ANALYSIS OF HOSPITAL QUALITY OF HYSTERECTOMY SURGERY

Stats 504

(Procedure 13)

Team members: Xiaomeng Du, Ye Guan, Mengjiao Zhang, Rui Zhang

Instructor: Professor Edward Rothman

November 11 2016

Analysis of Hospital Quality of Hysterectomy Surgery

1. Introduction

Surgical procedures quality for different hospitals can vary significantly, so it is of great importance to investigate the influence of various factors and how to evaluate the quality of surgical procedures. In this report, we studies a dataset for hysterectomy surgery from Michigan Surgical Quality Collaborative (MSQC) and our purpose is to learn what factors influence the quality of the procedure. The analysis results should give us insight into hospital rankings based on hysterectomy surgery, reveal potential methods for hospitals to improve the surgery quality, and may also give recommendations to help patients choose proper hospitals before surgery.

One biggest problems when comparing the surgical quality of the hospital is Simpson’s Paradox, what’s true in aggregate may not be true in any subset. The naive way of comparing the performance of the different hospitals would be calculating the rate of complications of each hospital. However due to Simpson's paradox this approach could lead to biased results since we ignore the other important variables that could influence the performance of the hospitals. For example, if a hospital is more likely to be assigned to patients with worse conditions, its patients are more likely to have complications after surgery, even if it has a good surgery quality. To deal with Simpson’s paradox, we use different ways of modeling to adjust for other factors when trying to compare the performance of different hospitals.

We have pre and post surgery data on 23815 patients of 68 hospitals. The approaches we use to study the influence of the covariates to the response are building logistic regression model, counter-factual method, and matching propensity score. The logistic regression model aims to discover what is the general relationship between the predictors and the response, and if we are choosing the correct variables to do our analysis. The counter-factual method and matching propensity score is aiming at comparing the differences between 68 hospitals.

We’ve found out the rankings for all hospitals regarding the surgery quality, potential factors that may affect surgery quality such as surgical approach, being a smoker or not, having cancer or not, and blood loss, etc.

Another issue needs to be addressed is data integrity. The data collection is totally relied on the hospital self-reporting. Since the reimbursement from the insurance company could be influenced by the performance of the hospitals, there is clear motivation to modify the post surgery reports of the patients. Specifically the smaller the hospital is, the fewer the cases, the bigger influence it will be to change one report of a patient since the overall performance will be changed dramatically if there are only a few cases. Thus smaller hospitals will have stronger motivation to modify the record by leaving out the patients that have complications and labeling them as ‘unfollowed’. To explore the data integrity, wefit a logistic regression model with 30 days follow-up variable (*followed_for_30_days*).

2. Material and Methods

2.1 Data

The dataset is about the hysterectomy surgery in 68 hospitals. The dataset comes from Michigan Surgical Quality Collaborative study. We have in total 23815 observations with about 700 variables. These variables include patients’ basic information like health condition, pre surgery and pro surgery information.

2.2 Response and Predictor

The response variable we use to measure the quality of the surgery is a summary of the following complications: *flg_cmp_any* (any morbility), *flg_dead30* (death within 30 days of procedure), *flg_cmp_pneumonia* (pheumonia), *flg_util_readm* (readmission within 30 days), *flg_util_reop* (reoperation within 30 days), *flg_util_transfused* (RBC transfused), *val_los* (length of stay).
Other than *val_los*, all the other variables are coded 0 and 1, indicating no or yes. We code the *val_los* variable into another categorical variable with 0 indicating the length of stay is shorter or equal than 2 days, and 1 indicating the length of stay is longer than 2 days.

val_los (length of stay) is considered as a representative as the cost of surgery. The longer you stay in hospital, the more you cost for this surgery. The other variables are considered as a representative as the health condition after the surgery. If any one of them happens, it means the surgery does not help much to the health of the patient. Therefore, we conclude that our response represent the surgery quality well. Moreover, the proportion of the response being 1 is over 20%, which means this is indeed worth studying.

The covariates that we choose to analyze the response are as following:

Table 1. Covariates selection and manipulation

Variable name	**Meaning**	**Process**	**Comments**
Site_CID_160801	The IDs of hospital		
asa_class_id	ASA class	0: ASA = 1,2 (healthy) 1: ASA = 3,4,5 (sick)	Data with ASA = 7 is considered as missing value.

flg_cmb_cancer	Indicate whether the patients have cancer	0: No 1: Yes	
insurance_payment_type	Type of insurance	1: self-pay, medicaid and uninsured 0: O.W.	The patients with type: self-pay, medicaid or uninsured is considered as the poor.
specmn_weight_grams	Weight of cancer specimen	0: less than 80% quantile 1: O.W.	
e_surgical_approach	Surgical approach	1: Open wound 2: Laparoscopic 3: Robotic 4: Vaginal	The approach that is finally converted to open wound is counted as Open wound.
cervix_removal_method	Cervix removal method	factorize(6 types)	
vaginal_cuff_suture	The condition of cuff suture	factorize(6 types)	
flg_cmb_open_wound	Indicator of open wound	0: No 1: Yes	
flg_cmb_smoker	Indicator of smoker	0: No 1: Yes	
val_age	Age	(18, 40], (40, 55], (55, 95]	
val_surgtime	Surgical time	0: less than 75% quantile 1: O.W.	The value larger than 1000 is considered as outlier
val_bmi	Body mass index	0: less than 35 1: O.W.	
fluid_out_ebl_total	Blood loss during surgery	0: less than 75% quantile 1: O.W	

2.3 Methods

The approaches we use to study the influence of the covariates to the response are building logistic regression model, counterfactual method, and propensity score matching. The logistic regression model aims to discover what is the general relationship between the predictors and the response, and if we are choosing the correct variables to do our analysis. However, to study the effects of hospital, we need to include many interactions in the model. To simplify the model as well as dealt with the Simpson’s Paradox, we apply the counterfactual method and propensity score matching.

2.3.1 Modeling

The response is binary variable, whose expectation is the probability of being 1. Since a probability lies between 0 and 1, we consider a linear regression model for the log odds, therefore we choose to use logistic regression. The logistic regression model use the log odds of having a complication as response, and the exponential of the coefficients could be interpreted as the effects on the odds ratio, that is the change of odds ratio given other factors remaining the same.

We want to learn about how the covariates influence the probability of having a complication, as long as how the hospitals are different from each other. There are mainly two challenges.

First, since there are 68 hospitals, if we include the hospital id and the interaction term with each of the other covariates, the model will be gigantic and hard to interpret. Considering that the number of cases is a good indicator of the hospital size, and the hospital size is usually related to its quality, we summarize the number of cases of each hospital and use a variable, quantile, to indicate the quantile of case number of the hospital. The cut point we choose are 25%, 50%, and 75% quantiles. With only 4 labels, we are able to introduce the interaction term into the model. It turns out that none of the interaction terms are significant.

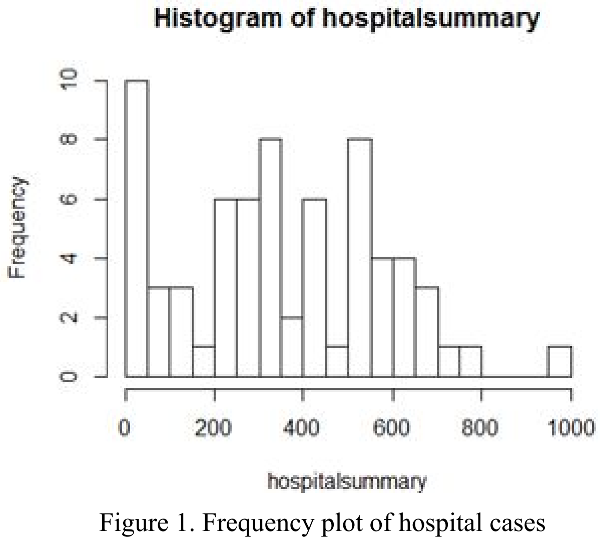


Figure 1. Frequency plot of hospital cases

The second challenge is that some of the variables are clearly correlated, and that creates collinearity within the predictor matrix. For example, the loss of blood is related to the surgery type, the surgery time and patients' BMI and other variables. To deal with this issue, we do regression model of specmn_weight_grams, val_surgtime and fluid_out_ebl_total on other covariates separately and use the residual of each model as our new predictor.

2.3.2 Counterfactual

In order to compare the hospital performance, we come up with another method, counterfactual. The key idea behind counterfactual is that if we assume the performance in different hospitals are similar, the model we construct based on one hospital should be the same as the model constructed on other hospitals. Therefore, we can use the model we develop based on one hospital to predict the performance of other hospitals, then compare the prediction and actual performance. The advantage of the counterfactual method compared with regression is that we are able to construct a simpler model, saving us from including the interaction term between hospital category and other covariates. Also with counterfactual method, we can deal with the Simpson's paradox.

First, we choose one hospital as the baseline to construct the model. The distribution of case numbers in different hospitals are shown below. We choose one hospital with most cases since it will construct a more accurate model. From the table below, the hospital with id _034_ has 854 observations, which is far more than other hospitals.

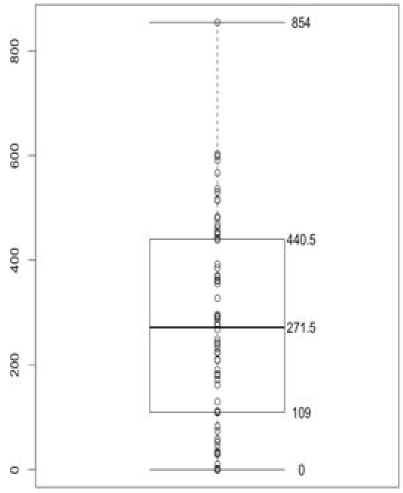


Figure 2. boxplot of hospital cases

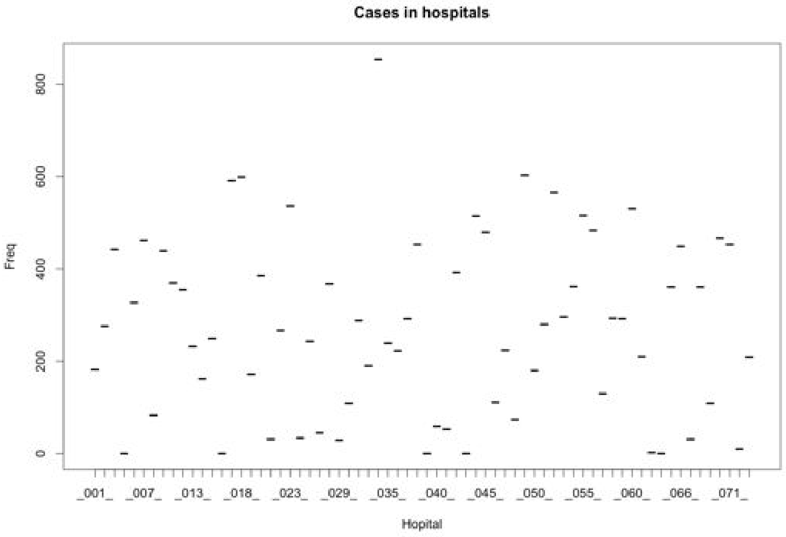


Figure 3. Histogram of hospital cases

Second, we fit the logistic regression model using data in hospital *034*. The predictors and response in this model are same as the modeling method does. The response is the comprehensive indicator of surgery performance, which is a 0-1 categorical variable. The predictor includes health condition of the patients, surgical approach, surgery process and insurance type.

