UCLA Statistics 141XP

Surgical Site Infection Project for Dr. Rootman

Professor Mahtash Esfandiari
Spring 2022

Nora Liu, Himani Yalamaddi, Jaehyeok Park, Genesis Qu,
Chae Yeon Lim, Yuelong Zhang

1 Abstract

Skin cancer excision encounters from 2015-2021 were analyzed to two ends: to determine the relationship between an incision procedure's location and the rate of surgical site infection, and to identify patient characteristics and comorbidities that may increase the chances of infection. Postoperative antibiotic prescriptions were used as a marker for surgical site infection. Ultimately, a mixed-effect logistic regression showed promise in prediction.

For data analysis, two models were fitted to our data: the random intercept model and the Firth logistic regression model. Moreover, ROC curve was used to evaluate which model performs better. From the model, we also found which variables are the best predictors for our outcome variable. Lastly, we addressed the limitations of the data provided and our models and recommendations for further investigation to retrieve the optimal results.

2 Problem statement

Before jumping into fitting the model, first, we cleaned data and explained the data based on the cleaned data that includes all encounters for a patient while removing encounters with multiple diagnoses and encounters with non-malignant cancer procedures. Since we kept only one encounter for a patient, we made a new independent variable called 'multiple encounters,' which is a categorical variable with two groups: patients with multiple encounters and patients with one encounter. Subsequently, we conducted explanatory data analysis based on our outcome variable and independent variables to get some general idea regarding our data.

Therefore, we sought to find out whether there is any factor that is important when making a prediction of infection after a surgical procedure. Through this explanatory data analysis, we could fetch two related questions about the likelihood of postoperative infections. The first question is, does tissue perfusion that is controlled for other possible risk factors decrease the risk of surgical site infections? Furthermore, the second question is, are there other factors that mediate SSI risk?

To seek the answers to our questions, we used two models. The first model was Firth logistic regression model. The Firth logistic regression model is a valid score function by adding a term that counteracts the first-order term from the asymptotic expansion of the bias of the maximum likelihood estimation. It is known for the analysis of binary outcomes with small samples, which fits our case. However, the ROC curve was not showing the optimal result for us; thus, we used the Random Intercept model to enhance our outcome.

The Random Intercept model is one of the linear mixed models but contains an error and intercept random effects. Thus we can observe that there is an error and intercept term in our constructed model equation. Using this model, we could retrieve a better ROC curve than the Firth logistic regression model has produced, indicating that the random intercept model is doing a good job analyzing our given data.

3 Schematics of Variables and Measurement

3.1 Data Preparation

Before we dive into the variables, we would like to make clear the steps taken to create the two data sets used in our analyses. Referring to Figure 1, the first data set (”full_v5.csv”) includes all every encounter that a patient had (regarding a skin lesion), while the second data set (”last_v4.csv”) only includes the last encounter for each patient in the data set. All variables in both data sets are identical, except ”multiple encounter” in ”last_v4.csv,” which indicates whether a patient had multiple encounters.

