BIOST/STAT 579
Final Project
Yuhan Wang

Statistical Analysis Plan

BACKGROUND

Venous thromboembolism (VTE) is a blood clot that forms within a vein and is a condition that can be fatal. Identifying clinical risk factors associated with increased risk of a subsequent VTE episode or death in those who have experienced VTE is important for improving treatment choices.

OBJECTIVE

The primary purpose of the cohort study is to investigate the risk factors of adverse VTE sequelae. For the part of statistical inference, we test the null hypothesis that there is no association between sex and the risk of VTE recurrence or death in the year after incident VTE. To better characterize the association, further analysis is conducted on how menopausal status modify the association between sex and risk. For the part of prediction, a prognostic model is built to predict whether individuals will experience VTE recurrence or death within one year after an incident VTE.

STUDY METHODS

This cohort study was undertaken using electronic medical records of 1,922 members, aged 30-89, of a health maintenance organization, who had experienced an incident non-fatal VTE event (DVT and/or PE) and who did not have cancer diagnosis and/or treatment in the two years prior to the VTE event. Each individual was followed for one year after the incident VTE event, during which VTE recurrence and death were recorded. The primary focus of this component of the study was adverse VTE sequelae under standard care within the first year after incident VTE. An extension cohort was also considered, consisting of 1,013 individuals from the initial cohort whose treatment on ACT terminated within the year following their incident VTE and who had not died or experienced recurrence during that year. These individuals were followed for a further one year during which subsequent VTE recurrence and death were recorded.

STATISTICAL PRINCIPLES

The primary outcome of the study was the risk of VTE recurrence or death in the year after incident VTE.

All statistical models were chosen a priori. For primary analysis on the association between sex and the risk of VTE recurrence or death in the year after incident VTE, logistic regression model was applied. The outcome is the indicator of VTE recurrence or death in the year after incident VTE (recvte=1|death=1). The model includes sex as the main predictor. Covariates included are age, body mass index, post-menopausal status, oral contraceptive use, hormone replacement therapy, anticoagulation therapy, and previous cardiovascular disease. The point estimate and confidence interval for the coefficients of sex in logistic regression model measures the log odds ratio of VTE recurrence or death between male and female group. P value indicates the statistical significance of the association.

For secondary analysis on the association between menopausal status and the risk of VTE recurrence or death in the year after incident VTE, logistic regression model was applied. Because menopausal status does not apply to the male, a subset of female patients is selected as the dataset for secondary analysis. The regression model includes post-menopausal status as the main predictor. Covariates included are age, body mass index, oral contraceptive use, hormone replacement therapy, anticoagulation therapy, and previous cardiovascular disease. The point estimate and confidence interval for the coefficients of menopausal status in logistic regression model will be reported. P value indicates the statistical significance of the association.

For predictive modelling, logistic regression model will be applied. The main model will include all the baseline variables except participant id. Based on the result of main model, I will run several reduced models, which only retain predictors with $|Z|>2$, $|Z|>1.5$ in the main model, respectively. Expanded models will also be built, adding interaction terms between age and all other variables. ROC curve of these models will be drawn and compared. The model with highest ROC value will be the prognostic model. Furthermore, the prognostic model will be applied on the dataset of extension cohort, assessing the accuracy of identifying high risk individuals within two years.

Result

Patients

During the first year of study, a total of 1922 patients who had experienced an incident non-fatal VTE event (DVT and/or PE) was followed. There are 1680 individuals who had not died or experienced recurrence during that year. There are 242 individuals who had died or experienced recurrence during that year. The demographic and baseline clinical characteristics between two groups were presented in Table 1.