Third, we use the fitted logistic model to predict the surgery performance in all the other hospitals. We can obtain the prediction of probability of complications from the logistic regression model. Then, we can do paired t-test with predicted and actual performance in every hospital, then we can derive a t-statistics from each t-test.

Forth, we can use the t-statistics we derive before to rank the hospitals. Positive t-statistics indicates actual performance is worse than prediction, which indicates performance in that hospital is worse than the baseline hospital. Otherwise, negative t-statistics indicates better performance than the baseline hospital. The smaller the t-statistics is, the better the hospital is.

2.3.3 Propensity Score Matching

Another approach we can use to deal with Simpson paradox is propensity score matching (PSM). PSM is widely used for sample selection by matching and stratification processes. PSM before randomization and seeking the best matching pairs will enhance balance on covariates, reduce sample size requirement as well as improve the hypothesis test power.

To investigate the mortality rate differences between various hospitals, we need to consider the disproportion issues happened in different hospitals. For example, more severely sick patients may select top-ranking hospitals, thus these hospitals would have higher mortality rates compared to those less popular hospitals. Moreover, the distribution balance of the treatment and control groups are also very crucial in experimental designs for solid comparison results.

2.3.3.1 Sample Balancing

If we choose hospital 1 as standard, then we have 123 patients from hospital 1 and 13365 patients from other hospitals after pre-processing our MSQC data. To balance the size of treatment(patients in hospital 1) and control groups(patients in the other hospitals), we increase the treatment group size by resampling with replacement from treatment group until the ratio between the sizes of treatment and control group is one to three.

2.3.3.2 Selection of Predictors and Response

To pair up patients from treatment and control groups, we select several important pre-operation variables to indicate health condition of patients before surgery. The predictors include ASA ranking (asa_class_id), age (val_age), body weight (val_bmi), the insurance type (e_insurance_type), whether the patient is a smoker or has cancer diagnosed before surgery or not (flg_cmb_smoker, flg_cmb_cancer).

We also select several variables to define the outcome of surgical procedures. The response variables include if there are any complications (flg_cmp_any), pneumonia (flg_cmp_pneumonia), readmission (flg_until_readm), reoperation (flg_until_reop), transfusion (flg_until_transfused) or death within 30 days, as well as length of stay in hospital (val_los) after surgery.

2.3.3.3 Matching with propensity scores.

By fitting a logistic regression model with the covariates and response for certain treatment group and control group, we can compute the propensity score(the probability that a patient goes to treatment groups, some hospital). For each patient in the treatment group, we can match, in the control groups, one or more patients with the same propensity score and similar covariates.

2.3.3.4 Rank hospitals by Z-scores

Table 2. Contingency table

Stratum i	Response = 1	Response = 0	Total
Hospital 1	k_i	$K_i - k_i$	K_i
Control group	n - k	$N_i + k_i - n_i - K_i$	$N_i - K_i$
Total	n_i	$N_i - n_i$	N_i

With propensity scores and matched patients, we can stratified the patients (both treatment group and control group) by propensity scores. Within a stratum, we can make a 2×2 contingency table (For example, Table 2). Under the assumption that the hospital is independent to the outcome of the surgery, k_i follows the hypergeometric distribution with mean $n_i\frac{K_i}{N_i}$, and variance $n_i\frac{K_i(N_i-K_i)(N_i-n_i)}{N_i^2(N_i-1)}$. A Z-score for stratum i can be computed as following:

$$Z_i = \frac{k_i - n_i\frac{K_i}{N_i}}{\sqrt{n_i\frac{K_i(N_i-K_i)(N_i-n_i)}{N_i^{2}(N_i-1)}}}$$

Z_i can show that within stratum i, comparing to the control group, how the treatment group performs. In table 2, if Z-score is large, it means comparing to Control group, the patients who go to hospital 1 is more likely to result in bad surgery outcome (respond = 1). On the contrary, if Z-score is small, it means the patients who go to hospital 1 is more likely to result in good surgery outcome (respond = 0). Combining all the strata, assuming the independence between different stratum, we can have an overall Z-score for one hospital as following:

$$Z_{overall} = \frac{\sum_{i=1}^{S} k_i - n_i \frac{K_i}{N_i}}{\sqrt{\sum_{i=1}^{S} n_i \frac{K_i (N_i - K_i)(N_i - n_i)}{N_i^{2} (N_i - 1)}}}$$

$Z_{overall}$ of a hospital shows the overall quality of the hospital, comparing to the control groups. With this overall Z-score, we can rank all hospitals.

3. Results

3.1 Modeling

3.1.1 Interpretation of the model

According to the modeling method described above, we can fit the logistic model. The estimates of the covariate coefficient are shown in table 3. From the model, we can find that

a) All the coefficients of quantile groups are significant and increasingly positive, showing that the bigger the size of the hospital, the more likely the patient will have a complication.

b) The ASA score is significant, showing that the sicker the patient is the more likely they will get a complication.

c) Similarly, having cancer will increase the probability of the patient has a complication.

d) If the patient has Medicaid or self-paid insurance or uninsured, the patient is more likely to have a complication. This is expected because these patients are more likely to be poor people that avoids going to the hospital until they are very sick due to the large medical care bills.

e) The surgery type is significant, and the larger wound the surgery brings, the more likely the patient will have a complication. Also, the blood loss residual and surgery time are significant, meaning the more blood the patient loss and the longer the surgery takes, the more likely the patient is going to have a complication. These conclusions together suggest that potential improvement of the surgery can be made by controlling the blood loss and surgery time as long as using the minimum invasive surgery type.

f) The age groups variable is significant, more interestingly, the patient of age between (40,55), (0,40) and (55,95) has increasing probability to have a complication given all the other conditions are the same. The oldest patients are most likely to have a complication, but the middle aged people are safer than the youngest people. Further medical study could be done to analysis why this is true. One possible explanation is that people who have to have hysterectomy are more likely to have a very serious disease, while the middle aged people are more likely to have a common illness due to their age.

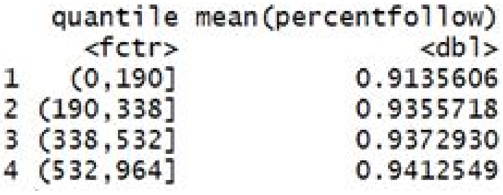
Table 3. Covariate coefficient of logistic model

Covariate	Estimate	Exponential Coefficient(e^{β})	Standard Error	P-value	
Intercept	-3.8494	0.0213	0.2209	<2e-16	***
quantile(190, 338]	0.4111	1.5085	0.1614	0.010845	*
quantile(338, 532]	0.6515	1.9184	0.1532	2.12e-05	***
quantile(532, 964]	0.6971	2.0080	0.1519	4.47e-06	***
asa_class_id1	0.6326	1.8826	0.0581	<2e-16	***
flg_cmb_cancer1	0.6433	1.9028	0.3646	0.0776	.
insurance_payment_type1	0.2774	1.3197	0.0672	3.70e-05	***
e_surgical_approachOpen	2.4685	11.8049	0.1447	<2e-16	***
e_surgical_approachRobo	-0.2590	0.7718	0.1418	0.067766	.
e_surgical_approachVagi	-0.0411	0.9597	0.1476	0.780687	
val_age(40, 50]	-0.2188	0.8035	0.0662	0.000955	***
val_age(55, 95]	0.2783	1.3208	0.0760	0.000249	***
val_surgtime	0.0079	1.0079	0.000424	<2e-16	***
bloodres	0.0011	1.0011	0.00012	<2e-16	***

3.1.2 Data Integrity Analysis

According to the logistic regression model, the bigger the hospital is, the more likely their patients will have a complication given other conditions are the same. It is hard to interpret this observation, since bigger hospitals generally tends to have better quality of medical care. Thus further investigation is necessary to analyze the difference between the hospitals.

Other than analysis on models, it is necessary to question the reliability of the data itself since the data collection is dependent on the staff of the hospitals. If we compare the average follow-up-for-30-days rate between the four different sizes of hospitals, we have the following result:



```
    quantile mean(percentfollow)
      <fctr>               <dbl>
1    (0,190]           0.9135606
2  (190,338]           0.9355718
3  (338,532]           0.9372930
4  (532,964]           0.9412549
```

Figure 4. The relationship between 30 days follow up and hospital sizes

The following up rate is clearly increasing with the size of the hospital. This means the smaller hospitals is more likely to drop patients for further follow ups. This is suspicious and makes us wonder if they intend to drop the patients with complications to lower their risks.

We build a model with response being the follow-up for 30 days indicator (*followed_for_30_days*). We use only the data of the patients who have no complications, since this part of the patients are the ones that we suspect that are left out of the study to lower the risk of the hospital. Also, the patients with complications could drop out of following-ups before they reach 30 days because of deaths. If this model shows similar effect with the model we built to explain the response of complications, we know that patients with higher risks of compilations are more likely to be left out of the study; hence the data we collected may be biased.

Followings are the key results of the model (full results are in Appendix):

Table 4. Model results for covariates

Covariate	Estimate	Exponential Coefficient	Standard Error	P-value	
bmires	-9.683e-03	0.9904	4.914e-03	0.04877	*
bloodres	-1.433e-04	0.9999	7.127e-05	0.04442	*
flg_cmb_smoker1	-3.461e-01	0.7074	8.310e-2	3.12e-05	***

It seems that the patients have bigger BMI, lost more blood during surgery, and are smokers are more likely to be dropped out of the study, and the effect is significant. This is consistent with what we found in the previous model where the response is the indicator of the complications.

3.2 Counterfactual

3.2.1 Logistic regression

First, we compare the covariate distribution in the baseline hospital *034* with all the hospitals. Some of the categorical variables are listed below. We can find that cancer ratio in hospital *034* has greatly difference from all hospitals, ratio of cancer patients is low in the hospital *034*. Surgical approach varies in hospital *034* and all hospitals as well. Number of cases using laparoscopic and vaginal surgical approaches are relatively low in hospital *034*. Feature of patients going to hospital *034* may be slightly different from patients of all hospitals. Therefore, the full model including all the hospitals may produce different models as single hospital does, which indicates possibility of Simpson's paradox.

Using the data from baseline hospital *034*, we fit a logistic regression model. We find that the covariate *asa_class_id, age of the patients* and *blood lost* are very significant, *smoke* is also significant. It shows that the patients' health condition and age has a great effect of the surgery performance. It is relatively explicit since healthy and young patients will have high possibilities of success.

3.2.2 Hospital rank

Based on the t-statistics, we can get the rank of the hospitals in the table below. In order to test the robustness of the model, we also change the baseline of the hospital. We choose hospital *002* as the baseline and do the counterfactual analysis again. The rank comparison between these two different baseline are shown below. We find that the ranks are generally the same using different baselines. It gives us strong evidence that the counterfactual method is robust.