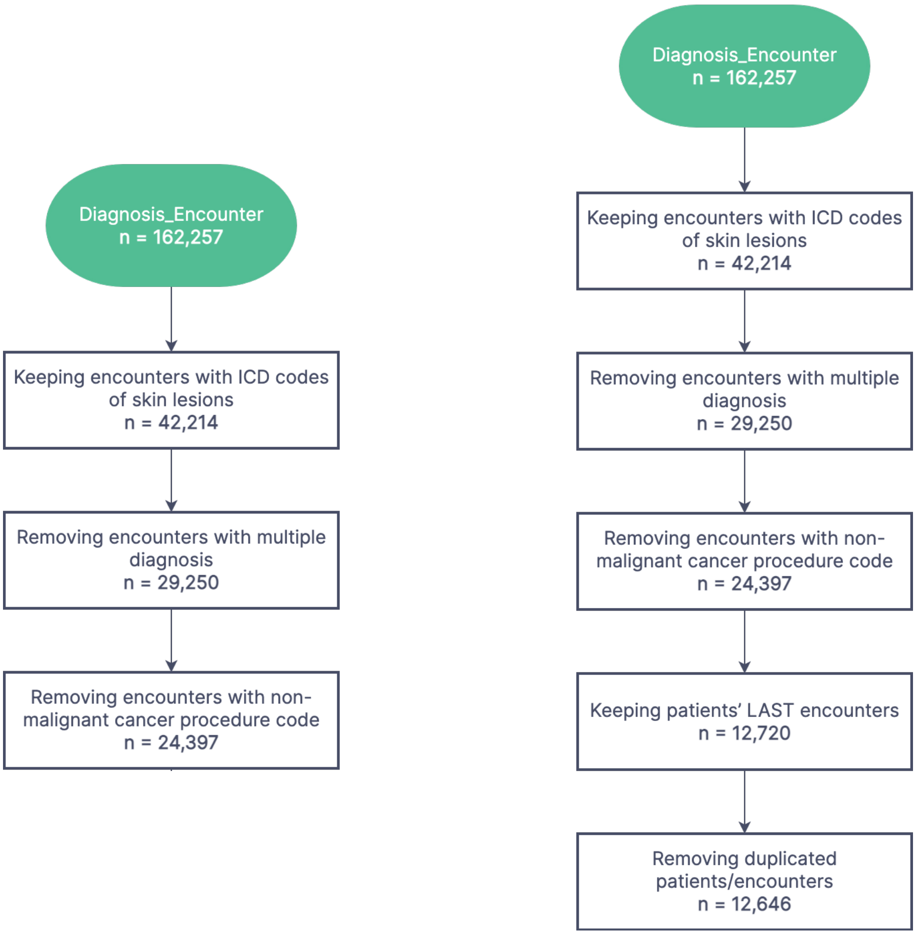

Diagnosis_Encounter
n = 162,257
Keeping encounters with ICD codes of skin lesions
n = 42,214
Removing encounters with multiple diagnosis
n = 29,250
Removing encounters with non-malignant cancer procedure code
n = 24,397
Diagnosis_Encounter
n = 162,257
Keeping encounters with ICD codes of skin lesions
n = 42,214
Removing encounters with multiple diagnosis
n = 29,250
Removing encounters with non-malignant cancer procedure code
n = 24,397
Keeping patients’ LAST encounters
n = 12,720
Removing duplicated patients/encounters
n = 12,646

Figure 1: Data Preparation Flowchart. The left one is the flowchart for all the encounters of skin lesion of patients, while the right one is the flowchart for the data set of just last encounter of patients. The preparation steps are mostly the same while the latter has two more steps.

