# SMI assessment 2

11326396

February 6, 2024

## 1 Exploratory Data Analysis (EDA)

First, I load the library (mice, dplyr, ggplot2), and load the data. Using the summary() to get an overview of the data, I found out that some patients don't have a diagnosis of heffpox (the value is equal to 0). So I deleted the rows of data on patients without heffpox, as the task is to examine whether milnepan is effective in reducing the risk of death among patients with heffpox. Then I set some categorical variables to factor type, and change the variable name of milnepan to treat, which can be better described.

```
dat <- dat %>%
  select(heffpox, sex, age, bmi, smoking, diabetes, milnepan, icu, death7day) %>%
  filter(heffpox == 1) %>%
  mutate(treat = as.factor(milnepan),
         sex = as.factor(sex),
         smoking = as.factor(smoking),
         diabetes = as.factor(diabetes),
         icu = as.factor(icu),
         death7day = as.factor(death7day)) %>%
  select(-heffpox, -milnepan)
```

Finally, I got the data with a dimension of 8219 * 8. The basic summary of each variable is shown in turn (Table 1), From this, we can see that there is missing data in smoking and bmi variables.

| Variable | Summary (N = 8219) | Missing Data, n (%) |
|---|---|---|
| sex, n (%) | | 0 (0%) |
| male | 3964 (48.23%) | |
| female | 4255 (51.77%) | |
| smoking, n (%) | | 370 (4.50%) |
| Yes | 2016 (24.53%) | |
| No | 5833 (70.97%) | |
| diabetes, n (%) | | 0 (0%) |
| Yes | 799 (9.72%) | |
| No | 7420 (90.28%) | |
| icu, n (%) | | 0 (0%) |
| Yes | 7341 (89.32%) | |
| No | 878 (10.68%) | |
| treat, n (%) | | 0 (0%) |
| Yes | 3852 (46.87%) | |
| No | 4367 (53.13%) | |
| death7day, n (%) | | 0 (0%) |
| Yes | 3195 (38.87%) | |
| No | 5024 (61.13%) | |
| age, mean (min/max) | 44.45 (18/97) | |
| bmi, mean (min/max) | 27.94 (14.98/42.80) | 2708 (32.95%) |

Table 1: Baseline summary table of the variables in the dataset, along with summaries of the number of missing data in each

From Table 1, We can see some missing data in the bmi variable, up to 32.95% of the whole data, and a few missing data in the smoking variable, accounting for only 4.50%, and other variables in this data do not have any missing data. the gender ratio of patients is relatively balanced, with females accounting for 51.77%, and the majority of patients are non-smokers, accounting for 70.97%, and only a small proportion (9.72%) of people suffer from diabetes. What's more, nearly half of the patients (46.87%) have been prescribed milnepan, and 10.68% of the patients are moved to intensive care, more than 1/3 of patients die within 7 days of admission, accounting for 38.87%.

To delve deeper into the missing data, I first view if the missingness co-occurs across all variables.

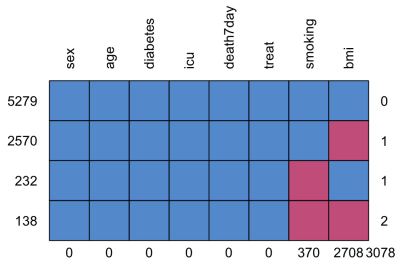
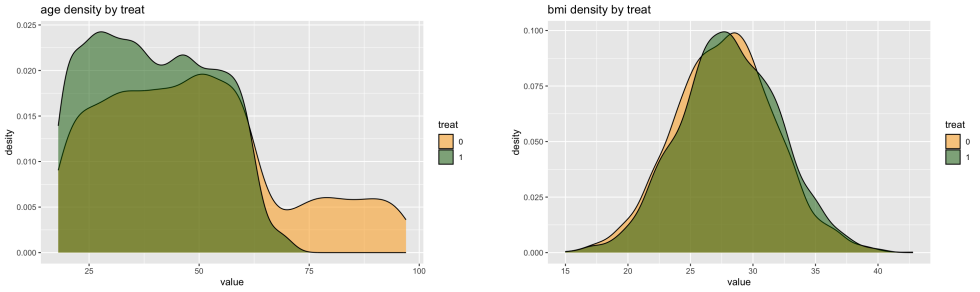



Figure 1: missing data patterns

From Figure 1, it seems that there is a small amount of co-occurrence in missing data between BMI and smoking (138), as it's not so significant, I'm ignoring this part for now.