Table 5. Rankings and statistics for all hospitals

rank	hospital	statistics	rank	hospital	statistics	rank	hospital	statistics
1	_024_	-0.82	22	_027_	2.44	43	_058_	5.55
2	_067_	-0.78	23	_001_	2.58	44	_071_	5.66
3	_072_	-0.74	24	_040_	2.76	45	_012_	5.89
4	_061_	-0.63	25	_037_	2.99	46	_003_	5.91
5	_041_	-0.50	26	_054_	3.09	47	_032_	6.58

6	_002_	0.00	27	_059_	3.60	48	_014_	6.64
7	_069_	0.10	28	_009_	3.65	49	_028_	7.39
8	_048_	0.13	29	_070_	3.65	50	_038_	7.77
9	_022_	0.19	30	_033_	3.87	51	_051_	7.77
10	_008_	0.57	31	_010_	4.09	52	_035_	7.98
11	_042_	0.82	32	_005_	4.37	53	_017_	8.00
12	_013_	0.83	33	_044_	4.52	54	_046_	8.23
13	_030_	0.88	34	_007_	4.65	55	_065_	8.76
14	_073_	0.95	35	_050_	4.66	56	_060_	9.13
15	_047_	0.97	36	_056_	4.95	57	_066_	9.14
16	_057_	1.05	37	_053_	4.99	58	_018_	9.96
17	_019_	1.25	38	_052_	5.05	59	_020_	11.47
18	_029_	1.58	39	_045_	5.08	60	_034_	15.56
19	_021_	1.63	40	_049_	5.15	61	_023_	16.11
20	_025_	1.70	41	_015_	5.34	62	_068_	16.83
21	_055_	1.83	42	_036_	5.44	63	_062_	46.08

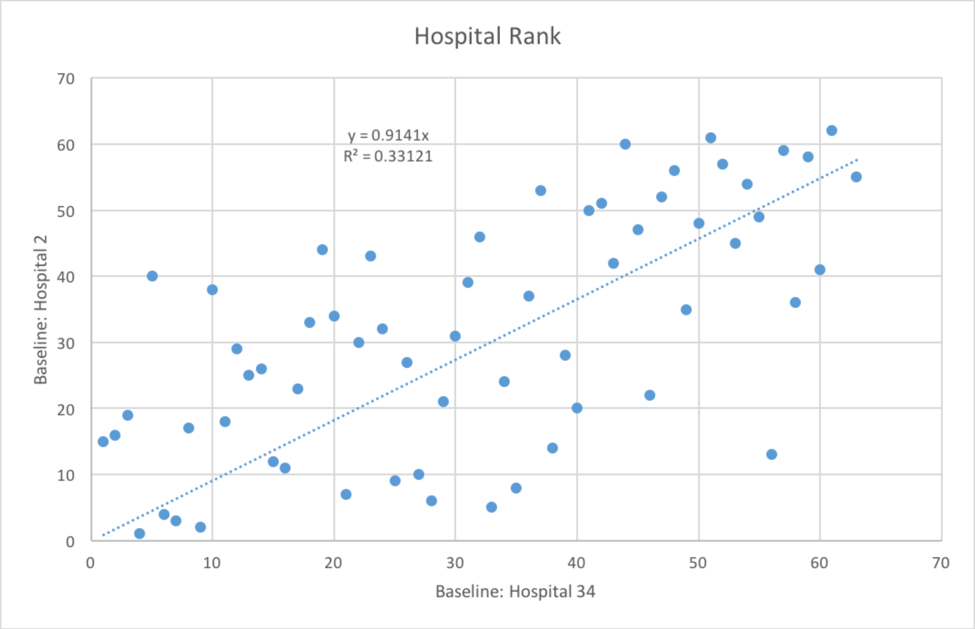


Figure 5. Comparison between the hospital rankings with two different baselines

3.3 Matching

3.3.1 The relationship between hospital and responds

In this part we are trying to find the connection between hospitals and responds, i.e. the quality of the surgery. Following, I will use hospital 18 (with 371 patients) as an example.

First, let's see the how well our propensity score matching deals with the Simpson's Paradox. To see that, we can compare the distributions of the covariates in the pre-matched samples and the ones in the matched stratum.

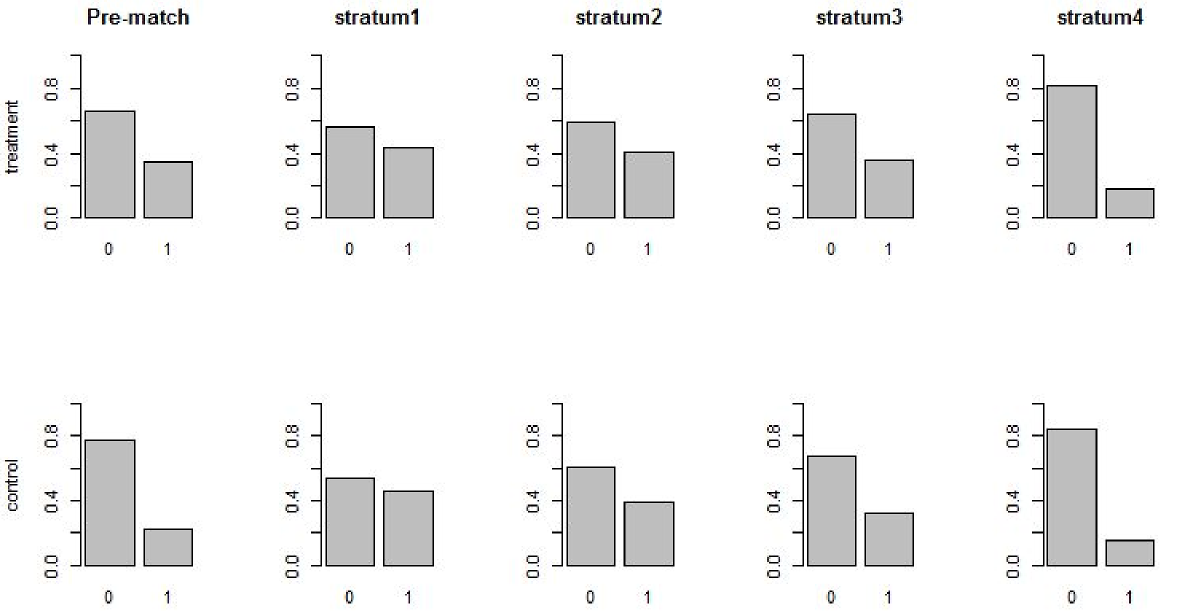


Figure 6. ASA (asa_class_id) distributions before and after matching

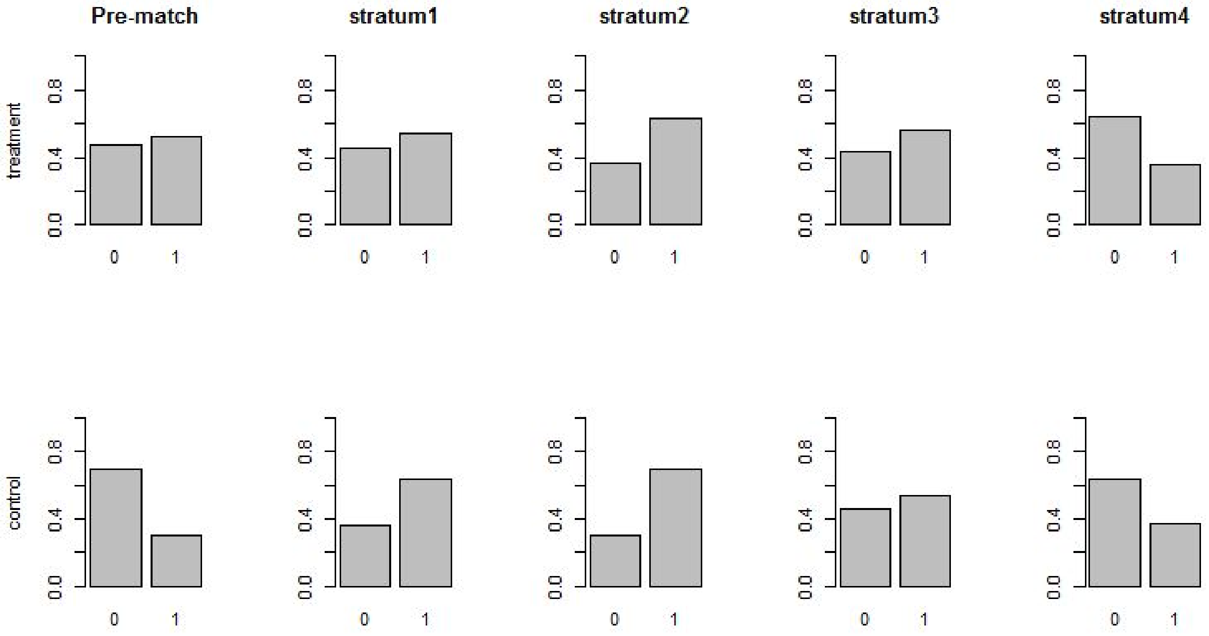


Figure 7. Blood lost (fluid_out_ebl_total) distributions before and after matching

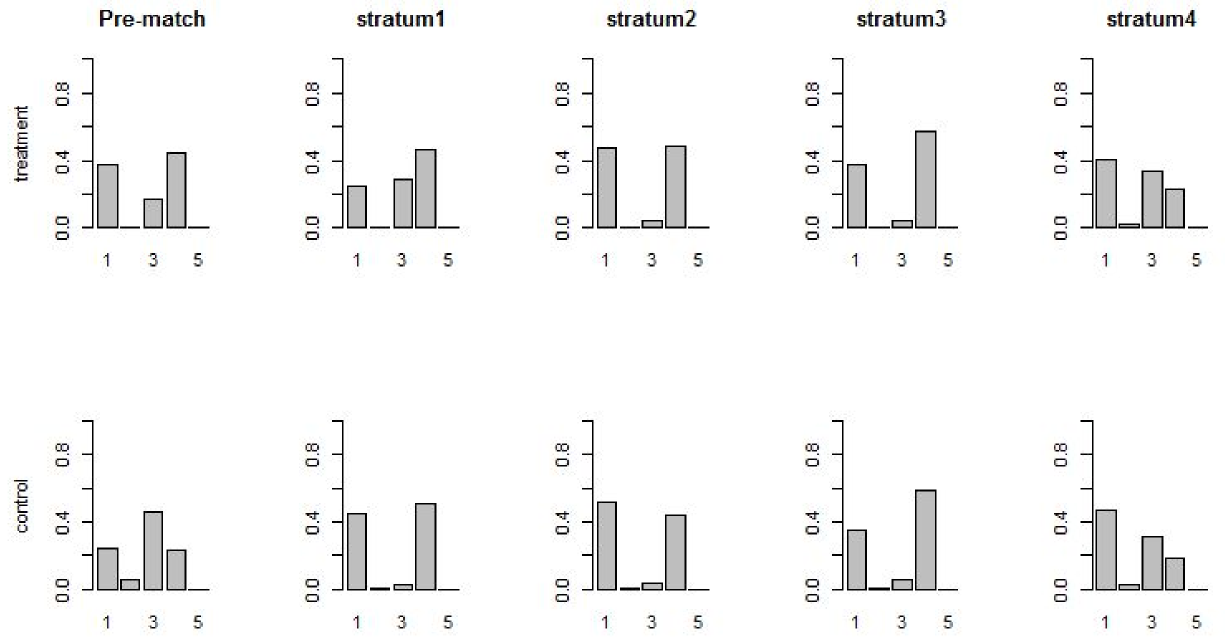


Figure 8.Surgical approach (e_surgical_approach) distributions before and after matching

