Final Project

Pinyan Liu

Abstract

Objective: A venous thromboembolism (VTE) is a blood clot that forms within a vein and is a condition that can be fatal. In order to identify clinical risk factors associated with increased risk of subsequent VTE episode or death in those who have experienced VTE is important for improving treatment choices. **Methods:** In order to identify whether the risk of VTE recurrence or death in the year after incident VTE differs by sex, logistics regression models were conducted separately for recurrence or death, adjusted for age, sex, obesity, oral contraceptive use and hormone replacement therapy. for question 2, logistic regression models were fit to study the role of menopausal status playing in differences in risk by sex. For question 3, developing a prognostic model to predict whether individuals will experience VTE recurrence or death within one year after an incident VTE and then assess them by ROC curves and AUC area. **Results:** Data from 1,922 participants in the initial cohort and 1,013 individuals in the extension cohort were analyzed. For question 1, there is no obvious evidence showing that the risk of VTE recurrence or death in the year after incident VTE differs by sex. For question 2, menopausal status modifies the association between the risk of VTE recurrence or death and sex. For question 3 and 4, the prognostic model to predict whether individuals will experience VTE death within one year after an incident VTE is more accurate than that of VTE recurrence. **Conclusion:** Clinical risk factors associated with increased risk of subsequent VTE episode or death in those who have experienced VTE are more complicated than what we could imagine. More analyses need to be done to the role of sex-based variables playing in the scientific question.

1.Background

A venous thromboembolism (VTE) is a blood clot that forms within a vein and is a condition that can be fatal. Deep vein thrombosis (DVT), which is a specific, and more serious, type of VTE most often occurring in the leg. The clot can block blood flow back to the heart and may damage the valves in the vein. Also, the clot may detach and travel to major organs such as the lungs, resulting in a pulmonary embolism (PE). Individuals with a first episode of VTE are known to be at increased risk of a subsequent VTE episode and/or death.

Anticoagulant therapy (ACT) is one of the typical treatments for VTE. Duration of ACT therapy after VTE varies, but, when prescribed, it is usually for at least 3 months. While ACT is known to reduce risk of recurrent VTE, it also increases risk

of hemorrhage. The two risks, recurrence and bleeding, need to be balanced against each other and the optimal duration of ACT after VTE remains a clinical dilemma.

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. Active cancer and its treatment are known to increase the risk of a first VTE event. The risk of first VTE is also reported to increase with age, obesity, oral contraceptive use and hormone replacement therapy, but the association of these and other factors, such as sex, with adverse sequelae such as VTE recurrence and/or death among individuals who have already had a first VTE episode is unclear.

2.Statistical Analysis Plan

The cohort study aimed at investigating risk of adverse VTE sequelae 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. A recurrent VTE event was defined as a clot at a new location at any time after the incident event or at the incident location at least 14 days after the incident event. The primary focus of this component of the study was adverse VTE sequelae under standard care within the first year after incident VTE. To provide information on risk of adverse VTE sequelae after completion of ACT, 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. The study design is shown in *Figure 1*.

Univariate descriptive statistics will be provided for all baseline variables including type of incident VTE, age at incident VTE (years), sex, race, post-menopausal status, body mass index, anticoagulation therapy at baseline, smoking status, previous cardiovascular disease, hormone replacement therapy use at baseline, and oral contraceptive use at baseline. Statistics presented will be number and percent of cases for categorical variables and mean and standard deviation for continuous variables. Variables will be subset into two groups defined by sex.