Table 1: baseline characteristic of subjects stratified by indicator of recurrence or death within one year

	No recurrence or death (N=1680)	recurrence or death (N=242)	Overall (N=1922)
Type of incident VTE			
DVT	806 (48.0%)	117 (48.3%)	923 (48.0%)
PE	582 (34.6%)	77 (31.8%)	659 (34.3%)
both	292 (17.4%)	48 (19.8%)	340 (17.7%)
Age at incident VTE (years)			
Mean (SD)	64.5 (14.2)	71.4 (13.4)	65.3 (14.3)
Median [Min, Max]	64.9 [30.1, 89.9]	73.8 [31.3, 89.8]	66.2 [30.1, 89.9]
sex			
Female	913 (54.3%)	131 (54.1%)	1044 (54.3%)
Male	767 (45.7%)	111 (45.9%)	878 (45.7%)
race			
African American	87 (5.2%)	11 (4.5%)	98 (5.1%)
White	1516 (90.2%)	221 (91.3%)	1737 (90.4%)
other	63 (3.8%)	7 (2.9%)	70 (3.6%)
unknown	14 (0.8%)	3 (1.2%)	17 (0.9%)
Post-menopausal status			
No	167 (9.9%)	7 (2.9%)	174 (9.1%)
Post	716 (42.6%)	122 (50.4%)	838 (43.6%)
Peri	22 (1.3%)	2 (0.8%)	24 (1.2%)
Missing	775 (46.1%)	111 (45.9%)	886 (46.1%)
Body mass index (kg/m^2)			
Mean (SD)	31.2 (7.61)	29.9 (8.90)	31.1 (7.79)
Median [Min, Max]	30.0 [14.3, 97.7]	28.6 [15.5, 71.2]	29.8 [14.3, 97.7]
Missing	5 (0.3%)	3 (1.2%)	8 (0.4%)
Anticoagulation therapy at baseline			
No	16 (1.0%)	20 (8.3%)	36 (1.9%)
Yes	1664 (99.0%)	222 (91.7%)	1886 (98.1%)
Smoking status			
never smoker	852 (50.7%)	106 (43.8%)	958 (49.8%)
current smoker	132 (7.9%)	22 (9.1%)	154 (8.0%)
former smoker	690 (41.1%)	111 (45.9%)	801 (41.7%)
Missing	6 (0.4%)	3 (1.2%)	9 (0.5%)
Previous cardiovascular disease			
No	1323 (78.8%)	151 (62.4%)	1474 (76.7%)
Yes	357 (21.2%)	91 (37.6%)	448 (23.3%)
Hormone replacement therapy use at baseline			
No	811 (48.3%)	118 (48.8%)	929 (48.3%)
Yes	102 (6.1%)	13 (5.4%)	115 (6.0%)
Missing	767 (45.7%)	111 (45.9%)	878 (45.7%)
Oral contraceptive use at baseline			
No	837 (49.8%)	129 (53.3%)	966 (50.3%)
Yes	76 (4.5%)	2 (0.8%)	78 (4.1%)
Missing	767 (45.7%)	111 (45.9%)	878 (45.7%)

Outcome

Table 2: Transformed coefficient for predictors and p value in logistic regression model

	Transformed coefficient	95% CI	P value
Sex	3.16	(0.62, 16.22)	0.167
Age at incident VTE (years)	1.03	(1.012, 1.043)	<0.01*
HRT(yes=1,no=0)	1.07	(0.56, 2.07)	0.830
OC (yes=1, no=0)	2.66	(0.44, 15.97)	0.284
BMI (kg/m^2)	0.99	(0.97, 1.02)	0.645
ACT (yes=1, no=0)	0.10	(0.05,0.21)	<0.01*
Pstmp * I(sex=0)			
Pstmp=1	1.086	(0.35, 3.32)	0.88
Pstmp=2	1.98	(0.37, 10.58)	0.42
Pstmp=.	--	--	<0.01*
priorCVD (yes=1,no=0)	1.73	(1.28, 2.33)	<0.01*

From logistic regression model, the odds of recurrence or death is 3.16 (95% CI: 0.62, 16.22) folds higher in male group than female group. With two sided p value 0.167, we have no strong evidence to reject the null hypothesis that there is no association between the risk of VTE recurrence or death in the year after incident VTE and sex.