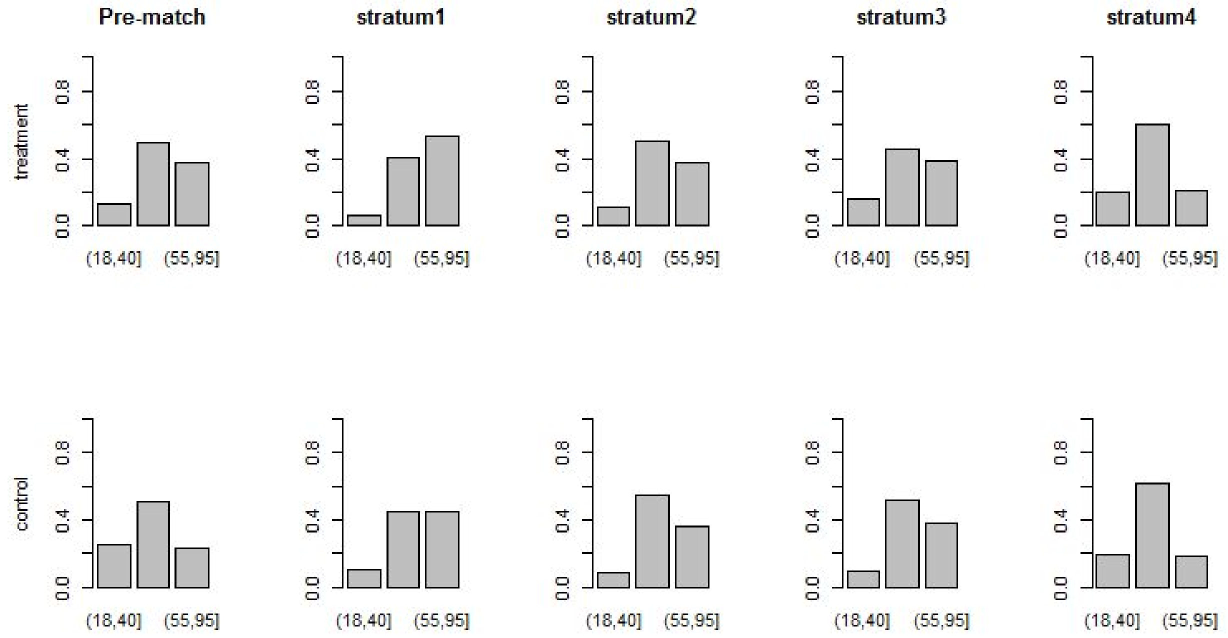
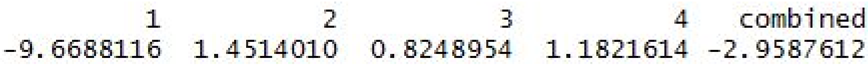


Figure 9. Age (val_age) distributions before and after matching

We can see, in the pre-matched samples, the distribution of one covariates in the treatment group is quite different from the control groups. After propensity score matching, in different stratum, the distribution is quite different, which is reasonable because the patients in different stratum should have different conditions. By comparing the distributions between treatment group and control group within the same stratum, we can see they are very much the same. Therefore, we can conclude that, within the stratum, we are indeed comparing apples to apples and Simpson's Paradox is being successfully treated.

Second, we can have a feel on the surgical operation quality by seeing the Z-scores of them.

Table 6. Z-scores for surgical operation quality



```
         1          2          3          4   combined
-9.6688116  1.4514010  0.8248954  1.1821614 -2.9587612
```

Group 1: propensity score larger than 75% quantile of all propensity score
Group 2: propensity score larger than 50% quantile and less than 75% quantile of all propensity score
Group 3: propensity score larger than 25% quantile and less than 50% quantile of all propensity score
Group 4: propensity score less than 25% quantile of all propensity score

We can see, in group one hospital 18 has significantly less "response = 0" than the control group. Although in the other groups the statistics show that hospital 18 has more "response = 1" (not significantly), combining all the strata, the hospital 18 is still significantly better than the control group.

With similar analysis, we can compare the results of all hospitals. The complete rankings are as below.

Table 7. Hospital rankings with Z-scores

rank	hospital	statistics	rank	hospital	statistics	rank	hospital	statistics
1	_072_	-38.04	22	_008_	-7.74	43	_012_	1.37
2	_041_	-31.74	23	_047_	-6.58	44	_060_	1.66
3	_061_	-23.06	24	_070_	-6.18	45	_053_	2.06
4	_067_	-22.96	25	_071_	-5.30	46	_029_	2.13
5	_030_	-22.91	26	_059_	-4.04	47	_007_	4.55
6	_022_	-22.53	27	_009_	-3.48	48	_036_	5.46
7	_024_	-21.12	28	_010_	-3.13	49	_015_	5.56

8	_027_	-20.19	29	_018_	-2.96	50	_051_	5.76
9	_069_	-17.78	30	_054_	-2.95	51	_014_	5.93
10	_048_	-17.60	31	_049_	-1.70	52	_065_	6.12
11	_002_	-17.09	32	_032_	-1.60	53	_057_	6.55
12	_001_	-16.89	33	_003_	-1.60	54	_038_	6.77
13	_050_	-15.98	34	_021_	-1.44	55	_034_	7.31
14	_073_	-15.65	35	_019_	-0.93	56	_017_	7.78
15	_013_	-13.36	36	_058_	-0.57	57	_020_	7.91
16	_040_	-13.20	37	_028_	0.51	58	_035_	9.45
17	_044_	-11.50	38	_052_	0.52	59	_066_	10.05
18	_055_	-10.69	39	_056_	0.70	60	_023_	14.82
19	_033_	-9.81	40	_025_	0.74	61	_068_	23.05
20	_037_	-8.05	41	_005_	0.81	62	_046_	25.95
21	_042_	-8.03	42	_045_	1.04			

3.3.2 Relationship between surgical approach preference and the ranking of a hospital

Other than revealing the ranking of hospitals, PSM method can also provide insights into the effect of surgical approach on quality of surgical procedures. Surgical approaches are of great importance because patients may have very diverse tolerance responses to different approaches, and then result in various outcomes including length of stay in hospital, mortality rate, etc. We classified the surgical approaches involved in our study into five groups: 1. Open; 2. Laparoscopic; 3. Robotic; 4. Vaginal; 5. Other.

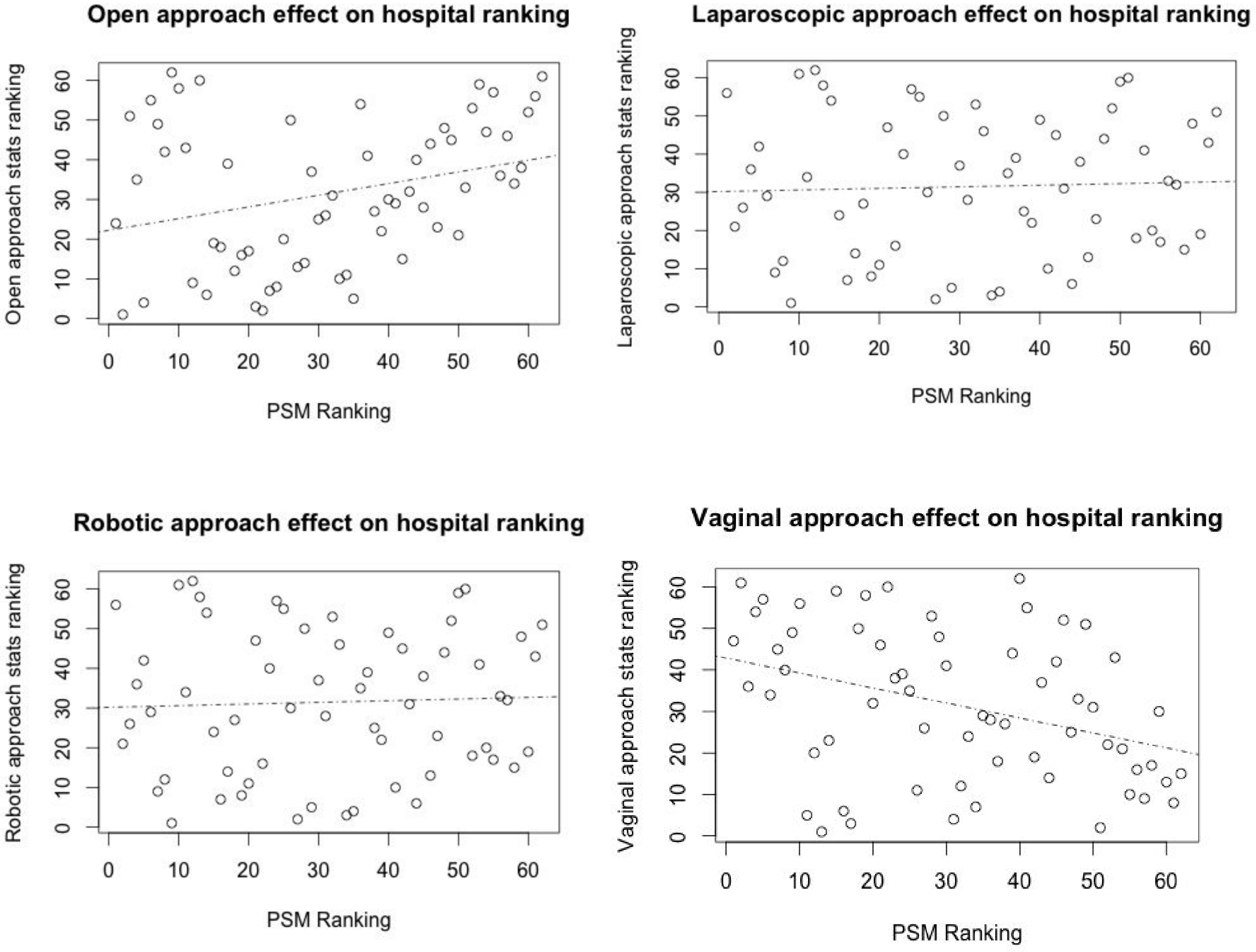
Since we only have one patient has “Other” surgical approach so we discard this approach and only take insights into the other four kinds of surgical approaches. We’ve applied PSM method and conclude that the ranking for surgical approaches are:

Vaginal > Robotic > Laparoscopic > Open.

Open approach has the highest risk for patients and should be avoided during surgery. The priority approach is vaginal method. However, if a hospital is lack of funding and not well equipped with advanced tools, then this hospital may have a relatively worse procedure quality compared to other hospitals with robotics, etc.