3.2 Schematics

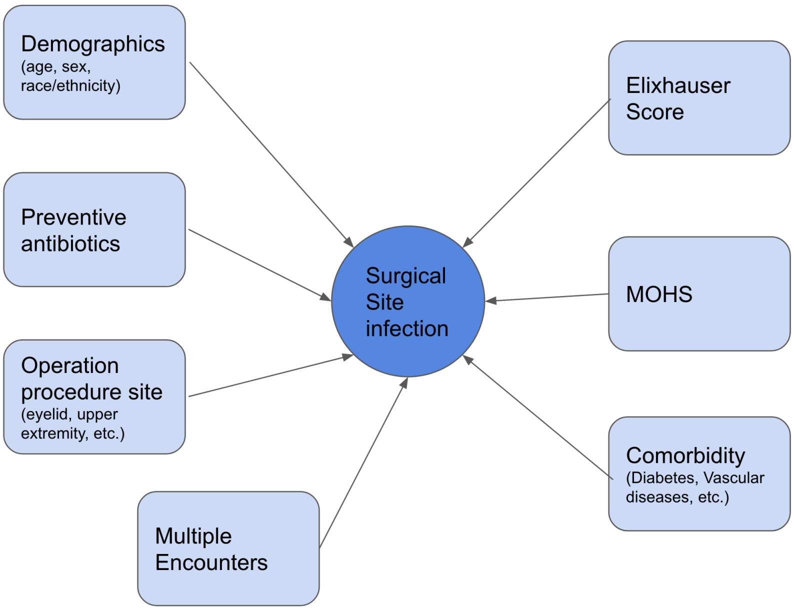


Figure 2: Schematic of Variables

The main focus of our project is to investigate the whether or not operation procedure location correlates with the risk of developing surgical site infection. We also want to determine if there are other factors (such as preventive antibiotics, size of excision, vascular disease) that mediate this risk. Each block in Figure 2 represents one group of variables. Our major concern, procedure location is a categorical variable with multiple levels, where each level correspond to one particular body part or a group of related body parts. Some example would be "eyelid", "upper extremity", etc.

3.3 Variable Description and Measurement

Here we would elaborate on each variable and how they are measured in this project

- **Response variable: Surgical site infections**: binary categorical variable. This is measured by whether the patients are prescribed antibiotics within 30 days after the procedure. We would look at each drug's start date, corresponding procedure date, and whether it is an antibiotics to decide whether the patients are prescribed antibiotics within 30 days after the procedure. (file: *encounter_diagnosis.csv, Medications.csv, Antibiotics.csv)*

- **Independent variables**:

  – **Patient demographics**: patients' age (numerical), sex (categorical), race/ethnicity (categorical). These are directly available from the patients' encounter records (file: *encounter_diagnosis.csv*)

  – **Procedure site**: categorical variable, the location of the procedure. These are extracted from the skin lesion ICD-10 code. These codes correspond to patients' diagnosis, so would have information on operation or procedure sites. Originally procedure site has 11 categories: "Eyelid", "Chest/abdomen", "Lower extremity" ,"Ears", "Nose", "Mouth/Lips/Pharynx", "Head/face/neck (unspecified)", "Upper extremity", "Other/unspecified", "Genitourinary", "Anus". However, since the counts for each category is skewed, we decide to combine a few categories based on recommendations: "Anus" with "Genitourinary", "Ears", "Nose", and "Mouth/Lips/Pharynx" are together; so we end up getting 8 levels for procedure sites. (file: *encounter_diagnosis.csv, CPT and ICD10 Codes.xlsx*)

  – **MOHS/Excision Size**: categorical, whether the procedure is MOHS (1) or non-MOHS (0), and the size of the excision for non-MOHS procedures. This information is available in the procedure dataset (file: *Procedures.csv*)

- **Elixhauser Score**: numerical variable named elix_vw_score, which is the Van Walraven score of patients about comorbidities. The patients check boxes for the provided list of diseases they have. With some calculations, higher score indicates the patients suffer from more diseases. Patients with no records in Elixhauser score means they do not check any box in the provided list of Elixhauser diseases. Missing scores are replaced with minimum scores. (file: *elixhauser.csv*)
- **Preventive antibiotics**: binary categorical variable. This variable indicates whether the prescribed antibiotics within 7 days before the procedure, 1 for yes and 0 for no. This variable is extracted in the similar way to the response variable. We look at whether the medication, which needs to be antibiotics, start within 7 days before the procedure. (file: *encounter_diagnosis.csv, Medications.csv, Antibiotics.csv)*
- **Multiple Encounters**: binary categorical variable. This variable indicates whether a patient had multiple encounters, 1 for multiple encounters, 0 for single encounter. Since we only keep the last encounter for each patient, it is important to keep track of how many encounters they have before. This can be counted by the number of times the patient id appears in the enconter dataset before this last encounter (file: *encounter_diagnosis.csv*)
- **Comorbidity**: More precisely speaking, this is a group of binary categorical variable, including diabetes, immunocompromised, autoimmune, tobacco, and vascular diseases. Patients get 1 if they have records indicating that they havie these diagnosis in the encounter, 0 otherwise. This information is extracted from the ICD-9 and -10 codes from the encounter dataset. We can use this information to keep track of what diagnosis this patient has ever received. (file: *encounter_diagnosis.csv, CPT and ICD10 Codes.xlsx*)