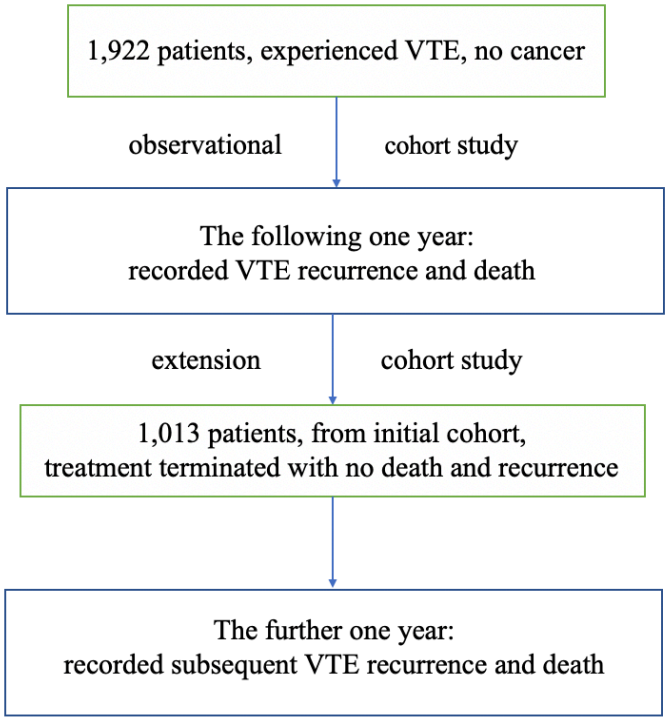

Figure 1. The design of study.

For the missing values in the dataset, all "." indicating missingness will be replaced by NA. Then, the way to deal with it depends on how much missing data do we have in the dataset. If there is less than 10% missing data, it's reasonable to exclude them, as they won't have significant impact on the result. Otherwise, if there is more than 10% missing data, it's more rational to use specific models to cope with it, depending on what kind of missingness do we have. To determine the type of missingness, first, graphical exploration will be done separately for MCAR, MAR and NI. By plotting the missing values and non-missing values, if they are completely random and overlapping the whole way, then they will be treated as MCAR. If they also have overlapping missingness and observations, but not completely random, for example, more missing values as time passes by. Then, this will be treated as MAR, as the missingness may depend on the observed values. Last, if missing and observed values are totally not overlapped, then this is NI. Besides, additional exploratory analysis using data in wide format could also be done to make sure what type of missingness do we have. After confirming the type of missingness, I will choose multiple imputations accounting for clustering and "brute-force" pooling of results.

The primary purpose of the study is to investigate is there evidence that the risk of VTE recurrence or death in the year after incident VTE differs by sex. First, table

with frequency of recurrence or death within the year for female and male will be presented, which is the exploratory analysis. Then, we will fit two logistic models taking recurrence or death in the year after incident VTE as outcomes separately. The reason why we treated recurrence or death separately is that these two are totally two different outcomes and would involve different biological pathways and physiological activities. In other words, VTE recurrence may not lead to death. Therefore, the risk of VTE recurrence is not identical to the risk of VTE death. Sex will be considered as the predictor. In addition, as we have already known that the risk of first VTE is reported to increase with age, sex, obesity, oral contraceptive use and hormone replacement therapy, considering the potential impacts on the recurrence VTE and adverse sequelae, we include these factors as covariates in these two models. The point estimate and 95% confidence interval from a regression of risk difference of recurrence or death on sex can be interpreted as the association between the changes in the two measures for patients with similar baseline characteristics.

Question 2 aims to find the role of menopausal status playing in differences in risk by sex. For female, post-menopausal have three status, 0=No; 1=Post; 2=Peri. For all the male, as they don't have post-menopausal status, so they all have missing values for this variable. Then, a new variable "pstmp2" will be created that I will relabel the male with "3", which is another category for post-menopausal status. Then, two logistic regression models will be fitted which take recurrence or death in the year after incident VTE as outcomes separately. Then, we will include "sex" and "pstmp2" as covariates. GLM models will be fitted with interaction between sex and pstmp2 and no interaction. The point estimate, 95% confidence interval and P-values from regressions will be interpreted to investigate the role of menopausal status playing differ by sex.

Question 3 proposes to develop and assess a prognostic model to predict whether individuals will experience VTE recurrence or death within one year after an incident VTE. First, dataset will be randomly separated into two parts, the training data and the test data. Then, I will use training data first to develop the prognostic model. As age, sex, obesity, oral contraceptive use and hormone replacement therapy are all considered potentially related with risk of VTE and its recurrence. And in addition, race, body mass index and other factors like smoking are one's basic characteristics, so for this prediction model, I will choose to incorporate all of baseline characteristics in it. Then, the outcome is also the risk of recurrence or death. To assess the prognostic model, ROC curves will be drawn separately for these two models and then area under the curve will be compared with the lowest standard 0.5.