Secondary analysis

Table 3: Transformed coefficient for predictors and p value in logistic regression model within female patients

	Transformed coefficient	95% CI	P value
Age at incident VTE (years)	1.02	(0.998, 1.046)	0.07
HRT(yes=1,no=0)	0.91	(0.47, 1.75)	0.78
OC (yes=1, no=0)	0.37	(0.06, 2.27)	0.28
BMI (kg/m^2)	1.00	(0.97, 1.04)	0.60
ACT (yes=1, no=0)	0.10	(0.04,0.25)	<0.01*
Pstmp * I(sex=0)			
Pstmp=1	1.35	(0.39, 3.4.69)	0.64
Pstmp=2	1.98	(0.37, 10.51)	0.42
priorCVD (yes=1,no=0)	1.76	(1.16, 2.67)	<0.01*

From logistic regression model, the odds of recurrence or death is 1.35 (95% CI: 0.39, 3.4.69) folds higher in Pre-status group than no status group. The odds of recurrence or death is 1.98 (95% CI: 0.37, 10.51) folds higher in Post status group than no status group. With two sided p value 0.64 and 0.42, respectively, we have no strong evidence to reject the null hypothesis that there is no association between the risk of VTE recurrence or death in the year after incident VTE and post-menopausal status.

Prediction

Outcome: binary variable, indicator of recurrence or death within one year

Basic model

	Odds ratio	95% CI	P value
Age	1.03	(1.02, 1.04)	<0.01
Act	0.12	(0.06, 0.24)	<0.01
Priorcvd	1.68	(1.25, 2.26)	<0.01

Model 1

	Odds ratio	95% CI	P value
Age	1.03	(1.02, 1.05)	<0.01
Act	0.11	(0.06, 0.23)	<0.01
Priorcvd	1.67	(1.23, 2.26)	<0.01
Smoker			
Current smoker	1.78	(1.04, 3.03)	0.03
Former smoker	1.16	(0.86, 1.56)	0.32
Vte_type			
PE	0.83	(0.69,1.14)	0.25
Both	1.16	(0.81,1.67)	0.42

Model 2

	Odds ratio	95% CI	P value
Act	0.10	(0.04, 0.25)	<0.01
Priorcvd	1.88	(1.24, 2.83)	<0.01
Pstmp			
Peri	3.47	(1.54, 7.83)	<0.01
post	2.39	(0.46, 12.49)	0.30
Hrt	0.81	(0.43, 1.53)	0.52

Model 3

	Odds ratio	95% CI	P value
Age	1.03	(0.98, 1.09))	0.22
Act	0.16	(0.004,6.14)	0.33
Priorcvd	0.37	(0.04, 3.38)	0.38
Hrt (no=0, yes=1)	0.89	(0.46, 1.72)	0.74
Age*act	0.99	(0.94, 1.05)	0.85
Age*priorcvd	1.02	(0.99, 1.05)	0.17

From the comparison of AUC score of the four models above, model 3 with interaction terms performs the best with AUC=0.632, basic model performs the second with AUC=0.629, model 1 the third with AUC=0.620, model 2 the last with AUC=0.592.