4 Exploratory Data Analysis

4.1 Response Variable: Surgical site infections

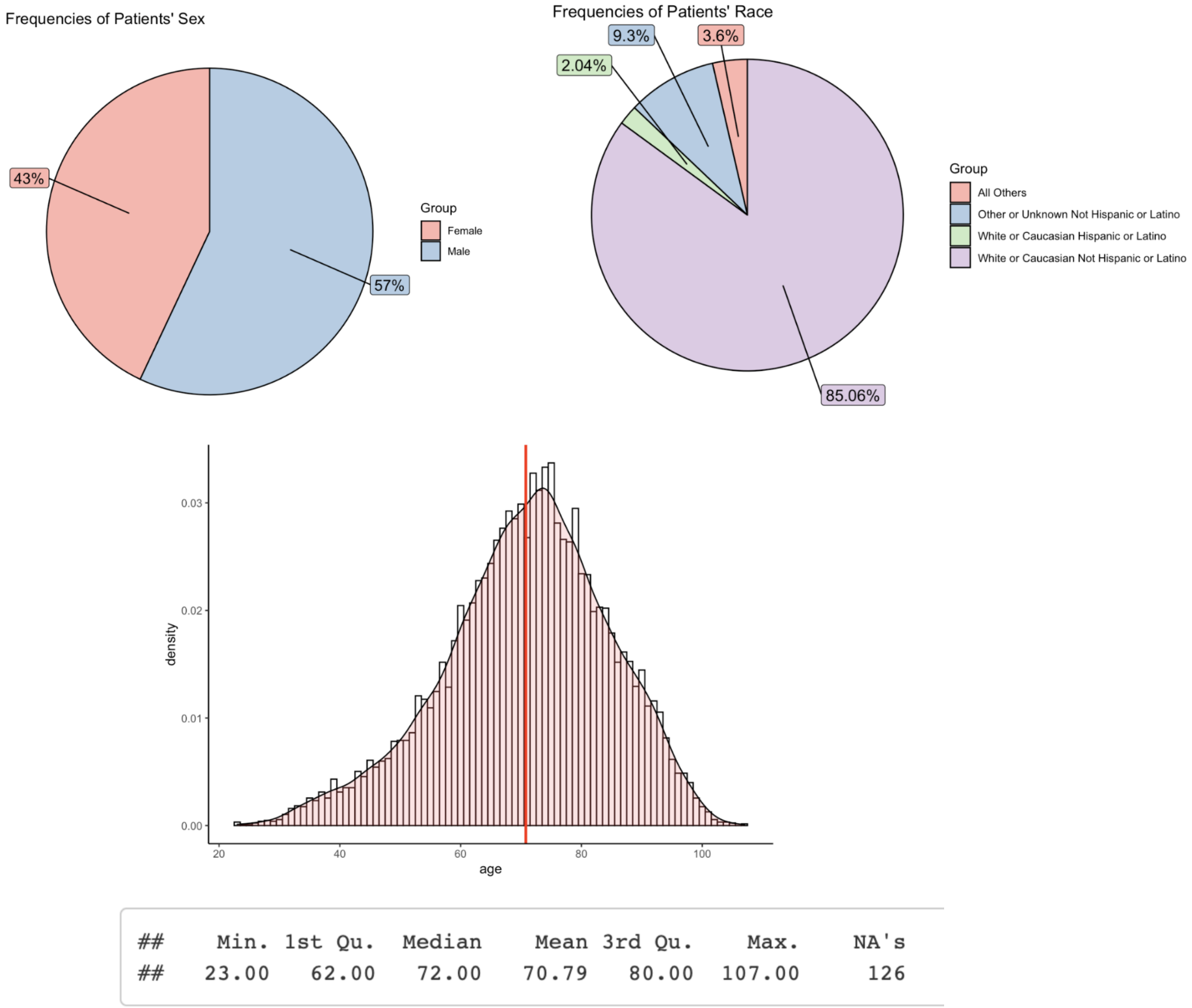
infected	not infected
12551	95

Table 1: Count of response variable

Table 1 is the count of the surgical site infection (SSI). We have a very skewed dataset with only 0.75% infection rate. This would affect our choice of statistical analysis methods later.