In terms of other exploratory analyses, given the variable treat that we are interested in, to get a deeper analysis of continuous variables, a density plot would show the distribution of age and bmi variables by treat.




(a) Age distribution (b) bmi distribution

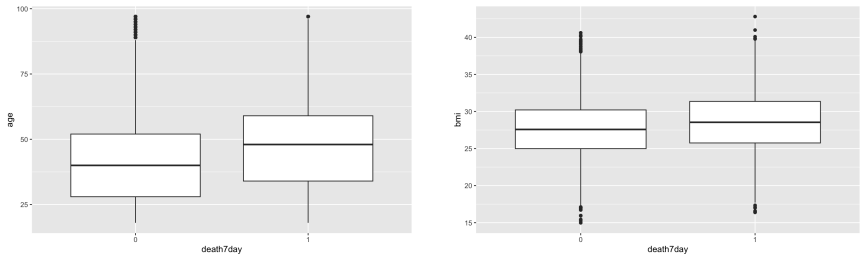
Figure 2: Density plot for continuous variables a)age and b)bmi

From Figure 2, the distribution of age spreads out across the range, with a wide base [18, 97], the age of half of the patients is in the range [30, 55], indicating that most patients are younger, and the long right tail suggests a smaller number of older patients in the data. The distribution of body mass index appears to be approximately normal, the peak is around 30, which is on the border between overweight and obese according to the standard BMI categories, indicating that most patients are relatively obese, and the tails on both ends are relatively thin, indicating that extreme BMI values are less common.

# 2 DAG

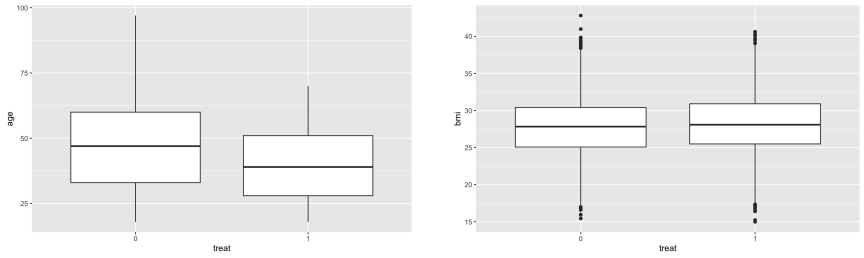
DAG is a good way to identify confounders in causal inference. To check if all variables (except exposure and outcome) are potential confounders, I draw a DAG thinking about 1) the temporal order of variables, in my opinions, it's (age, bmi, sex) ->(diabetes, smoking) ->treat ->icu ->death7day. 2)Using the contingency table and Pearson's Chi-squared test to analyze the association among all

binary variables (which is too much, so I show it in the DAG as if the p-value between two binary variables is <0.01, I will add the association (arrow) to the DAG), and using the regression model and boxplots to analyze the association among 2 continuous variables, shown in Figure 3 and Figure 4.




(a) Age distribution (b) bmi distribution

Figure 3: Boxplot for continuous variables a)age and b)bmi by the outcome death7day




(a) Age distribution (b) bmi distribution

Figure 4: Boxplot for continuous variables a)age and b)bmi by the exposure treat

From Figure 3, we can see that the age distributions differ in the outcome death7day, and also differ in the exposure treat, indicating that age may be the potential confounder. From Figure 4, the bmi distributions have relatively less difference in the outcome death7day, and little difference in the exposure treat, but the p-value is still <0.01, indicating that age may be the potential confounder.