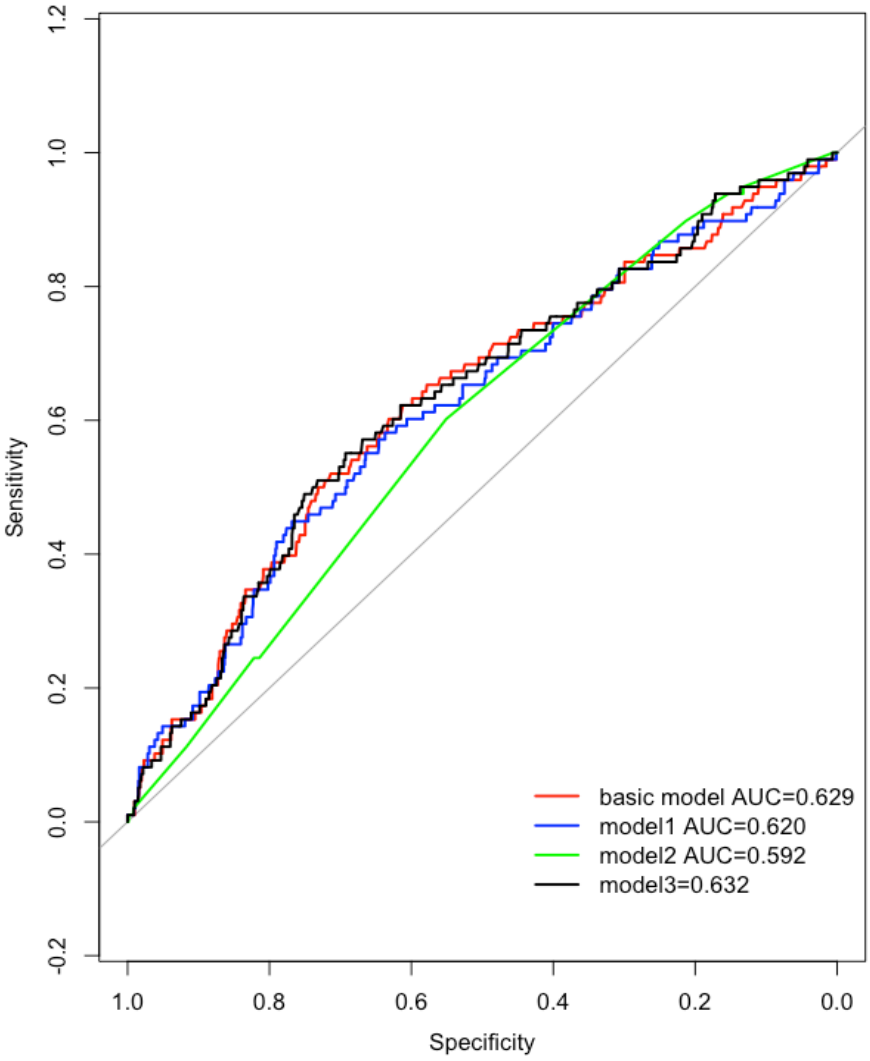

Figure 1 :
comparison of ROC curves of four predicting models

Therefore, from the analysis above, I choose model 3 as the predicting model, which includes age, act, priorcvd, hrt as covariates. Interaction terms between age and act, age and priorcvd are also included.

Limitation

From figure 4 and figure 5, we cannot see obvious pattern of missing data except missing due to sex. Some characteristic variables do not apply to the male, like menopausal status. Despite of that, there is still possibility of non ignorable missing, which cannot be fully solved. Further sensitivity analysis is needed. For prediction model, data of cohort within one follow up year are used as training set. Data of extension cohort from first follow up year to second follow up year are used as test set. However, the outcome variables in training set and test set are different. In training set, we predict the risk of recurrence or death in year 1 following incident VTE. In test set, we predict the risk of recurrence or death in year 2 following incident VTE. It cannot be ignored that the risk of recurrence or death will change over time. Therefore, the performance of models on extent cohort may not necessarily represent the true accuracy of prediction of risk within one year following incident VTE. To build a more accurate model, new dataset should be collected as test set, which means the enrollment of new patients. It will be more expensive than just using extent cohort.