4.2 Predictor: Demographic information

Figure 3 shows the distribution of basic demographic information of patients included in the study. We have 57% male and 43% female. The predominant race is white or Caucasian not Hispanic, holding 85% in the patient population. We have a relatively balanced age distribution with mean and median at around 70. 126 patients have missing data on age



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   23.00   62.00   72.00   70.79   80.00  107.00     126
```

Figure 3: Data visualization of basic demographic information: patient sex, race, and age

4.3 Predictor: Grouped Procedure Site

Procedure site is in fact our top priority concern in this study. Due to relatively high imbalance in procedure site count, we combined a few sites based on recommendation. Figure 4 is the bar plot of percentage of SSI for each grouped procedure site. Table 2 gives the exact number of people in each category of SSI VS procedure locations. Both give us a good indication that Anus/Genitourinary has the highest proportion of SSI, followed by Ears/Nose/Mouth/Lips/Pharynx, and Head/Face/Neck. If we make eyelid the base level for procedure site, we might see positive coefficients for Anus/Genitourinary

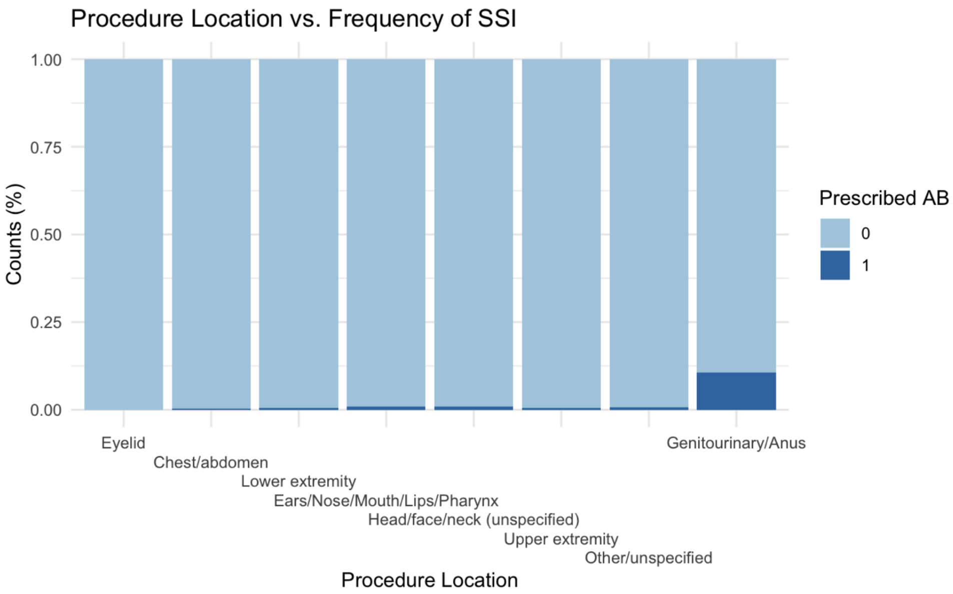


Figure 4: Caption

Table 2: Infection VS Grouped Procedure Site

	not infected	infected
Eyelid	312	0
Chest/abdomen	1181	3
Lower extremity	1342	8
Ears/Nose/Mouth/Lips/Pharynx 2079	21	
Head/face/neck (unspecified) 5564	48	
Upper extremity	1767	8
Other/unspecified	264	2
Genitourinary/Anus	42	5

4.4 Predictor: MOHS

MOHS is a special type of excision method where the surgeons ensure the exact outline of the tumor is cut off. The excised section is observed under microscope until the whole tumor is removed cleanly. Usually MOHS procedure involves very small section, but there is no clear definition on this.