After the above analysis, I drew a DAG shown in Figure 5, which I think that icu is a mediator, as 798 patients of all 878 patients having accepted icu died within 7 days, accounting for 90.89%, indicating the close association between icu and death7day, in addition, life experience is also a reason. The only variable that is not statistically associated with any other variables is sex, as the logistic regression model and Chi-squared test do not show any significant relationship. The relationship among other variables is shown in Figure 5.

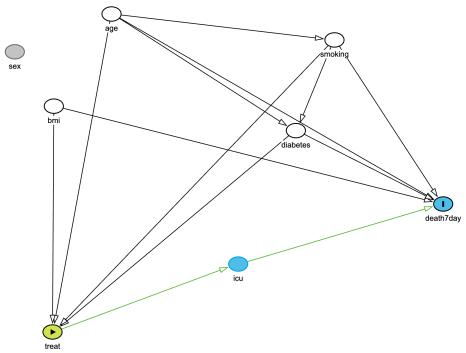



Figure 5: DAG of all variables

# 3 Deal with missing data

## 3.1 Assumptions of MAR mechanism

From the patterns across the variables in Figure 1, we can see that there seems like no dependence in missing data between BMI and smoking. Then investigate if missingness correlates with other variables, I used logistic regression where the dependent variable (bmi, smoking) is an indicator of missingness and the independent variables include the outcome and other predictors.

| Variable | Coefficient Estimate (se) | p-value |
|---|---|---|
| Intercept | 27.412 (0.180) | <2e-16 |
| sex (Male vs Female) | -0.029 (0.110) | 0.791 |
| age | -0.001 (0.004) | 0.786 |
| smoking (Yes vs No) | -0.138 (0.127) | 0.280 |
| diabetes (Yes vs No) | 0.159 (0.163) | 0.332 |
| treat (Yes vs No) | 0.476 (0.113) | 2.58e-05 |
| icu (Yes vs No) | 0.147 (0.193) | 0.445 |
| death7day (Yes vs No) | 0.918 (0.123) | 1.17e-13 |

Table 2: Results from the linear regression model of other variables on bmi.

The result of Table 2 suggests that treat variables and death7day variables have a significant relationship with the bmi variables, as the p-values <0.05.

| Variable | Coefficient Estimate (se) | p-value |
|---|---|---|
| Intercept | -1.730e+00 (2.452e-01) | 1.71e-12 |
| sex (Male vs Female) | 6.337e-02 (6.434e-02) | 0.3246 |
| age | 1.862e-02 (1.998e-03) | <2e-16 |
| bmi (Yes vs No) | -8.934e-03 (8.081e-03) | 0.2689 |
| diabetes (Yes vs No) | 2.277e-01 (9.108e-02) | 0.0124 |
| treat (Yes vs No) | 1.804e-02 (6.715e-02) | 0.7882 |
| icu (Yes vs No) | 2.185e-01 (1.088e-01) | 0.0447 |
| death7day (Yes vs No) | -4.112e-05 (7.253e-02) | 0.9995 |

Table 3: Results from the logistic regression model of other variables on smoking.

The result of Table 3 suggests that age variables, diabetes variables, and icu variables have a significant relationship with the smoking variables, as the p-values <0.05.

After the above analysis, we can make the assumption that the missing data were missing at random (MAR). About the assumption of multivariate normality, from the age density, the age variable does not conform to the normal distribution, I tried some transformations, but they did not work out, so I decided to use multiple imputations including predictive mean matching, which are robust to deviations from normality.