Question 4 aims to investigate whether the prognostic model would identify high risk individuals in the extension cohort accurately. In this question, we just treat the DataFile3 as the test data, and then fit the prediction model developed in question 3. ROC curve will also be drawn to show the accuracy of the prognostic model and then area under the curve will be calculated. The larger the area is, the more accurate the prognostic model will be.

3. Result

3.1 Descriptive statistics

	female (N=1044)	male (N=878)	Overall (N=1922)
Type of incident VTE			
1: DVT	479 (45.9%)	444 (50.6%)	923 (48.0%)
2: PE	403 (38.6%)	256 (29.2%)	659 (34.3%)
3: both	162 (15.5%)	178 (20.3%)	340 (17.7%)
age			
Mean (SD)	65.9 (15.5)	64.6 (12.7)	65.3 (14.3)
Race			
0: African-American	57 (5.5%)	41 (4.7%)	98 (5.1%)
1: White/Caucasian	933 (89.4%)	804 (91.6%)	1737 (90.4%)
2: Other	47 (4.5%)	23 (2.6%)	70 (3.6%)
9: Unknown	7 (0.7%)	10 (1.1%)	17 (0.9%)
Post-menopausal status			
0	174 (16.7%)	0 (0%)	174 (9.1%)
1	838 (80.3%)	0 (0%)	838 (43.6%)
2	24 (2.3%)	0 (0%)	24 (1.2%)
. (Missing)	8 (0.8%)	878 (100%)	886 (46.1%)
Body mass index (kg/m2)			
Mean (SD)	31.5 (8.46)	30.5 (6.87)	31.1 (7.79)
Missing	5 (0.5%)	3 (0.3%)	8 (0.4%)
Anticoagulation therapy at baseline			
No	21 (2.0%)	15 (1.7%)	36 (1.9%)

	female (N=1044)	male (N=878)	Overall (N=1922)
Yes	1023 (98.0%)	863 (98.3%)	1886 (98.1%)
Smoking status			
. missing value	7 (0.7%)	2 (0.2%)	9 (0.5%)
0: never smoker	585 (56.0%)	373 (42.5%)	958 (49.8%)
1: current smoker	83 (8.0%)	71 (8.1%)	154 (8.0%)
2: former smoker	369 (35.3%)	432 (49.2%)	801 (41.7%)
Previous cardiovascular disease			
No	828 (79.3%)	646 (73.6%)	1474 (76.7%)
Yes	216 (20.7%)	232 (26.4%)	448 (23.3%)
Hormone replacement therapy use at baseline			
. missing value	0 (0%)	878 (100%)	878 (45.7%)
No	929 (89.0%)	0 (0%)	929 (48.3%)
Yes	115 (11.0%)	0 (0%)	115 (6.0%)
Oral contraceptive use at baseline			
. missing	0 (0%)	878 (100%)	878 (45.7%)
No	966 (92.5%)	0 (0%)	966 (50.3%)
Yes	78 (7.5%)	0 (0%)	78 (4.1%)

Table 1. baseline characteristics by different sex. Data are mean (SD) or number (%). Include information for race, age and baseline characteristics for both female and male patients.

From Table 1, we could see that there were 1044 female and 878 male. As the study focused mainly on differences between male and female, so we group the baseline characteristics by sex. For age and body mass index, female and male are approximately the same, with the average age around 65 years old and mean body mass index around 31 kg/m2. For anticoagulation therapy at baseline, more than 98% female and male already had it. For smoking status, only 8% of them are current smoker, the rest of them are equally never smoker or former smoker. Besides, the majority of patients had previous cardiovascular disease. For hormone replacement therapy and oral contraceptive use at baseline, male didn't have these therapies, and for female, most of them had both two therapy.

3.2 Primary outcomes

From the exploratory analysis in Table 2, it is obvious that the risk of VTE recurrence or death in the year after incident VTE has almost no difference between female and male, especially in recurrence.