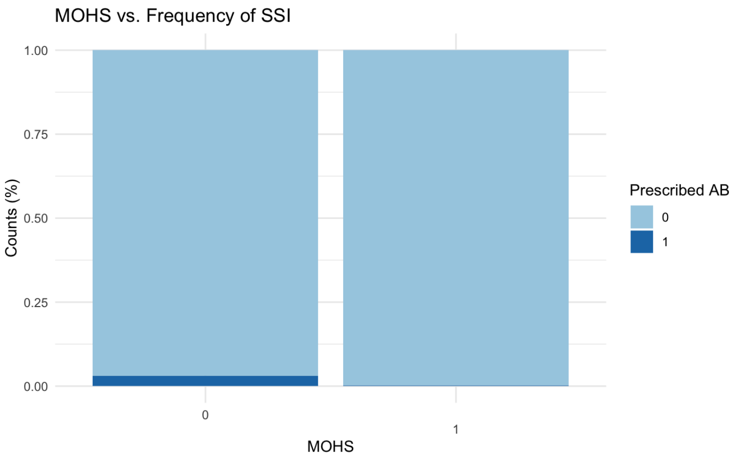


Figure 5: Boxplot of MOHS procedure Antibiotics against SSI

	non-infected	infected
non-MOHS	2164	70
MOHS	10387	25

Table 3: Count of MOHS

Pearson's Chi-squared test with Yates' continuity correction
data: mohs and SSI
X-squared = 202.65, df = 1, p-value < 2.2e-16

From table 3 we do see that non-MOHS procedures show a higher infection: non-infection ratio than MOHS procedures. The chi-squared test also gives a p-value less than 0.05, indicating that MOHS and SSI are not independent, so we therefore include MOHS as a predictor in our model

4.5 Predictor: Elixhauser Score

Elixhauser score is a score about comorbidities. Each patient would check the boxes for a series of diseases, including diabetes, etc. The higher the elixhauser score, the more diseases the patients suffer from.

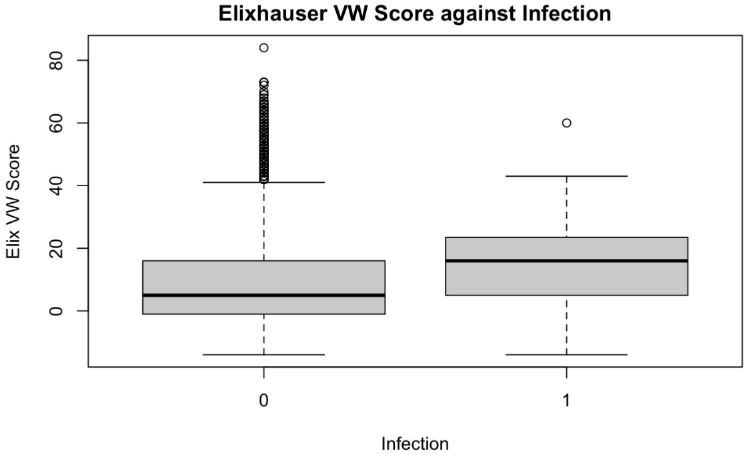


Figure 6: Boxplot of Elixhauser VW score against SSI

Figure 6 shows that the infected patients have a higher median Elixhauser vw score than non-infected patients. This observation is further corroborated by Welch Two Sample t-test with p value < 0.5. The mean Elixhauser score for infected patients is significantly higher than non-infected patients .Infected patients have a mean score of 14.42, about 6 points higher than non-infected patients (7.90)

Welch Two Sample t-test
data: elix_vw_score by antibiotics_after_procedure_less_thirty
t = -4.4843, df = 95.72, p-value = 2.032e-05
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval: -9.404590 -3.633197
sample estimates:
mean in group 0 mean in group 1
7.902159 14.421053

4.6 Predictor: Preventive Antibiotics

Preventive antibiotics is defined by whether the patients are prescribed antibiotics within 7 days before the procedure. This is a binary categorical variable.