## 3.2 multiple imputation

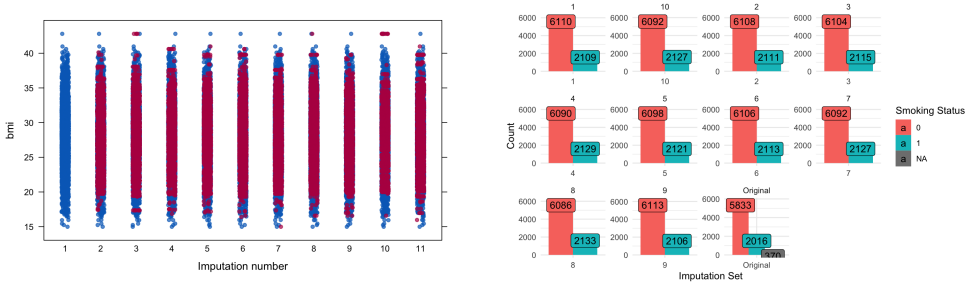
Use multiple imputations to get 10 imputed datasets.

```
imp_dat <- mice(dat, m = 10, maxit = 5)
```

From the above code, we can see that 10 imputed datasets are generated, the maximum number of iterations is 5, "pmm" is used for all continuous variables (by default), and "polyreg" (Polytomous logistic regression) is used (by default) for all categorical variables.

Then imputed data distribution for bmi and smoking is shown in Figure 6. We can that a) The density of the bmi distribution appears to be consistent across each imputed data set, showing similar variability. This indicates that the multiple imputation process produces a consistent distribution of imputed values. There are no obvious outliers or differences, which means that the imputation process may have well maintained the overall statistical properties of the bmi data. b) From these histograms

of imputed smoking data, the number of non-smokers and smokers changes slightly in each imputation but overall remains consistent. This also shows that although each imputation may produce slightly different results, the overall data structure and proportions are maintained.




(a) imputed bmi data distribution (b) imputed smoking data distribution

Figure 6: Comparison of 10 imputed data and original data for bmi and smoking variables

In summary, the above analysis shows that the imputed data distribution in the 10 datasets looks more appropriate, so we analyze each dataset separately accounting for confounders, and finally summarize the results (compare with the association between treat and death7day to estimate the causal effect).

# 4 Approaches 1 to estimate causal effects: outcome modeling

Although in the DAG section, I think that sex variable is an irrelevant variable, I still include it in the logistic regression model considering the accuracy of the model.

```
analysis_results <- lapply(1:10, function(i) {
  dataset <- complete(imp_dat, action = i)
  fit <- glm(death7day ~  sex + age + bmi + diabetes + treat + icu + smoking, data =
    dataset, family = "binomial")
  return(fit) })
lapply(analysis_results, summary)
```

Viewing the results for all models, I get the p-values for all variables in 10 regression models in Table 4. We can see that 1) The p-values for sex and smoking variables in 10 models are much larger than 0.05, indicating that they consistently have no significant impact on death7day. 2) Age and icu variables consistently have a significant impact on death7day, as the p-values in 10 models are smaller than 2e-16. 3) The p-values for bmi and diabetes variables have slight changes among the 10 models, however, they still consistently have a significant impact on death7day.

| | p-value | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 | Model 9 | Model 10 |
| Intercept | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 |
| sex | 0.9935 | 0.7515 | 0.9862 | 0.8456 | 0.8640 | 0.8844 | 0.9258 | 0.9501 | 0.9526 | 0.8734 |
| age | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 |
| bmi | 1.60e-13 | 1.08e-09 | <2e-16 | 1.99e-06 | 4.89e-15 | <2e-16 | <2e-16 | <2e-16 | 8.71e-16 | <2e-16 |
| diabetes | 5.25e-07 | 2.34e-07 | 9.17e-07 | 3.02e-07 | 2.22e-07 | 7.37e-07 | 3.59e-07 | 1.59e-07 | 3.50e-07 | 2.65e-07 |
| treat | 0.0007 | 0.0022 | 0.0002 | 0.0016 | 0.0008 | 0.0002 | 0.0007 | 0.0007 | 0.0007 | 0.0003 |
| icu | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 |
| smoking | 0.5210 | 0.5548 | 0.3748 | 0.6233 | 0.9120 | 0.8934 | 0.2657 | 0.9989 | 0.6986 | 0.4716 |