Table 2. Risk of VTE recurrence or death by different sex. Data are number (%).

	female (N=1044)	male (N=878)	Overall (N=1922)
Recurrence			
0	987 (94.5%)	830 (94.5%)	1817 (94.5%)
1	57 (5.5%)	48 (5.5%)	105 (5.5%)
Death			
0	961 (92.0%)	813 (92.6%)	1774 (92.3%)
1	83 (8.0%)	65 (7.4%)	148 (7.7%)

Then, we use "uwIntroStats" package in R to fit two logistic linear regression models, two outcomes are risk of recurrence and death separately. Coefficients of two models are presented in Table 3 and Table 4. From tables, we could see that for both the risk of VTE recurrence or death in the year after incident VTE, age has P-Value larger than 0.05. So we have not enough evidence to reject the null hypothesis that there is any difference between female and male. The odds of having the risk of VTE recurrence for female is almost the same for male, while the odds of having the risk of VTE death in the year for male is 10% higher than that for female. Then, we could also identify that, the adjusted covariates, like age, BMI, oral contraceptive use at baseline and hormone replacement therapy use at baseline all have no scientific significance to affect the risk of VTE recurrence. While for the risk of VTE death, it seems that age and hormone replacement have P-Value smaller than 0.05, which shows that they are related with the risk of VTE death in the year after incident VTE differs by sex and should be adjusted.

Table 3. Primary outcomes for recurrence and adjusted characteristics.

	Odds ratio	95% CI	P-Value
Sex	0.97	(0.64,1.5)	0.90
Age	1.0	(0.99,1.0)	0.90
BMI	1.0	(1.0,1.04)	0.11

Oral contraceptive use at baseline	0.44	(0.094,2.1)	0.30
Hormone replacement therapy use	1.1	(0.48,2.6)	0.80

Table 4. Primary outcomes for death and adjusted characteristics.

	Odds ratio	95% CI	P-Value
Sex	1.1	(0.72,1.5)	0.79
Age	1.1	(1.0,1.1)	<0.00005*
BMI	0.98	(0.95,1.0)	0.39
Oral contraceptive use at baseline	0.66	(0.28,1.5)	0.33
Hormone replacement therapy use	1.4e-07	(7.2e-07,2.6e-06)	<0.00005*

3.3 Secondary outcomes

3.3.1 Question 2

Before analyses, because all male have missing values of menopausal status, in order to incorporate them into the model and study the role of menopausal by sex, I created a new variable which "3" indicates male, and the rest of them are just the same. Additionally, we could not just include male into the "no menopausal status", because for female, the assumption of having no menopausal status is that they already have the time of menstruation. But for male, they will never have it. Therefore, there must exist some different metabolism and bio-pathways in the body, thus affect the risk of VTE recurrence or death. In this way, we could study different levels of menopausal status, including none, post, peri and male.

Then, after creating the new variable, I fit two logistic regression model incorporating only sex and menopausal status as the covariates, and also using death and recurrence of VTE as the outcome separately. Because male will never have menopausal status, we choose not include the interaction term between sex and menopausal status. In table 5 and table 6, we could find that sex have P-value smaller than 0.05, which means that we are confident to reject the null hypothesis that there is no association between risk of VTE recurrence or death and sex. Then, for menopausal status, it has P-Value smaller than 0.05 overall. Therefore, we could reject the null hypothesis that there is no association between risk of VTE recurrence or death and menopausal status. Then, we find that for both models, male indicated by "3" in menopausal status have P-Value smaller than 0.00005, although this variable only have 1 degree of freedom and cannot represent the menopausal status,

it could reflect that there does exist obvious difference of menopausal status between sex in differences in risk by sex.

In female, menopausal status plays unobvious role to effect the risk of VTE recurrence, while whether it is post menopausal status or peri menopausal status does effect the risk of VTE death. Then, combined female and male, menopausal status modifies the association between the risk of VTE recurrence or death and sex.

Table 5. Coefficient estimates of Question 2's model for risk of VTE recurrence

	Odds ratio	95% CI	P-Value
Intercept	0.036	(0.016,0.081)	<0.00005*
Sex	3.3e+5	(1.5e+5,7.1e+5)	<0.00005*
Menopausal status			<0.00005*
1: post	1.7	0.73	0.21
2: peri	2.6	0.48	0.27
3: male	2.9e-06	1.7e-06	<0.00005*

Table 6. Coefficient estimates of Question 2's model for risk of VTE death

	Odds ratio	95% CI	P-Value
Intercept	0.012	(0.0029,0.047)	<0.00005*
Sex	1.25e+06	(6.1e+5,2.6e+6)	<0.00005*
Menopausal status			<0.00005*
1: post	9.2	(2.2,37)	0.0021*
2: peri	5.5e-06	(1.3e-06,2.4e-05)	<0.00005*
3: male	5.5e-06	(1.2e-06,2.6e-05)	<0.00005*

3.3.2 Question 3

In order to do prediction, firstly, datafile 2 was subset into training data and test data, which took up 75% and 25% of datafile 2 separately. Firstly, I just incorporated all baseline characteristics variables in the prediction model, however, AUC area under the ROC curve is pretty low. Then, I chose to delete some of them until I got the largest AUC area. The procedure is the same for both VTE recurrence and death.