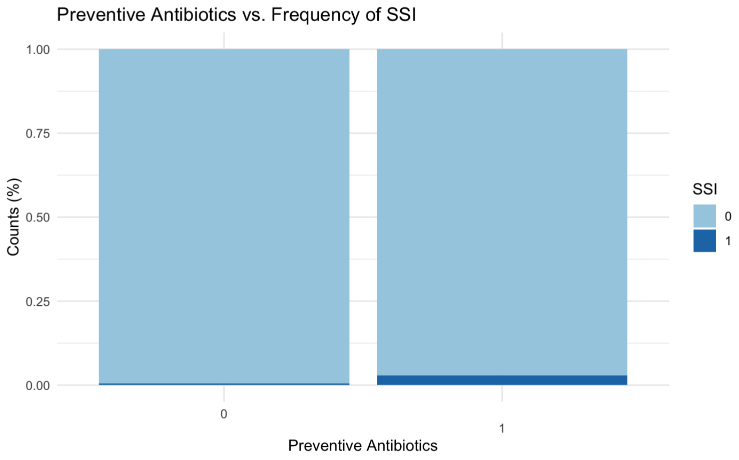


Figure 7: Bar plot of the use of Preventive Antibiotics against SSI

	non-infected	infected
No Preventive Antibiotics	11284	57
Preventive Antibiotics	1267	38

Table 4: Count of Preventive Antibiotics against SSI

Figure 7 and table 4 shows that patients who are prescribed preventive antibiotics have higher proportion of SSI

Pearson's Chi-squared test with Yates' continuity correction
data: preventive_antibiotics and antibiotics_after_procedure_less_thirty
X-squared = 87.911, df = 1, p-value < 2.2e-16

The chi-squared test also gives a p-value less than 0.05, indicating that preventive antibiotics and SSI are not independent, so we therefore include preventive antibiotics as a predictor in our model

4.7 Predictor: Multiple Encounter

Multiple encounters is a binary category variable with 1 representing the patient has more than 1 encounter and 0 meaning the patient only has 1 encounter. We group this encounter variable this way to create a more balanced subset.

	non-infected	infected
Single Encounter	8070	78
Multiple Encounter	4481	17

Table 5: Count of Encounter times VS SSI

Figure 8 shows that patients with single encounter have higher proportion of SSI, which is further confirmed by chi-square test. The chi-squared test gives a p-value less than 0.05, indicating that whether the patients have multiple encounters and SSI are not independent, so we therefore include this variable as a predictor in our model

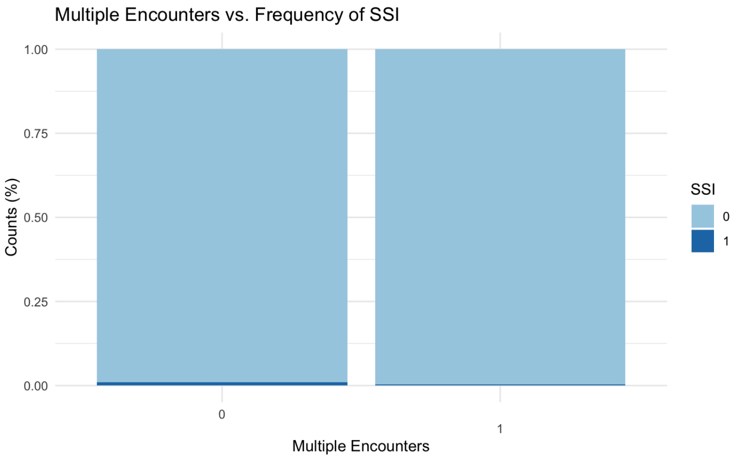


Figure 8: Bar plot of Multiple Encounter against SSI

Pearson's Chi-squared test with Yates' continuity correction
data: multiple_encounters and antibiotics_after_procedure_less_thirty
X-squared = 12.281, df = 1, p-value = 0.0004576

4.8 Predictor: Comorbidity: Diabetes