Table 4: P-values of all variables on 10 regression models

To perform a deeper analysis of the exposure treat's coefficients, I collect related data in 10 models, shown in Figure 7, we can see that 1) the estimated coefficient of the treat variable is negative in all 10 models, indicating that treat and the occurrence of death7day are negatively related. 2)The standard error of treat is relatively stable across all models, which indicates that the estimate is highly accurate and does not change much across different imputed data sets. 3) The p-values in all models are small, indicating that the effect of treat is statistically significant, as the p-value is less than 0.01.

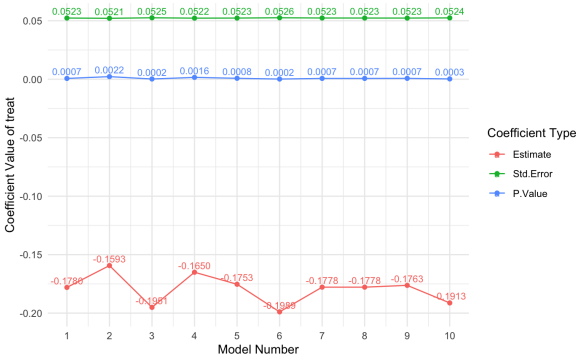



Figure 7: Line chart of the coefficients of treat on 10 models

Overall, the above analysis provides strong evidence that from all imputed datasets there is a negative relationship between treat and death7day, and that this result is statistically significant. This means that treat may be an effective factor in reducing the probability of death7day.

Finally, we have to compare the summaries between the association model before modeling and the pool coefficients on the 10 models to estimate the causal effect. The model summary between treat and death7day before modeling is shown in Figure 8. The model estimates and standard errors from all imputed datasets were combined using the pool function to obtain an overall estimate of these parameters, and the summary is shown in Figure 9.

```r
pooled_results <- pool(analysis_results)
summary(pooled_results)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.27047    0.03054  -8.856   <2e-16 ***
treat1      -0.39824    0.04574  -8.708   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10984  on 8218  degrees of freedom
Residual deviance: 10907  on 8217  degrees of freedom
AIC: 10911

Number of Fisher Scoring iterations: 4
```

Figure 8: The summary of the logistic regression model before matching

| | term | estimate | std.error | statistic | df | p.value |
|---|---|---|---|---|---|---|
| 1 | (Intercept) | -3.130218713 | 0.41161801 | -7.6046689 | 14.91931 | 1.653224e-06 |
| 2 | sex1 | -0.006139792 | 0.05027173 | -0.1221321 | 7504.36392 | 9.027977e-01 |
| 3 | age | 0.021427275 | 0.00149307 | 14.3511540 | 3640.20228 | 1.784743e-45 |
| 4 | bmi | 0.052875610 | 0.01451167 | 3.6436603 | 13.58095 | 2.782078e-03 |
| 5 | diabetes1 | 0.420965205 | 0.08311465 | 5.0648739 | 7370.13744 | 4.185901e-07 |
| 6 | treat1 | -0.179480564 | 0.05396120 | -3.3261038 | 1926.68653 | 8.972340e-04 |
| 7 | icu1 | 3.032205393 | 0.12173975 | 24.9072756 | 8041.78552 | 5.653468e-132 |
| 8 | smoking1 | 0.029239251 | 0.06152299 | 0.4752573 | 547.29856 | 6.347932e-01 |