Table and Figure

Table 4: comparison of recurrence or death rate by sex during 1st year and 2nd year

	Female	Male	Overall
1 year follow up	0.125	0.126	0.126
2 year follow up	0.096	0.097	0.097

Table 5: comparison of recurrence or death rate by Post-menopausal status in female

Menopausal status	No	Post	Peri	All Female
1 year follow up	0.040	0.145	0.083	0.125
2 year follow up	0.048	0.113	0.053	0.096

Figure 2: data distribution of numeric variable

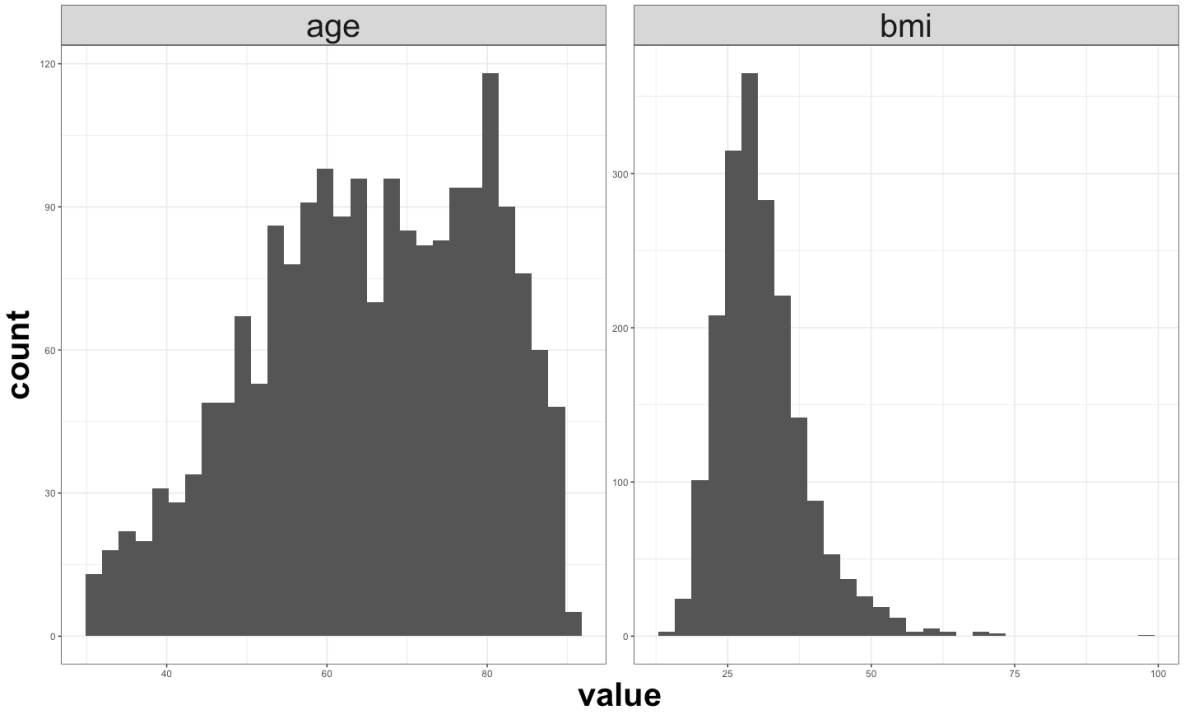

Figure 3: Time after incident VTE to recurrent event or death (days) in 1st year

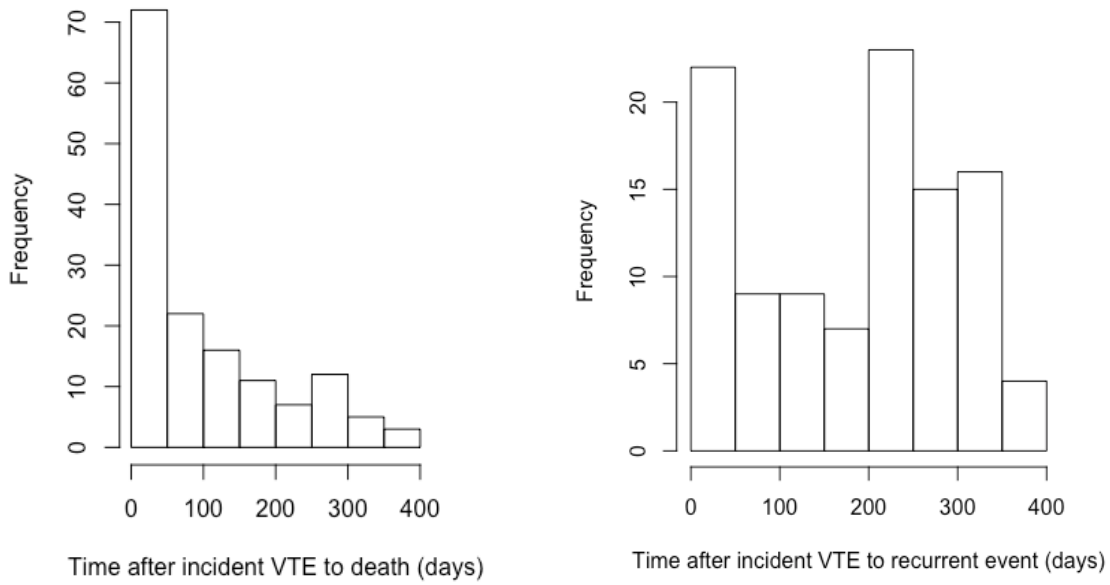

Figure 4: Time after incident VTE to recurrent event or death (days) in extension cohort