We've also searched for some information regarding surgical approach selection online and it seems that hospitals indeed prefer to use vaginal surgical approach other than other methods. If vaginal approach is not applicable, the alternative approaches are laparoscopic or robotic assisted laparoscopic methods.

Z scores are calculated to evaluate the surgical approach preferences for different hospitals. Higher z scores mean bigger chances to apply the specific surgical approach. We define high z scores to have lower rankings of statistics.



Open approach effect on hospital ranking

Laparoscopic approach effect on hospital ranking

Robotic approach effect on hospital ranking

Vaginal approach effect on hospital ranking

Figure 10. Surgical approach effect on hospital ranking

From the above plots, we can see that those hospitals have high preference for open usually have low PSM ranking, while the hospitals that prefer vaginal approach will have high PSM ranking. The relationships between robotics/laparoscopic approaches and PSM rankings are not very clear.

Moreover, it is interesting to see that we are capable to gain some knowledge of hospital ranking based on the surgical approach distribution (Figure 11). From the following distributions, we know that Hospital 046 has a lot of surgeries using open approach, Hospital 041 and 030 have a significant amount of surgeries using vaginal approach and robotic approach surgeries occupy the majority of surgeries in Hospital 002. Since open approach is a bad approach compared to vaginal and robotic approaches, we suspect that Hospital 046 should have bad quality of surgery while Hospital 041, Hospital 030, Hospital 002 should have better quality of surgery thus they should rank higher.

From the ranking result we got from PSM, we know that the rankings of Hospital 046, Hospital 041, Hospital 030 and Hospital 002 are 62nd, 2nd, 5th, and 11th, respectively. The rankings are consistent with our guess. However, this guess has Simpson's paradox issue due to the reason that there are so many factors that influence the quality of surgical procedures. To avoid this problem, we explored the effect of surgical approach on the ranking for one hospital further using PSM method. The results are shown in Table 8.

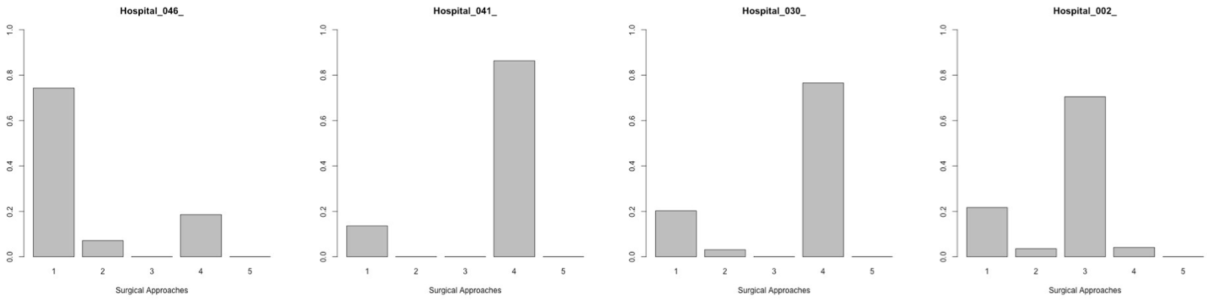


Figure 11. Surgical approach distribution for different hospitals

Table 8. Z scores for four different hospitals

	046	**041**	**030**	**002**
Open	26.9758	-37.2841	-20.2269	3.7894
Laparoscopic	9.1636	-9.6421	1.0598	-2.1772
Robotic	-32.9425	-18.5219	-27.4674	14.8357
Vaginal	-11.4906	45.1793	33.3499	-23.1942
Ranking	62	2	5	11

Table 8 indicates the z scores for different approaches of each hospital. For Hospital 046, we can see that open surgical approach is significant regarding to other methods and its ranking is very low, which is 62nd. Hospital 041 and 030 both have significant statistics for vaginal approach, which is also the approach with top ranking, and the rankings for these two hospitals are 2nd and 5th. Hospital 002 has a stronger preference for robotic approach, which is a relatively better approach compared to laparoscopic and open methods, but not as good as vaginal approach. The ranking for Hospital 002 is 11th.

To summarize, selection of surgical approach has an important impact on surgical procedure quality. Hospitals with more surgeries using open approach tend to have worse surgical procedure quality while the ones with more surgeries using vaginal or robotic approaches may have better surgical procedure quality.

4. Conclusion

In this project, we successfully apply three methods to study the influential factors on the surgery quality. Using modeling method, with the regression coefficients, we can quantify how the important factors change will affect. For example, blood loss is negative correlated with surgery performance. Using open wound methods will largely lower the quality of the surgery. Besides, we investigated the relationship between the follow-up and the size of the hospital and found out that the patient who smokes and with large BMI and large blood loss are easier to be dropped.

Applying counterfactual method, we build a simpler model for our base line hospital. By comparing other hospitals with the "baseline", we can rank all the hospitals. We also tried changing different baseline hospital and found out this method is stable.

Propensity score matching method are used to deal with Simpson's paradox. By showing the distribution of the covariates in each strata, we can see the patients in treatment group are successfully matched the patients with the similar covariates in the control group. By computing Z-scores, we can quantify the difference between the quality of each hospitals and the control group (all hospitals). We've also studied the relationship between hospital and surgical approach to see if there is any approach preference for hospitals. From our results, we got the approach distributions for one hospital, thus we can see the connection between surgical approach and hospital ranking.

The following figure shows that the rankings from counterfactual and the ones from propensity score matchings are very similar, therefore robust.

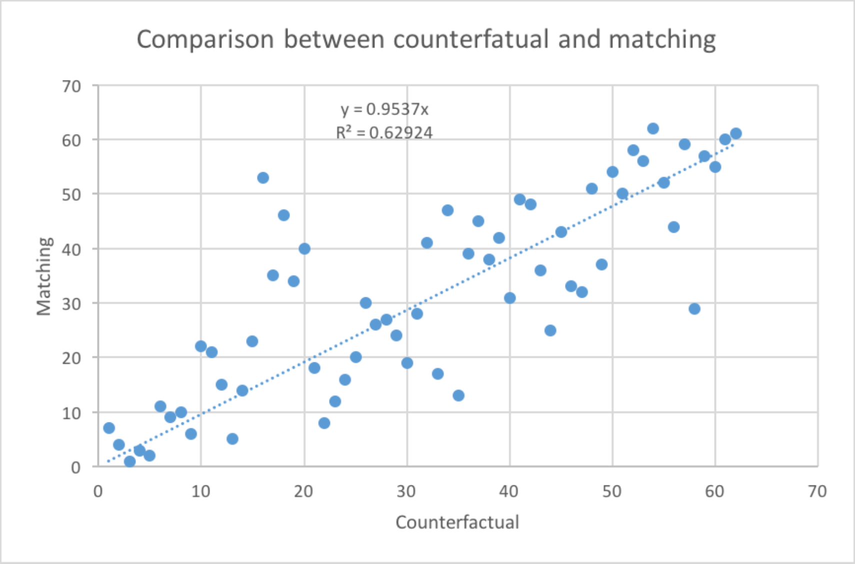
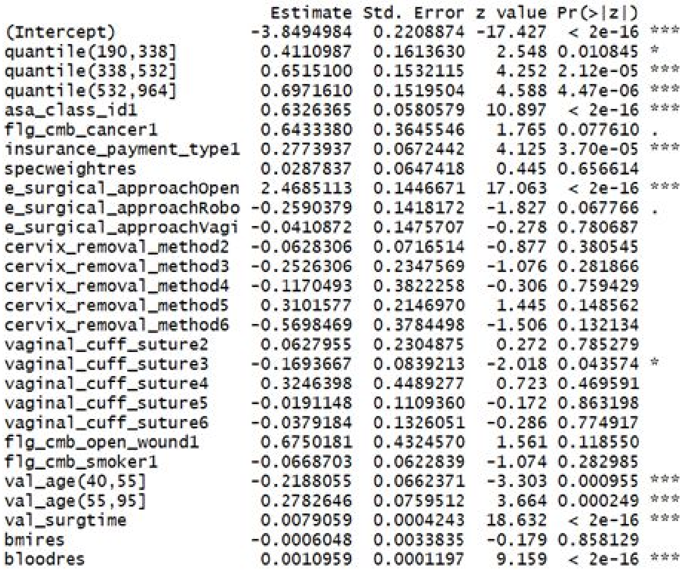


Figure 12. Comparison between the hospital rankings from counter-factual and PSM approaches

Combining all the results, we conclude that hospitals 002, 024, 022, 041, 048, 061, 067, 069 performs better hysterectomy surgeries. As for the surgery approach, vaginal is the best approach and open wound is the worst one. Hence we recommend hospital 048, 041, 067 which prefer vaginal approach and perform better surgery. The worst hospital, we think, is hospital 046, which prefers open wound and perform bad surgery.

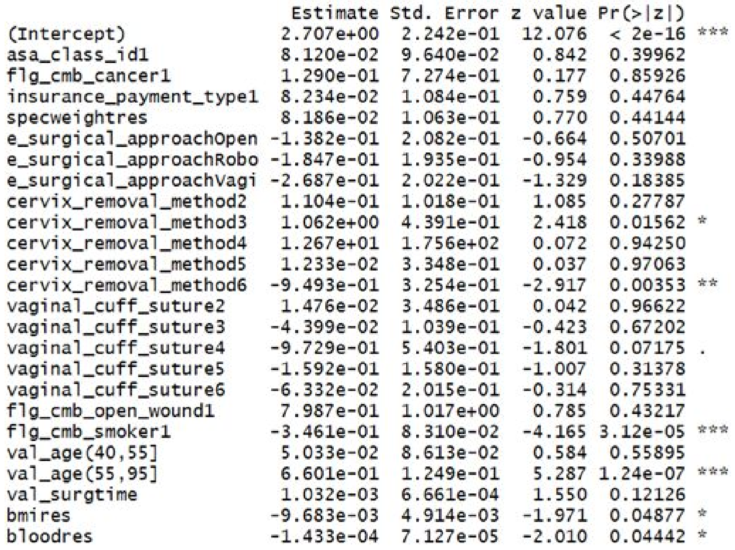
5. Appendix

A1



```
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -3.8494984  0.2208874 -17.427  < 2e-16 ***
quantile(190,338]         0.4110987  0.1613630   2.548 0.010845 *  
quantile(338,532]         0.6515100  0.1532115   4.252 2.12e-05 ***
quantile(532,964]         0.6971610  0.1519504   4.588 4.47e-06 ***
asa_class_id1             0.6326365  0.0580579  10.897  < 2e-16 ***
flg_cmb_cancer1           0.6433380  0.3645546   1.765 0.077610 .  
insurance_payment_type1   0.2773937  0.0672442   4.125 3.70e-05 ***
specweightres             0.0287837  0.0647418   0.445 0.656614    
e_surgical_approachOpen   2.4685113  0.1446671  17.063  < 2e-16 ***
e_surgical_approachRobo  -0.2590379  0.1418172  -1.827 0.067766 .  
e_surgical_approachVagi  -0.0410872  0.1475707  -0.278 0.780687    
cervix_removal_method2   -0.0628306  0.0716514  -0.877 0.380545    
cervix_removal_method3   -0.2526306  0.2347569  -1.076 0.281866    
cervix_removal_method4   -0.1170493  0.3822258  -0.306 0.759429    
cervix_removal_method5    0.3101577  0.2146970   1.445 0.148562    
cervix_removal_method6   -0.5698469  0.3784498  -1.506 0.132134    
vaginal_cuff_suture2      0.0627955  0.2304875   0.272 0.785279    
vaginal_cuff_suture3     -0.1693667  0.0839213  -2.018 0.043574 *  
vaginal_cuff_suture4      0.3246398  0.4489277   0.723 0.469591    
vaginal_cuff_suture5     -0.0191148  0.1109360  -0.172 0.863198    
vaginal_cuff_suture6     -0.0379184  0.1326051  -0.286 0.774917    
flg_cmb_open_wound1       0.6750181  0.4324570   1.561 0.118550    
flg_cmb_smoker1          -0.0668703  0.0622839  -1.074 0.282985    
val_age(40,55]           -0.2188055  0.0662371  -3.303 0.000955 ***
val_age(55,95]            0.2782646  0.0759512   3.664 0.000249 ***
val_surgtime              0.0079059  0.0004243  18.632  < 2e-16 ***
bmires                   -0.0006048  0.0033835  -0.179 0.858129    
bloodres                  0.0010959  0.0001197   9.159  < 2e-16 ***
```