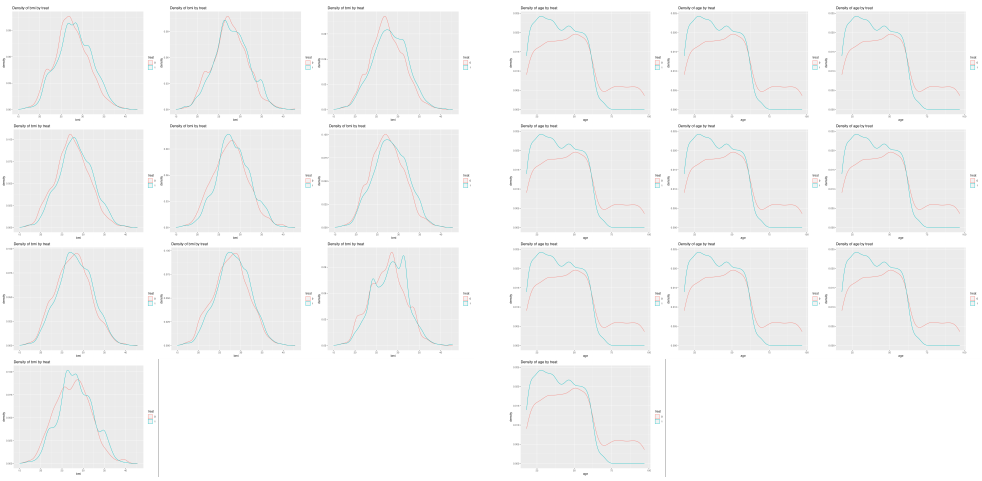
Figure 9: The model coefficients of all covariates after pooling

Compare the summaries between the association model before modeling and the pool coefficients on the 10 models, after controlling all potential confounders, 1) The estimate for treat in the direct model is more negative (from -0.39825 to -0.1795) than in the pooled results. This suggests that the treatment effect might be overestimated when not accounting for missing data and confounders. 2) In both models, the treat variable is statistically significant, indicating that the treatment effect is robust across different methods of handling missing data. 3) The magnitude of the treatment effect (as indicated by the estimate) is lower in the pooled results. This may suggest that after accounting for missing data and confounders through imputation, the effect of treatment is less pronounced than initially observed. 4) The standard errors for treat in the pooled results are larger than in the direct model, which reflects the additional uncertainty introduced by the imputation process.

# 5 Approaches 2 to estimate causal effects: PS matching

## 5.1 observational studies

To perform PS matching, we have to check the overlap first. There must be some overlap in the distributions to avoid having groups with zero/100% chance to be treated (positivity assumption). So I plot the density of bmi and age by treat on 10 imputed datasets to check their overlaps, as shown in Figure 10.



(a) Distribution of bmi by treat on 10 imputed datasets

(b) Distribution of age by treat on 10 imputed datasets

Figure 10: Distribution of a) bmi and b) age by treat on 10 imputed datasets

The overlaps of two variables on treat seem appropriate, which shows that positivity assumption is established. The similarity of the distributions between treated and untreated of bmi variables suggests this may not be an important confounder, and all distributions of age are the same because the age variables have no imputed data, he similarity of the distributions between treated and untreated of age variables suggests this may not be an important confounder either.

## 5.2 PS before matching vs after matching

Under the assumption of no interaction, fit a propensity score model using logistic regression, including all the 5 potential confounders only except the mediater icu (which is mentioned in the DAG section), as there is not really anything lost by adjusting for too many variables, but a lot can be lost by adjusting for too little.