Then, for model 1 taking VTE recurrence as the outcome, the final prediction model took type of incident VTE, age, sex, post-menopausal status, previous cardiovascular disease, hormone replacement therapy and oral contraceptive use at baseline into considerations. From table 7, we could see that only type of incident VTE has P-Value smaller than 0.05, showing scientific significance. Then, the AUC area under the ROC curve is 0.542, not pretty larger than minimum 0.5. Therefore, this prediction model to predict whether individuals will experience VTE recurrence within one year after an incident VTE is not pretty reasonable.

In addition, for model 2 taking VTE death as the outcome, compared with model 1, the final prediction model took additional BMI and smoking status into considerations. From table 8, we could see that both age and previous cardiovascular disease had P-Value smaller than 0.05, showing scientific significance, and could reject the null hypothesis that there is no association between the risk of VTE death and age or previous cardiovascular disease. In addition, the AUC area under the ROC curve is 0.67, which is better than model 1, indicating that this prediction model to predict whether individuals will experience VTE death within one year after an incident VTE is more accurate.

Table 7. Coefficient estimates of Question 3's prediction model for risk of VTE recurrence

	Odds ratio	95% CI	P-Value
Intercept	0.083	(0.022,0.32)	0.0003*
Vte_type2: PE	0.40	(0.21,0.74)	0.003*
Vte_type3: both	0.84	(0.46,1.5)	0.6
Age	0.99	(0.97,1.0)	0.6
Sex	3.4e+5	(0.0, inf)	0.9
Pstmp1: post	1.5	(0.43,5.0)	0.5
Pstmp2: peri	3.5	(0.6,20)	0.2
Pstmp3: male	3.3e-06	(0.0,inf)	0.9
Priorcvd	1.2	(0.69,2.2)	0.5
Hormone replacement therapy1: yes	0.90	(0.34,2.4)	0.8
Oral	0.49	(0.093,2.6)	0.4

contraceptive1: yes			

Table 8. Coefficient estimates of Question 3's model for risk of VTE death

	Odds ratio	95% CI	P-Value
Intercept	4.3e-10	(0.0,inf)	1.0
Vte_type2: PE	0.97	(0.61,1.5)	0.9
Vte_type3: both	1.2	(0.67,2.0)	0.6
Age	1.06	(1.0,1.1)	1.1e-07***
Sex	1.3e+06	(0.0,inf)	1.0
Pstmp1: post	0.55	(0.11,2.7)	0.5
Pstmp2: peri	4.9e-07	(0.0,inf)	1.0
Pstmp3: male	3.6e-07	(0.0,inf)	1.0
Priorcvd	2.2	(1.4,3.3)	0.0002***
BMI	0.99	(0.95,1.2)	0.4
Hormone replacement therapy1: yes	0.25	(0.058,1.1)	0.06
Oral contraceptive1: yes	5.4e-07	(0.0,inf)	1.0
Smoker0: never	4.4e+06	(0.0,inf)	1.0
Smoker1: current	1.0e+07	(0.0,inf)	1.0
Smoker2: former	6.5e+06	(0.0,inf)	1.0