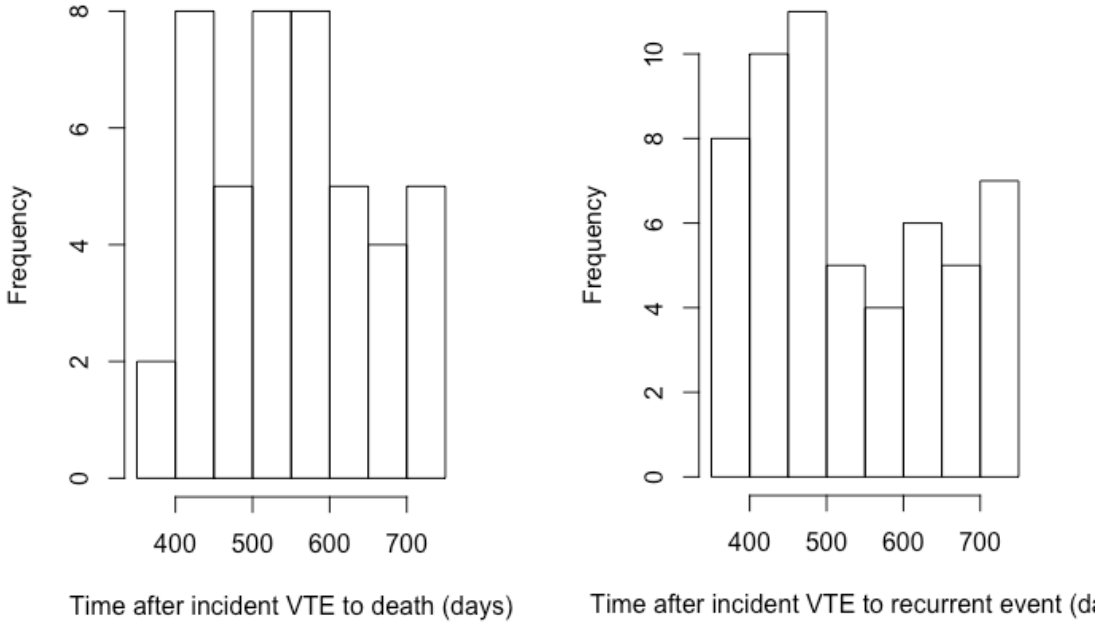

Figure 5: pattern of missing data