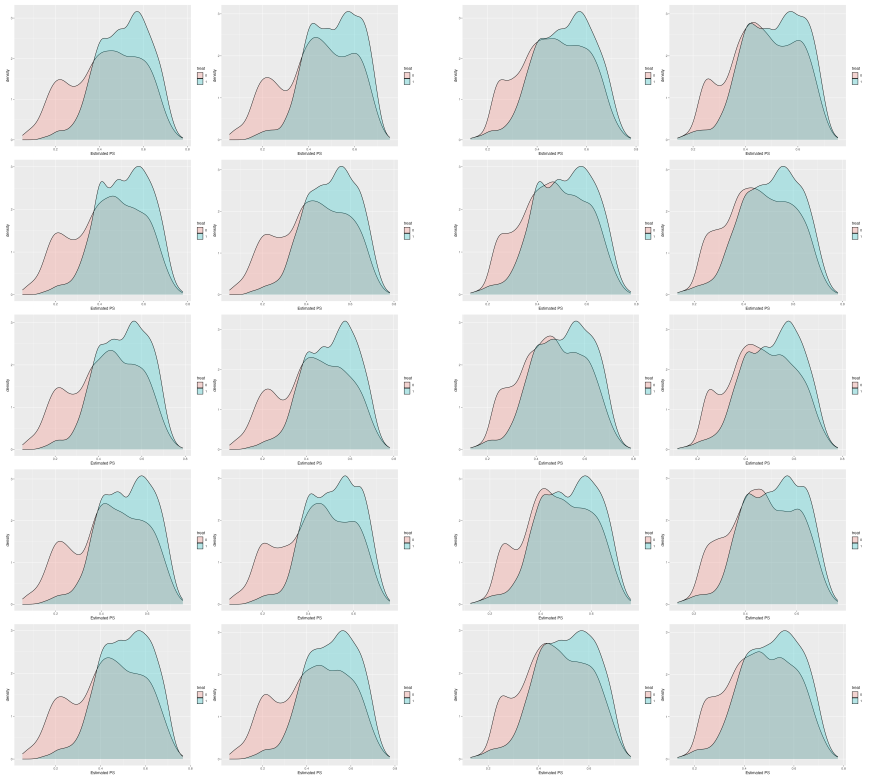
```
ps_plots <- lapply(1:10, function(i) {
  dataset <- complete(imp_dat, action = i)
  # calculate the propensity score
  ps_model <- glm(treat ~ age + bmi + diabetes + smoking + sex,
                  data = dataset, family = "binomial")
  dataset$pscore <- predict(ps_model, type = "response")
  d.ps <- data.frame(dataset, ps = ps_model$fitted)
  ggplot(d.ps, aes(x = ps, fill=treat)) +
  geom_density(alpha = 0.25) +
  xlab("Estimated PS")})
```

Take this regression model to estimate propensity scores, as shown in Figure 11 a), the overlap in propensity scores between the two groups (treated and control) appears to be reasonable, such that we can proceed with matching.

Match the two groups by PS using the function matchit() with nearest neighbor matching, and then plot the PS densities for the two treated and control groups, as shown in Figure 11 b).

```
ps_match_plots <- lapply(1:10, function(i) {
  dataset <- complete(imp_dat, action = i)
  m.out <- matchit(treat ~ age + bmi + diabetes + smoking + sex,
```

```
                  data = dataset, method = 'nearest')
m.dat <- match.data(m.out,  distance = 'pscore')
ggplot(m.dat, aes(x = pscore, fill = treat)) +
  geom_density(alpha = 0.25) +
  xlab("Estimated PS")})
```



(a) PS by treat before matching (b) PS by treat after matching

Figure 11: Comparison of PS by treat before PS matching and after matching

From Figure 11, we can see that the overlaps are much improved after PS matching on 10 imputed datasets, although they are still not perfect.

So now I can estimate the causal effect of the milnepan treatment on mortality rate within 7 days by comparing the model summaries before matching and after matching.

(1) As the imputed data does not include treat variable and death7day variable, so we only need to choose one imputed dataset randomly to estimate the association (unmatched) between treat and death7day for comparison, the model summary is shown in Figure 12.

```
dataset_1 <- complete(imp_dat, action = 1)
model_1 <- glm(death7day ~ treat, data = dataset_1, family = "binomial")
summary(model_1)}
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.27047    0.03054  -8.856   <2e-16 ***
treat1      -0.39824    0.04574  -8.708   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10984  on 8218  degrees of freedom
Residual deviance: 10907  on 8217  degrees of freedom
AIC: 10911

Number of Fisher Scoring iterations: 4
```

