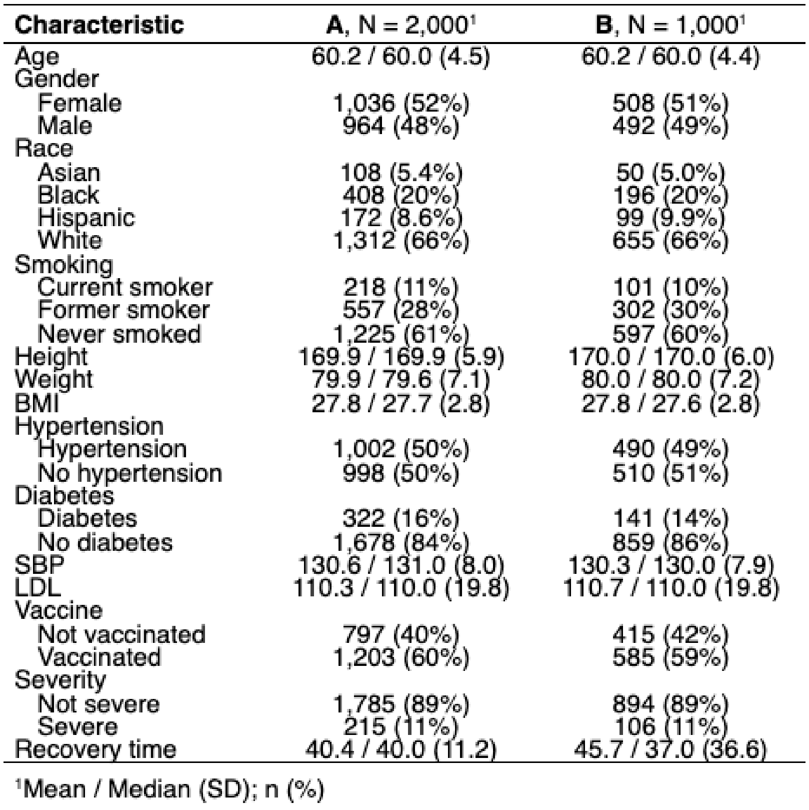
Final Report

Chen Liang (cl4469), Xinyi Shang (xs2529), Yuki Joyama (yj2803)

Exploratory Analysis and Data Visualization

The information of COVID-19 recovery time and other variables (id, gender, race, smoking history, height, weight, body mass index (BMI), history of hypertension and diabetes, systolic blood pressure (SBP), LDL cholesterol (LDL), vaccination status at the time of infection) is collected from two existing cohort studies. Baseline characteristics are presented in Table 1, showing that almost all characteristics are similar between the two study groups, except for COVID-19 recovery time.

Table 1: Baseline Characteristics



Characteristic	**A**, N = 2,000[1]	**B**, N = 1,000[1]
Age	60.2 / 60.0 (4.5)	60.2 / 60.0 (4.4)
Gender		
Female	1,036 (52%)	508 (51%)
Male	964 (48%)	492 (49%)
Race		
Asian	108 (5.4%)	50 (5.0%)
Black	408 (20%)	196 (20%)
Hispanic	172 (8.6%)	99 (9.9%)
White	1,312 (66%)	655 (66%)
Smoking		
Current smoker	218 (11%)	101 (10%)
Former smoker	557 (28%)	302 (30%)
Never smoked	1,225 (61%)	597 (60%)
Height	169.9 / 169.9 (5.9)	170.0 / 170.0 (6.0)
Weight	79.9 / 79.6 (7.1)	80.0 / 80.0 (7.2)
BMI	27.8 / 27.7 (2.8)	27.8 / 27.6 (2.8)
Hypertension		
Hypertension	1,002 (50%)	490 (49%)
No hypertension	998 (50%)	510 (51%)
Diabetes		
Diabetes	322 (16%)	141 (14%)
No diabetes	1,678 (84%)	859 (86%)
SBP	130.6 / 131.0 (8.0)	130.3 / 130.0 (7.9)
LDL	110.3 / 110.0 (19.8)	110.7 / 110.0 (19.8)
Vaccine		
Not vaccinated	797 (40%)	415 (42%)
Vaccinated	1,203 (60%)	585 (59%)
Severity		
Not severe	1,785 (89%)	894 (89%)
Severe	215 (11%)	106 (11%)
Recovery time	40.4 / 40.0 (11.2)	45.7 / 37.0 (36.6)

[1] Mean / Median (SD); n (%)

Model Training

In this section, describe the models you used to predict the time to recovery from COVID-19. Briefly state the assumptions made by using the models. Provide a

detailed description of the model training procedure and how you obtained the final model.

Results

Our final MARS model is as follows:

$\hat{y}$ = 22.435 + 3.574 × h(30.3 - bmi) + 9.783 × h(bmi - 30.3) * studyB + -6.264 × vaccine + 2.991 × h(164 - height) * h(bmi - 30.3) * studyB + 4.898 × h(bmi - 25.7) + -2.64 × h(87.6 - weight) * h(bmi - 30.3) * studyB
where $h(.)$ is hinge function.

The summary of the final MARS model is shown in Table 2.

Figure 2 illustrates that study B, BMI, height, weight, and vaccination status have the non-zero importance value in the final model.

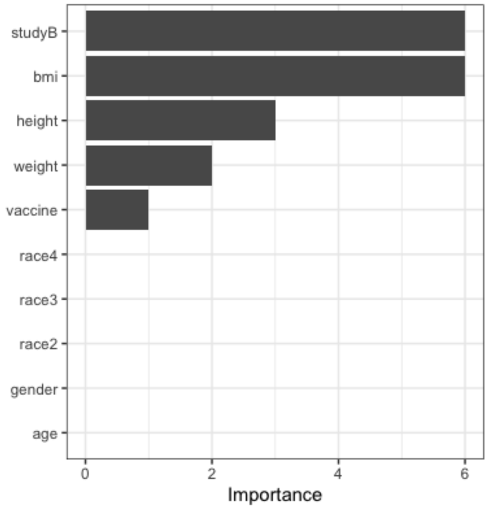


Figure 1: Variance Importance Plot

Table 2: Summary of the MARS model

Equation	Coefficients
(Intercept)	22.435204
vaccine	-6.264022
h(bmi-25.7)	4.898496
h(30.3-bmi)	3.574364
h(bmi-30.3) * studyB	9.782606
h(164-height) * h(bmi-30.3) * studyB	2.990502
h(87.6-weight) * h(bmi-30.3) * studyB	-2.640353

Conclusions

In this section, summarize your findings from the model analysis and discuss the insights gained into predicting time to recovery from COVID-19.

Additional Considerations

In your modeling efforts, did you include "study" as a predictor variable? Provide a rationale for your decision, considering the variable's relevance and potential impact on model accuracy and interpretability.