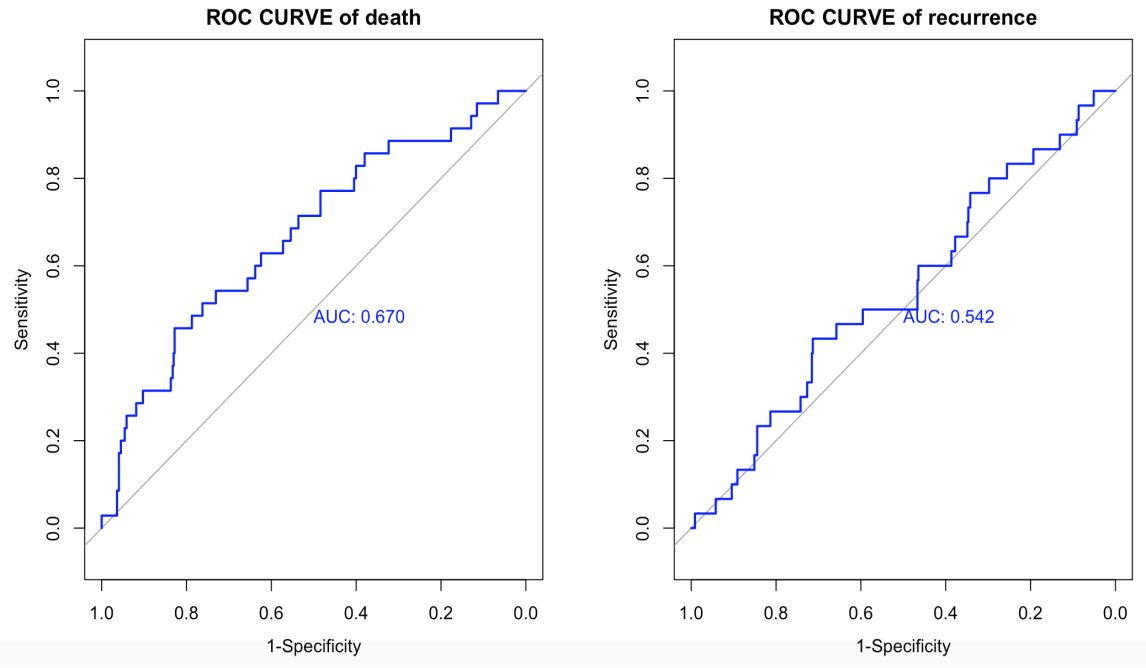

Figure 2. ROC curve of question 3.

3.3.3 Question 4

In order to test whether the prognostic model accurately identify high risk individuals in the extension cohort, I just treated the datafile 3 as the training data, and applied prediction models of VTE recurrence or death to the dataset. In addition, I combined datafile 3 with datafile 1 and 2 by participants "id" variable, so that we could make consistent that the prediction model also used the 1,013 individuals in the extension cohort. Then, two ROC curves were drawn separately for VTE recurrence and death in figure 3.

In figure 3, we could see that the right panel is the ROC curve of VTE recurrence, which has AUC area 0.51. This is not a ideal area for the ROC curve, because it is only a little bit larger than 0.5. Results indicated that model 1 not only can not predict accurately VTE recurrence within one year after an incident VTE but also not perform well to predict that in the following a further one year. Then, the left panel is the ROC curve of VTE death, which has AUC area 0.781. This is a more well-performed ROC curve and perform better than model 2 did in question 3. Results indicated that prediction model 2 performs pretty well in accurately identifying high risk individuals having VTE death in the following a further one year after 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.