Figure 12: The summary of the logistic regression model before matching

(2) Use matched data to model the association between treat and death7day, to estimate the causal effect, subject to assumptions:

```
model_summaries <- list()
for(i in 1:10) {
  dataset_i <- complete(imp_dat, action = i)
  m.out_i <- matchit(treat ~ age + bmi + diabetes + smoking +
                       sex, data = dataset_i, method ='nearest')
  m.dat_i <- match.data(m.out_i,  distance = 'pscore')
  model_i <- glm(death7day ~ treat, data = m.dat_i, family = "binomial")
  model_summaries[[i]] <- summary(model_i)
}
for(i in 1:10) {
  cat(paste("Summary for model", i, ":\n"))
  print(model_summaries[[i]])
  cat("\n\n")
}
```

Simplified summaries for 10 regression models are shown in Table 5.

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 | Model 9 | Model 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Estimate(treat) | -0.3376 | -0.3322 | -0.3450 | -0.3354 | -0.3312 | -0.3504 | -0.3344 | -0.3269 | -0.3408 | -0.3365 |
| Std. error(treat) | 0.04718 | 0.04719 | 0.04717 | 0.04719 | 0.04719 | 0.04716 | 0.04719 | 0.0472 | 0.04718 | 0.04718 |
| p-value(treat) | 8.37e-13 | 1.92e-12 | 2.57e-13 | 1.17e-12 | 2.26e-12 | 1.09e-13 | 1.38e-12 | 4.34e-12 | 5.06e-13 | 9.89e-13 |
| AIC | 10172 | 10169 | 10177 | 10171 | 10168 | 10180 | 10170 | 10165 | 10174 | 10171 |

Table 5: Summaries of 10 regression models after PS matching

Compare the summaries between the direct model before matching and the simplified summaries on 10 regression models after matching, after controlling all potential confounders, 1) The estimate for treat in the direct model is more negative (from <2e-16 to the range [1.09e-13 4.34e-12]) than in the matched results. This suggests that the treatment effect might be a little overestimated when not accounting for missing data and confounders. 2) In both models, the treat variable is statistically significant, indicating that the treatment effect is robust across different methods of handling missing data. 3) The magnitude of the treatment effect is lower in the matched results ( the estimate is from -0.3982 to the range [-0.3504, -0.3269]). This may suggest that after accounting for missing data and confounders through imputation, the effect of treatment is less pronounced than initially observed. 4) The standard errors for treat in the pooled results are a little larger than in the direct model, which reflects the additional uncertainty introduced by the imputation process.

# 6 Discussion

## 6.1 Advantages and disadvantages of outcome modeling

1) Advantages: Account for multiple confounders simultaneously.
Provide a direct estimate of the treatment effect, assuming the model is correctly specified.
Relatively straightforward to implement and interpret.
2) Disadvantages:
Can be sensitive to model misspecification, so if the DAG is not right, it may cause the model to fail.
May not fully account for unobserved confounding, which is not included in this dataset.

## 6.2 Advantages and disadvantages of PS matching

1) Advantages:
Create a matched sample that can be analyzed as if it were a randomized experiment.
Reduce the impact of confounding by balancing covariates across treatment groups.
It's non-parametric, not relying on the assumption of a particular statistical model.
2) Disadvantages:
Only account for observed confounders.
May discard a lot of data if the propensity scores do not overlap well.
Can be computationally intensive and requires careful diagnostics.

### 6.3 Consistency of outcome modeling and PS matching

1) Similarities

Both methods attempt to estimate the causal effect of treatment, rely on observed data, and require assumptions for causal interpretation.

Both methods provide similar estimates of the causal effects, which increases confidence in the findings.

2) Differences

Outcome modeling directly models the relationship between covariates and outcome which includes all observed potential confounders, while PS matching tries to mimic randomization by creating comparable groups, and if the mediater is not recognised correctly, it may import the error.

PS matching discards some unmatched cases, which leads to different sample sizes, and discarded information may be significant to estimate the causal effect, while the outcome modeling keeps all observed information.