A2



```
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)              2.707e+00  2.242e-01  12.076  < 2e-16 ***
asa_class_id1            8.120e-02  9.640e-02   0.842  0.39962    
flg_cmb_cancer1          1.290e-01  7.274e-01   0.177  0.85926    
insurance_payment_type1  8.234e-02  1.084e-01   0.759  0.44764    
specweightres            8.186e-02  1.063e-01   0.770  0.44144    
e_surgical_approachOpen -1.382e-01  2.082e-01  -0.664  0.50701    
e_surgical_approachRobo -1.847e-01  1.935e-01  -0.954  0.33988    
e_surgical_approachVagi -2.687e-01  2.022e-01  -1.329  0.18385    
cervix_removal_method2   1.104e-01  1.018e-01   1.085  0.27787    
cervix_removal_method3   1.062e+00  4.391e-01   2.418  0.01562 *  
cervix_removal_method4   1.267e+01  1.756e+02   0.072  0.94250    
cervix_removal_method5   1.233e-02  3.348e-01   0.037  0.97063    
cervix_removal_method6  -9.493e-01  3.254e-01  -2.917  0.00353 ** 
vaginal_cuff_suture2     1.476e-02  3.486e-01   0.042  0.96622    
vaginal_cuff_suture3    -4.399e-02  1.039e-01  -0.423  0.67202    
vaginal_cuff_suture4    -9.729e-01  5.403e-01  -1.801  0.07175 .  
vaginal_cuff_suture5    -1.592e-01  1.580e-01  -1.007  0.31378    
vaginal_cuff_suture6    -6.332e-02  2.015e-01  -0.314  0.75331    
flg_cmb_open_wound1      7.987e-01  1.017e+00   0.785  0.43217    
flg_cmb_smoker1         -3.461e-01  8.310e-02  -4.165 3.12e-05 ***
val_age(40,55]           5.033e-02  8.613e-02   0.584  0.55895    
val_age(55,95]           6.601e-01  1.249e-01   5.287 1.24e-07 ***
val_surgtime             1.032e-03  6.661e-04   1.550  0.12126    
bmires                  -9.683e-03  4.914e-03  -1.971  0.04877 *  
bloodres                -1.433e-04  7.127e-05  -2.010  0.04442 *  
```

