent, bias corrected, and ideal lines are in close proximity. To have a more objective comparison of the different models, performance metrics, the Brier score and C-index, were calculated. For these calculations, we chose the prediction models 2 and 3 from Table 8 as they resulted in the most number of variables eliminated. We selected models with less variables, because they are more explainable in comparison with models with more variables.

Performance metrics

Prediction Model	**Performance metric**	**Original (avg)***	**Corrected for overfitting (avg)***
2	Brier score	0.184	0.186
	C-index	0.739	0.733
3	Brier score	0.188	0.189
	C-index	0.727	0.724

*avg of the 10 imputed datasets

For the Brier scores, both model 2 and 3 have scores that are close to 0, demonstrating that the predicted probabilities are close to the actual outcomes. However, model 2 has a slightly lower score. For the C-index, while both models did not show excellent scores (closer to 1), the scores seem acceptable. However model 2 has a higher score than model 3. Therefore, in both the metrics, model 2 is the better performing one.

Chosen methods

Patient characteristics with familial hypercholesterolemia are described in Table 1. For continuous data, the mean with standard deviation (SD) is used for normal distribution and median with interquartile range (IQR) for skewed distribution. Several of the variables have missing values. The univariable regression analyses were performed with complete cases only, and the results are described in Table 2. The multivariable regression analyses were performed with imputed data. Performing multivariable regression with complete case only data reduces the sample size which in turn affects the power of this analysis. Furthermore, this can also result in a higher risk of bias, because the missing values may be related to the predictor or outcome variables and excluding these values can create a bias in the results. For the imputed data, we created 10 datasets with a seed of 12345 to increase reproducibility and to improve the accuracy of the averaged results of the imputed datasets. To select the variables for the prediction model, we performed backward and forward selection, elimination of variables with high multicollinearity, and bootstrapping. For bootstrapping, we made 1000 bootstrap samples and for each bootstrap sample a backward selection was performed. We then counted how many times the variables were selected in each bootstrap sample and repeated this for all the imputed datasets. We plotted these models on a calibration plot; however, to get more objective metrics to determine our final model, we calculated the average Brier score and C-index for model 2 and 3 with all 10 imputed datasets. As the results demonstrate, model 2 has the lower Brier score and C-index. Therefore, our final model is model 2 with the risk factors- sex, alcohol use, smoking, hypertension, glucose, HbA1c, diabetes, familiar Hc, Lpa, and homocysteine.