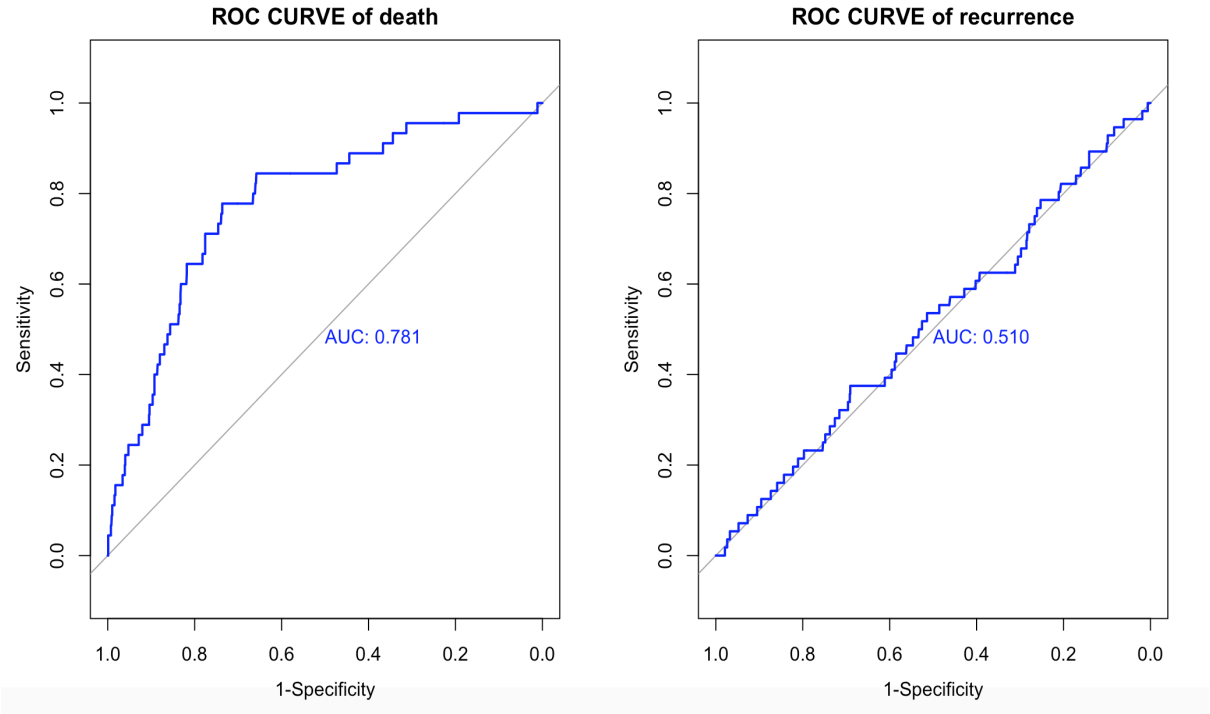

Figure 3. ROC curve of question 4.

4.Discussion

For question 1, there is no obvious evidence showing that the risk of VTE recurrence or death in the year after incident VTE differs by sex. While for the risk of VTE death, it seems that there is evidence showing that age and hormone replacement are related with the risk of VTE death in the year after incident VTE differs by sex. For question 2, as male have no menopausal status, so there won't be any interaction between that and sex. Therefore, we discussed this question based on subgroups of male and female. In female, menopausal status plays unobvious role to effect the risk of VTE recurrence, while whether it is post menopausal status or peri menopausal status does effect the risk of VTE death. Then, combined female and male, menopausal status modifies the association between the risk of VTE recurrence or death and sex. For question 3, model 1 taking VTE recurrence as the outcome, then the final prediction model took type of incident VTE, age, sex, post-menopausal status, previous cardiovascular disease, hormone replacement therapy and oral contraceptive use at baseline into considerations. This model has AUC area under the ROC curve 0.542. For model 2 taking VTE death as the outcome, compared with model 1, the final prediction model took additional BMI and smoking status into considerations. This model has AUC area under the ROC curve 0.67. Then, in question 4, after applying models developed in question 3 into the extension cohort, we got the AUC area under the ROC curve of VTE recurrence is

0.51, while the AUC area under the ROC curve of VTE death is 0.781, indicating that prognostic model to predict whether individuals will experience VTE death in the following further year is better than that of VTE recurrence.

However, there are still some limitations in the whole study. Firstly, based on the current knowledge, we only know that the risk of first VTE is reported to increase with age, obesity, oral contraceptive use and hormone replacement therapy, but the association of these and other factors, such as sex, with adverse sequelae such as VTE recurrence and/or death among individuals who have already had a first VTE episode is unclear. As we could infer that this is might also related with the recurrence or death of VTE, but this is just a guess. Therefore, considering them into the adjusted covariates may not be proper. Also, we may miss some other important factors which will contribute to the event. This is one of the reasons that the accuracy of prognostic models, especially for VTE recurrence is not pretty well, although I have tried best to get the largest AUC area.

Secondly, for some participant characteristics at baseline, for example, post-menopausal status, hormone replacement therapy use at baseline and oral contraceptive use at baseline, they do not have values for male. There are two ways I could think to cope with them. First, we could just drop out those missing values for male, but that variable is just a sex-specific one, which may not be appropriate. The other way is to revalue those missing values with a new category, indicating that this is for male especially. However, when we incorporated them as adjusted covariates, due to the singularities, all coefficients including standard error, P-Value would be NA in the results. Therefore, I just consider those sex-based variables may affect our models and analyses. And in the future, we have to identify more efficient methods to cope with them.