A3

```
setwd('D:\\my_courses\\STATS 504')
library(readstata13)
library(dplyr)
data=read.dta13('procedure13.dta')
summary(data[,c("flg_cmp_any","flg_dead30","flg_cmp_pneumonia","val_los","flg_util_readm","flg_util_reop","flg_util_transfused")])
response=data[,c("flg_cmp_any","flg_dead30","flg_cmp_pneumonia","val_los","flg_util_readm","flg_util_reop","flg_util_transfused")]
response$val_los=as.numeric(response$val_los>2)
cumu = rowSums(response[,],na.rm=T)



# covariates
covariate =
data[,c("site_cid_160801","asa_class_id","flg_cmb_cancer","insurance_payment_type","specmn_weight_grams","e_surgical_approach",
"cervix_removal_method","vaginal_cuff_suture",
           "flg_cmb_open_wound","flg_cmb_smoker","val_age","val_surgtime","val_bmi","fluid_out_ebl_total")]


# specmn weight
covariate$specmn_weight_grams =
as.numeric(covariate$specmn_weight_grams>quantile(data$specmn_weight_grams,na.rm=TRUE,0.8))

#cervix removal method
covariate$cervix_removal_method = as.factor(covariate$cervix_removal_method)

# vaginal_cuff_suture
covariate$vaginal_cuff_suture = as.factor(covariate$vaginal_cuff_suture)
# flg_cmb_open_wound
covariate$flg_cmb_open_wound = as.factor(covariate$flg_cmb_open_wound)
# flg_cmb_smoker
covariate$flg_cmb_smoker = as.factor(covariate$flg_cmb_smoker)

# age
```

```
covariate$val_age = cut(covariate$val_age,breaks = c(0,40,55,max(covariate$val_age)))

#val_surgtime
covariate$val_surgtime[covariate$val_surgtime>1000] = NA


#hospital
colnames(covariate)[1] = 'site_cid'
covariate$site_cid = as.factor(covariate$site_cid)

#asa_class_id: 0,1-category1; 2,3,4,5,6,7-category2
covariate$asa_class_id[covariate$asa_class_id==7]=NA
covariate$asa_class_id=as.factor(as.numeric(covariate$asa_class_id>=3))
summary(covariate$asa_class_id)
#flg_cmb_cancer:0,1,NA
covariate$flg_cmb_cancer = as.factor(covariate$flg_cmb_cancer)

#insurance_payment_type: category 1: 3 - Medicaid/8 - No Insurance; category 0: otherwise
covariate$insurance_payment_type          =          as.factor(as.numeric(covariate$insurance_payment_type==3          |
covariate$insurance_payment_type==8 | covariate$insurance_payment_type==9))

# e_surgical_approach: easy-2,3,7 medium-5,(9,10),11,12,13,14,17,18,21,22,23,24 hard-1,4,6,(8)
# 1 - Open
# 2 - Laparoscopic
# 3 - Laparoscopic, Hand-Assisted
# 4 - Laparoscopic, Converted to Open
# 5 - Robotic
# 6 - Robotic, Converted to Open
# 7 - Laparoscopic, Single Port (SIL)
# 8 - Open Endo
# 9 - Percutaneous Endo
# 10 - Percutaneous Endo, converted to Open Endo
# 11 - Vaginal
# 12 - Vaginal, lap-assisted
# 13 - Other (e.g. anal, transanal, transoral, etc.)

# > summary(covariate$e_surgical_approach)
# 1    2    3    4    5    6    7   11   12   13   14   17   18   21   22   23   24
# 5729 2143   40  434 9938  178   37 2857 2396    2    1    5    1    1    5   47    1

covariate[which(covariate$e_surgical_approach %in% c(2,3,7)),6] = 'Lapa'
covariate[which(covariate$e_surgical_approach %in% 5),6] = 'Robo'
covariate[which(covariate$e_surgical_approach %in% c(1,4,6)),6] = 'Open'
covariate[which(covariate$e_surgical_approach %in% c(11,12)),6] = 'Vagi'
covariate[which(covariate$e_surgical_approach %in% c(13,14,17,18,21,22,23,24)),6] = NA
covariate$e_surgical_approach = as.factor(covariate$e_surgical_approach)
summary(covariate$e_surgical_approach)

summary(covariate)
dim(covariate)
numcov = matrix(,ncol = 15, nrow = 23815)
for (i in 1:dim(covariate)[2]){
  numcov[,i] = as.numeric(covariate[,i])
}
```

```
summary(numcov)
colnames(numcov) = colnames(covariate)
rownames(numcov) = rownames(covariate)
cor(numcov, use = 'pairwise.complete.obs')

### blood colinearity
blood.model = lm(fluid_out_ebl_total~.,data = covariate[,!(colnames(covariate) %in% c("site_cid", "flg_cmb_open_wound",
'flg_cmb_smoker'))])

# blood.model = lm(fluid_out_ebl_total~val_surgtime + specmn_weight_grams + flg_cmb_open_wound + e_surgical_approach,data =
covariate)
summary(blood.model)
bloodres = covariate$fluid_out_ebl_total - predict(blood.model, covariate[,!(colnames(covariate) == "site_cid")])

###specmn_weight_grams
specweight.model = lm(specmn_weight_grams~.,data = covariate[,!(colnames(covariate) %in% c("site_cid",
'insurance_payment_type','cervix_removal_method','vaginal_cuff_suture','flg_cmb_open_wound'))])

# specweight.model = lm(specmn_weight_grams~val_surgtime + asa_class_id + val_bmi + e_surgical_approach, data = covariate)
summary(specweight.model)
specweightres = covariate$specmn_weight_grams - predict(specweight.model, covariate[,!(colnames(covariate) == "site_cid")])

### bmi
bmi.model = lm(val_bmi~val_surgtime + asa_class_id + specmn_weight_grams + e_surgical_approach, data =
covariate[,!(colnames(covariate) == "site_cid")])
bmi.model = lm(val_bmi~., data = covariate[,!(colnames(covariate) %in% c("site_cid", "insurance_payment_type",
'flg_cmb_open_wound'))])
summary(bmi.model)

bmires = covariate$val_bmi - predict(bmi.model, covariate[,!(colnames(covariate) == "site_cid")])


## covariate
covariate$bloodres = bloodres
covariate$specweightres = specweightres
covariate$bmires = bmires

# dataset
data.use = data.frame()
data.use = as.data.frame(as.numeric(cumu>0))
colnames(data.use) = 'response'
data.use$response = as.factor(data.use$response)
data.use = cbind(data.use,covariate)
summary(data.use)


#### complete rows:
data.comp = data.use[complete.cases(data.use),]
dim(data.comp)
#### fit in model.
logit.model.site <- glm(response ~ site_cid*(asa_class_id + flg_cmb_cancer + insurance_payment_type + specmn_weight_grams +
e_surgical_approach + cervix_removal_method + vaginal_cuff_suture + flg_cmb_open_wound + flg_cmb_smoker + val_age +
val_surgtime + val_bmi + bloodres) , family=binomial(link='logit'),data=data.comp)
coef.mat = summary(logit.model.site)$coefficient
```

```
View(coef.mat)
summary(coef.mat[,4]<0.05)
summary(logit.model.site)

View(coef.mat)
attributes(logit.model.site$coefficients)

logit.model <- glm(response ~ asa_class_id + flg_cmb_cancer + insurance_payment_type + specmn_weight_grams +
e_surgical_approach + cervix_removal_method + vaginal_cuff_suture + flg_cmb_open_wound + flg_cmb_smoker + val_age +
val_surgtime + val_bmi + fluid_out_ebl_total , family=binomial(link='logit'),data=data.comp)
summary(logit.model)

logit.model.2 <- glm(response ~ asa_class_id + flg_cmb_cancer + insurance_payment_type + specmn_weight_grams +
e_surgical_approach  + flg_cmb_open_wound + flg_cmb_smoker + val_age + val_surgtime + val_bmi + fluid_out_ebl_total ,
family=binomial(link='logit'),data=data.comp)
summary(logit.model.2)

logit.model.res <- glm(response ~ asa_class_id + flg_cmb_cancer + insurance_payment_type + specweightres + e_surgical_approach +
cervix_removal_method + vaginal_cuff_suture + flg_cmb_open_wound + flg_cmb_smoker + val_age + val_surgtime + bmires +
bloodres , family=binomial(link='logit'),data=data.comp)
summary(logit.model.res)

logit.model.2.res <- glm(response ~ asa_class_id + flg_cmb_cancer + insurance_payment_type + specweightres + e_surgical_approach
+ flg_cmb_open_wound + flg_cmb_smoker + val_age + val_surgtime + bmires + bloodres ,
family=binomial(link='logit'),data=data.comp)
summary(logit.model.2.res)$coefficient

logit.re.model <- glmer(response ~ asa_class_id + flg_cmb_cancer + insurance_payment_type + specmn_weight_type +
e_surgical_approach + cervix_removal_method + vaginal_cuff_suture + flg_cmb_open_wound + flg_cmb_smoker + val_age +
val_surgtime + val_bmi + bloodres + (1 | site_cid), data = data.comp, family = binomial)
summary(logit.re.model)

anova(logit.model,logit.model.2, test='Chisq')
anova(logit.model.res,logit.model.2.res, test='Chisq')


### investigating the missing values




followup = data$followed_for_30_days
data.use$followup = as.factor(followup)
data.healthy = data.use[data.use$response == '0',]
logit.model.follow <- glm(followup ~ asa_class_id + flg_cmb_cancer + insurance_payment_type + specmn_weight_grams +
e_surgical_approach + cervix_removal_method + vaginal_cuff_suture + flg_cmb_open_wound + flg_cmb_smoker + val_age +
val_surgtime + val_bmi + bloodres , family=binomial(link='logit'),data=data.use)
summary(logit.model.follow)
logit.model.follow.2 <- glm(followup ~ asa_class_id + flg_cmb_cancer + insurance_payment_type + specweightres +
e_surgical_approach + cervix_removal_method + vaginal_cuff_suture + flg_cmb_open_wound + flg_cmb_smoker + val_age +
val_surgtime + bmires + bloodres , family=binomial(link='logit'),data=data.use)
summary(logit.model.follow.2)
```

```
hospitalsummary = summary(as.factor(data.use$site_cid))
hist(hospitalsummary,breaks = 20)
sort(hospitalsummary)

data.miss = data.use
data.miss$followup = as.numeric(data.miss$followup)-1
data.miss$response = as.numeric(data.miss$response)-1

gb = group_by(data.miss,site_cid)
followrate = as.data.frame(summarise(gb,n(),sum(followup,na.rm = TRUE), sum(response,na.rm=TRUE)))
colnames(followrate) = c('site_cid','numofcase','numoffollowup','numofcomp')
followrate$percentfollow = followrate$numoffollowup/followrate$numofcase
followrate$percentresponse = followrate$numofcomp/followrate$numofcase

followrate$quantile = cut(followrate$numofcase,breaks = c(0,quantile(followrate$numofcase,c(0.25,0.5,0.75,1))))
quantile.gb = group_by(followrate,quantile)
summarise(quantile.gb,mean(percentfollow))

followrate$histlabel = cut(followrate$numofcase,breaks = c(0,160,470,700,1000))
histlabel.gb = group_by(followrate,histlabel)
summarise(histlabel.gb,mean(percentfollow))

data.new = left_join(data.use,followrate[,c('site_cid','quantile')],by = 'site_cid')
### fitting model for response
logit.model.res <- glm(response ~ quantile*( asa_class_id + flg_cmb_cancer + insurance_payment_type + specweightres +
e_surgical_approach + cervix_removal_method + vaginal_cuff_suture + flg_cmb_open_wound + flg_cmb_smoker + val_age +
val_surgtime + bmires + bloodres) , family=binomial(link='logit'),data=data.new)
summary(logit.model.res)
drop1(logit.model.res,test = 'Chisq')
logit.model.res.2 <- glm(response ~ quantile + asa_class_id + flg_cmb_cancer + insurance_payment_type + specweightres +
e_surgical_approach + cervix_removal_method + vaginal_cuff_suture + flg_cmb_open_wound + flg_cmb_smoker + val_age +
val_surgtime + bmires + bloodres , family=binomial(link='logit'),data=data.new)
summary(logit.model.res.2)
write.csv(summary(logit.model.res.2)$coef, 'logreg1.csv')

### fitting model for followup
logit.model.follow.res <- glm(followup ~ quantile*( asa_class_id + flg_cmb_cancer + insurance_payment_type + specweightres +
e_surgical_approach + cervix_removal_method + vaginal_cuff_suture + flg_cmb_open_wound + flg_cmb_smoker + val_age +
val_surgtime + bmires + bloodres) , family=binomial(link='logit'),data=data.new[data.new$response=='0',])
summary(logit.model.follow.res)
drop1(logit.model.follow.res,test = 'Chisq')
logit.model.follow.res.2 <- glm(followup ~ quantile + asa_class_id + flg_cmb_cancer + insurance_payment_type + specweightres +
e_surgical_approach + cervix_removal_method + vaginal_cuff_suture + flg_cmb_open_wound + flg_cmb_smoker + quantile*val_age +
val_surgtime + quantile*bmires + quantile*bloodres , family=binomial(link='logit'),data=data.new[data.new$response=='0',])
summary(logit.model.follow.res.2)
write.csv(summary(logit.model.follow.res.2)$coef, 'logreg2.csv')


#——————————————————————————————————————————(4)
counter-factual-----------------------------------------------
#category of hospitals
hospitals=levels(msqc$site_cid_160801)

#use the hospital of highest frequency as the standard
```

```
table=data.frame(table(msqc$site_cid_160801))
max=which.max(table$Freq)
#hospital = table[max,1]
hospital = table[2,1]


#function: logistic regression
control=msqc[msqc$site_cid_160801==hospital,]
logit=glm(response~ asa_class_id
+flg_cmb_cancer+insurance_payment_type+specmn_weight_type+e_surgical_approach+flg_cmb_open_wound+flg_cmb_smoker+val_
age+val_surgtime+val_bmi+fluid_out_ebl_total,family="binomial", data=control)

#logit1=glm(response~ asa_class_id
+flg_cmb_cancer+insurance_payment_type+specmn_weight_type+e_surgical_approach+flg_cmb_open_wound+flg_cmb_smoker+val_
age+val_surgtime+val_bmi+fluid_out_ebl_total,family="binomial", data=msqc)

#predict = round(predict(logit, control[,-c(1,13)],type="response"),3)
#t.test(predict,logit$y,paired=TRUE)
#t.test(predict,as.numeric(control[,13])-1,paired=TRUE)



t.stats=rep(0,length(hospitals))

for(i in 1:length(hospitals)){
  hospital = hospitals[i]
  treatment = msqc[msqc$site_cid_160801==hospital,]
  if(dim(treatment)[1]==0){
    print(hospital)
    next
  }
  predict = round(predict(logit, treatment[,-c(1,13)],type="response"),3)
  #predict = 1*(predict>0.5)
  #comp=data.frame(treatment[,13],predict)
  #paired t-test
  t=t.test(as.numeric(treatment[,13])-1,predict,paired=TRUE)