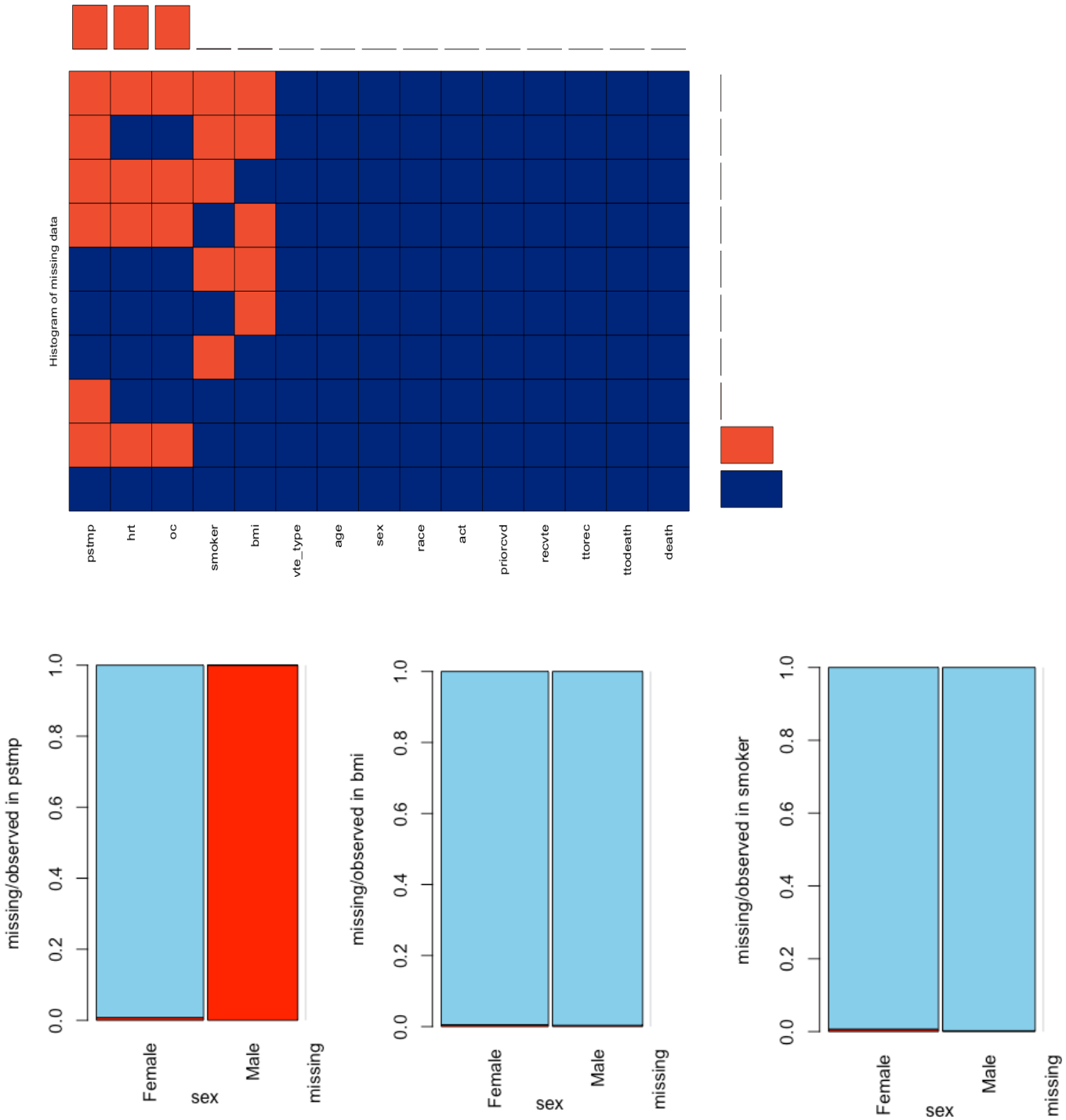

Figure 6: associations between missing and observed data