  #other statistics
  t.stats[i]=t$statistic
  if(i==max){
    print(t$statistic)
  }
}

score =data.frame(table[,1],t.stats)
names(score)=c('name','t-stats')
result = score[order(score$`t-stats`),]


write.csv(result, file ='score2.csv')


#------------------------plot the data------------------
boxplot(table[,2], staplewex = 1)
text(y=fivenum(table[,2]), labels =fivenum(table[,2]),x=1.25)
```

```
points(rep(1,length(table[,2])),table[,2])

#-----------------------summary of the variables-----------
barplot(table(control$asa_class_id),main='hospital _034_')
barplot(table(control$flg_cmb_cancer),main='hospital _034_')
barplot(table(control$insurance_payment_type),main='hospital _034_')
barplot(table(control$e_surgical_approach),main='hospital _034_')
barplot(table(control$flg_cmb_cancer),main='hospital _034_')
barplot(table(control$flg_cmb_cancer),main='hospital _034_')


## -------function---------
balance <- function(data, respond, treatment, control, n){
  treatment_group = data[data[, respond] == 1,]
  control_group = data[data[, respond] == 0,]
  treatment_group = treatment_group[sample(dim(treatment_group)[1], n * treatment, replace = TRUE),]
  control_group = control_group[sample(dim(control_group)[1], n * control),]
  return(rbind(treatment_group, control_group))
}

stratify_psm <- function(x){
  g = NULL
  for ( i in 1:length(x)){
    if ((sum(x[i] > x)/length(x)) > 0.75){
      g = c(g, 1)
    }else if((sum(x[i] > x)/length(x)) > 0.5){
      g = c(g, 2)
    }else if((sum(x[i] > x)/length(x)) > 0.25){
      g = c(g, 3)
    }else{
      g = c(g, 4)
    }
  }
  return(g)
}

stra_hypertest <- function(x, y, g){
  group_label = unique(g)
  stat = NULL
  Ut = 0
  Vt = 0
  for(i in group_label){
    k = sum((x[g==i]==1)&(y[g==i] ==1))
    tK = sum(x[g==i]==1)
    tD = sum(y[g==i])
    tAll = sum(g==i)
    U = k - tK*tD/tAll
    V = tD * tK * (tAll - tK) * (tAll - tD) / (tAll^2) / (tAll - 1)
    Ut = Ut + U
    Vt = Vt + V
    stat = c(stat, U / (V^(1/2)))
  }
  stat = c(stat, Ut / (Vt^(1/2)))
  names(stat) = c(group_label, "combined")
  return(stat)
}

## ---------------
# install.packages("readstata13")
library(readstata13)
mydata<-read.dta13("~/Documents/UM Class/Statistics Classes/504 Stats Consulting/Project/procedure13.dta")
## write.csv(mydata, file = "~/Documents/UM Class/Statistics Classes/504 Stats Consulting/Project/mydata.csv")
# install.packages("MatchIt")
library(MatchIt)

#--------------------------------------(1) Matching Variables-----------------------------------------------
covariate =
mydata[,c("site_cid_160801","asa_class_id","flg_cmb_cancer","e_insurance_type","specmn_weight_grams","e_surgical_approach","cer
vix_removal_method","vaginal_cuff_suture","flg_cmb_open_wound","flg_cmb_smoker","val_age","val_surgtime","val_bmi","fluid_out
_ebl_total")]
covariate_name =
c("e_surgical_approach","site_cid_160801","asa_class_id","flg_cmb_cancer","e_insurance_type","specmn_weight_grams","cervix_rem
oval_method","vaginal_cuff_suture","flg_cmb_open_wound","flg_cmb_smoker","val_age","val_surgtime","val_bmi","fluid_out_ebl_tot
al")

#------concatenate responds---------------------------------

respond_name = c("flg_cmp_any","flg_dead30","flg_cmp_pneumonia","val_los","flg_util_readm","flg_util_reop","flg_util_transfused")
respond_matrix = mydata[, respond_name]
respond_matrix$val_los = as.numeric(respond_matrix$val_los>2)
cumu = summary(as.factor(rowSums(respond_matrix[,],na.rm=T)))
sum(cumu[-1])/sum(cumu[])
respond = as.numeric(rowSums(respond_matrix, na.rm = TRUE) > 0)

covariate = cbind(covariate, respond)

# asa_class_id: 1,2->0; 3,4,5->1 and remove rows with id==7
matching_variables<-covariate[covariate$asa_class_id!=7,]
matching_variables$asa_class_id=as.factor(as.numeric(matching_variables$asa_class_id>2))


# count(mydata$asa_class_id)
#  x  freq
#1 1  2420
#2 2 16041
#3 3  5186
#4 4   157
#5 5     3
#6 7     8

#flg_cmb_cancer:0,1,NA
matching_variables$flg_cmb_cancer = as.factor(matching_variables$flg_cmb_cancer)

#e_insurance_type: All others->0; 3 (Medicaid)& 8 (self-pay) & 9 (Uninsured)->1
matching_variables$e_insurance_type = as.factor(as.numeric(matching_variables$e_insurance_type==3 |
matching_variables$e_insurance_type==8 | matching_variables$e_insurance_type==9))

#library(plyr)
```

```
#count(mydata$e_insurance_type)
#    x freq
#1   1 9825
#2   2 1356
#3   3 2017
#4   4  450
#5   5 7446
#6   6   31
#7   8  196
#8   9 1285
#9  10 1082

#specmn_weight_grams
matching_variables$specmn_weight_grams = as.factor(matching_variables$specmn_weight_grams >
quantile(matching_variables$specmn_weight_grams, 0.8,na.rm = TRUE))

# e_surgical_approach: 1 & 4 & 6 -> 1, 2 & 3 & 7 -> 2, 5 -> 3, 11 & 12 -> 4, (8 & 9 & 10 &) 13 & 14,17,18,21,22,23,24  -> 5

# 1 - Open
# 2 - Laparoscopic
# 3 - Laparoscopic, Hand-Assisted
# 4 - Laparoscopic, Converted to Open
# 5 - Robotic
# 6 - Robotic, Converted to Open
# 7 - Laparoscopic, Single Port (SIL)
# 8 - Open Endo
# 9 - Percutaneous Endo
# 10 - Percutaneous Endo, converted to Open Endo
# 11 - Vaginal
# 12 - Vaginal, lap-assisted
# 13 - Other (e.g. anal, transanal, transoral, etc.)

matching_variables$e_surgical_approach[matching_variables$e_surgical_approach==1|matching_variables$e_surgical_approach==4|ma
tching_variables$e_surgical_approach==6]<-'a'
matching_variables$e_surgical_approach[matching_variables$e_surgical_approach==2|matching_variables$e_surgical_approach==3|ma
tching_variables$e_surgical_approach==7]<-'b'
matching_variables$e_surgical_approach[matching_variables$e_surgical_approach==5]<-'c'
matching_variables$e_surgical_approach[matching_variables$e_surgical_approach==11
|matching_variables$e_surgical_approach==12]<-'d'
matching_variables$e_surgical_approach[matching_variables$e_surgical_approach==13 |matching_variables$e_surgical_approach==14
| matching_variables$e_surgical_approach==17 | matching_variables$e_surgical_approach==18
|matching_variables$e_surgical_approach==21 |matching_variables$e_surgical_approach==22
|matching_variables$e_surgical_approach==23 |matching_variables$e_surgical_approach==24]<-'e'
matching_variables$e_surgical_approach[matching_variables$e_surgical_approach=='a']<-1
matching_variables$e_surgical_approach[matching_variables$e_surgical_approach=='b']<-2
matching_variables$e_surgical_approach[matching_variables$e_surgical_approach=='c']<-3
matching_variables$e_surgical_approach[matching_variables$e_surgical_approach=='d']<-4
matching_variables$e_surgical_approach[matching_variables$e_surgical_approach=='e']<-5
matching_variables$e_surgical_approach = as.factor(matching_variables$e_surgical_approach)

# Smoker
matching_variables$flg_cmb_smoker=as.factor(matching_variables$flg_cmb_smoker)

# Age
```

```
matching_variables$val_age=cut(matching_variables$val_age,breaks=c(min(matching_variables$val_age),40,55,max(matching_variables$val_age)))

# Surgical time
matching_variables$val_surgtime=matching_variables$val_surgtime>=quantile(matching_variables$val_surgtime,probs=0.75,na.rm=T)
matching_variables$val_surgtime=as.factor(as.numeric((matching_variables$val_surgtime)))

# BMI
matching_variables$val_bmi=matching_variables$val_bmi>=35
matching_variables$val_bmi=as.factor(as.numeric((matching_variables$val_bmi)))

# Approximation for blood loss
matching_variables$fluid_out_ebl_total=matching_variables$fluid_out_ebl_total>=quantile(matching_variables$fluid_out_ebl_total,probs=0.75,na.rm=T)
matching_variables$fluid_out_ebl_total=as.factor(as.numeric((matching_variables$fluid_out_ebl_total)))

# Open wound or not
matching_variables$flg_cmb_open_wound=as.factor(matching_variables$flg_cmb_open_wound)

#cervix removal method
matching_variables$cervix_removal_method = as.factor(matching_variables$cervix_removal_method)

# vaginal_cuff_suture
matching_variables$vaginal_cuff_suture = as.factor(matching_variables$vaginal_cuff_suture)

# Hospital ID
matching_variables$site_cid_160801 = as.factor(matching_variables$site_cid_160801)

#--------------------------------------(2) Remove NA------------------------------------------------

clean_data<-matching_variables[complete.cases(matching_variables),]



#--------------------------------------(3) Response and PSM (Use preoperation variables) ------------------------------------------------

## Choose 62 different Hospitals as standard

PSM_nn<-list()

surgical_approach=sort(unique(clean_data$e_surgical_approach))
path = "~/Documents/UM Class/Statistics Classes/504 Stats Consulting/Project/"


for ( i in surgical_approach )
{
  clean_data$approach=ifelse(clean_data$e_surgical_approach==i,1,0)

  #plot prematch distribution of covariate
  dir.create(sprintf('~/Documents/UM Class/Statistics Classes/504 Stats Consulting/Project/new_results/groupcov/approach%s/', i))
  for (k in covariate_name[-1])
    {
    jpeg(file = paste(path,sprintf('new_results/groupcov/approach%s/prematch_cova_%s_treatment.jpeg',i,k), sep = ""))
    summ = summary(clean_data[(clean_data$approach==1),k])
    barplot(summ/sum(summ))
    dev.off()
    jpeg(file = paste(path,sprintf('new_results/groupcov/approach%s/prematch_cova_%s_control.jpeg',i,k), sep = ""))
    summ = summary(clean_data[(clean_data$approach==0),k])
    barplot(summ/sum(summ))
    dev.off()
  }




   # balance the sample size of treatment and control
  balance_data  = balance(data = clean_data, respond = "approach", treatment = 1, control = 3, n = floor(sum(clean_data$approach == 0)
/ 3))
  m.out=matchit(approach~site_cid_160801+asa_class_id+flg_cmb_cancer+e_insurance_type+specmn_weight_grams+
                  cervix_removal_method+vaginal_cuff_suture+flg_cmb_open_wound+
                  flg_cmb_smoker+val_age+val_bmi+val_surgtime+fluid_out_ebl_total
                ,data=balance_data,method="nearest", ratio = 1)

  psm_treat_names =rownames(m.out$match.matrix)
  psm_treat_ps = m.out$model$fitted.values[rownames(m.out$match.matrix)]
  psm_control_names = m.out$match.matrix[,1]
  psm_control_ps = m.out$model$fitted.values[psm_control_names]
  g = stratify_psm(psm_treat_ps)
  m.out$balance_data = rbind(balance_data[psm_treat_names,],balance_data[psm_control_names,])
  m.out$subclass = c(g,g)
  jpeg(file=paste(path,sprintf('new_results/Jitter/Jitter%s.jpeg',i), sep = ""))
  plot(m.out,type="jitter")
  dev.off()

  jpeg(file=paste(path,sprintf('new_results/Hist/Hist%s.jpeg',i), sep = ""))
  plot(m.out,type="hist")
  dev.off()

  PSM_nn[[length(PSM_nn)+1]]<-m.out

  match_PSM_nn<-match.data(m.out)
  write.csv(match_PSM_nn,file=paste(path,sprintf('new_results/match_data/matched_data%s.csv',i), sep=""))
  g = m.out$subclass
  for (j in unique(g)){
    for (k in covariate_name[-1])
      {
      jpeg(file = paste(path,sprintf('new_results/groupcov/match/approach%sgroup_%s_cova_%s_treatment.jpeg',i,j,k), sep = ""))
      summ = summary(m.out$balance_data[(m.out$balance_data$approach==1) & (g==j),k])
      barplot(summ/sum(summ))
      #hist(probability =TRUE, as.numeric(balance_data[(balance_data$hospital==1) & (g==j),k]))
      dev.off()
      jpeg(file = paste(path,sprintf('new_results/groupcov/match/approach%sgroup_%s_cova_%s_control.jpeg',i,j,k), sep = ""))
      summ = summary(m.out$balance_data[(m.out$balance_data$approach==0) & (g==j),k])
      barplot(summ/sum(summ))
      #hist(probability  = TRUE, (balance_data[(balance_data$hospital==0) & (g==j),k]))
      dev.off()
    }
  }
}

save(PSM_nn,file=paste(path, "new_results/PSM_nn_approach.RData", sep=""))
summary(PSM_nn[[1]])


## ------Do test on the table----------------------------
stat = list()
combined_stat = NULL
for (i in 1:length(PSM_nn)){
  y = PSM_nn[[i]]$balance_data$respond
  x = PSM_nn[[i]]$balance_data$approach
  g = PSM_nn[[i]]$subclass
  stat[[i]] = stra_hypertest(x = x, y = y, g = g)
  combined_stat = c(combined_stat, stat[[i]]["combined"])
  names(combined_stat)[i] = surgical_approach[i]
}
write.csv(sort(combined_stat),          file          =          "~/Documents/UM          Class/Statistics          Classes/504          Stats
Consulting/Project/new_results/combined_stat.csv")



#--------------------------------------(4) Plot distribution of surgical approaches for each hospital ------------------------------------------------

sort_siteid=sort(unique(clean_data$site_cid_160801))

for (i in sort_siteid)
    {
        approach.freq=table(clean_data$e_surgical_approach[clean_data$site_cid_160801==i])
                                                                                jpeg(file=sprintf('~/Documents/UM        Class/Statistics        Classes/504        Stats
Consulting/Project/new_results/Hospital_approach_barplots/Hospital%s.jpeg',i))
        barplot(approach.freq/sum(approach.freq),ylim=c(0,1),xlab="Surgical Approaches",main=sprintf('Hospital%s',i))
        dev.off()

    }
```