Figure 9 shows that patients with diabetes history in records have obviously higher proportion of SSI than non-diabetic patients. About 38% of diabetic patients have SSI, while only 0.7% non-diabetic patients suffer from SSI.

The chi-squared test also gives a p-value less than 0.05, indicating that diabetes and SSI are not independent, so we therefore include diabetes as a predictor in our model.

	non-infected	infected
No Diabete	12530	82
Diabete	21	13

Table 6: Count of Diabetes VS SSI

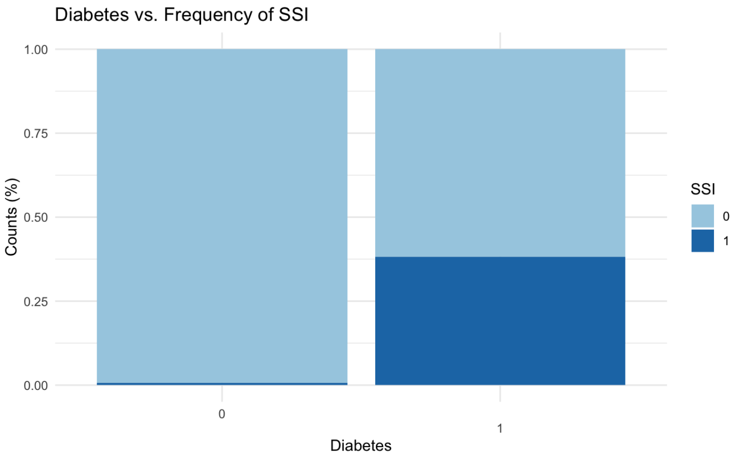


Figure 9: Bar plot of Diabetes against SSI

Pearson's Chi-squared test with Yates' continuity correction
data: Diabetes and antibiotics_after_procedure_less_thirty
X-squared = 593.04, df = 1, p-value < 2.2e-16

4.9 Predictor: Comorbidity: Vascular Diseases

Figure 10 shows that patients with vascular disease history in records have obviously higher proportion of SSI than non-vascular diseases patients. About 36% of vascular disease patients have SSI, while only 0.7% non-vascular disease patients suffer from SSI.

The chi-squared test also gives a p-value less than 0.05, indicating that vascular disease and SSI are not independent, so we therefore include vascular disease as a predictor in our model.

	non-infected	infected
No Vascular Diseases	12542	90
Vascular Diseases	9	5

Table 7: Count of Vascular Diseases VS SSI

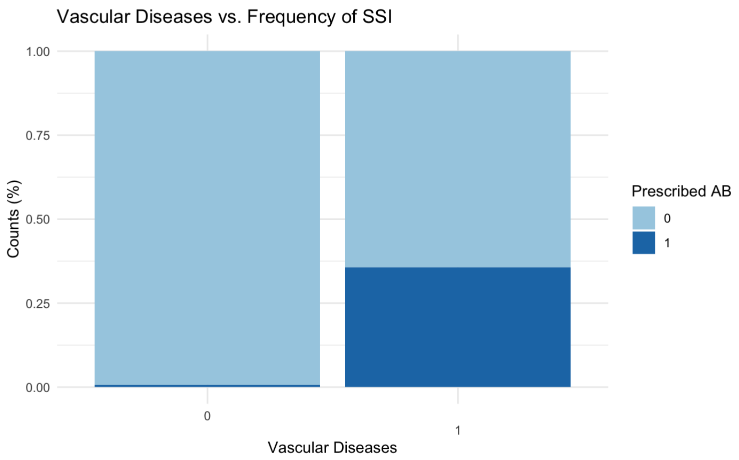


Figure 10: Bar plot of Vascular Diseases against SSI

Pearson's Chi-squared test with Yates' continuity correction
data: Vascular.disease and antibiotics_after_procedure_less_thirty
X-squared = 185.24, df = 1, p-value < 2.2e-16

4.10 Predictor: Comorbidity: Autoimmune Diseases

Figure 11 shows that patients with autoimmune disease history in records have obviously higher proportion of SSI than non-autoimmune disease patients. About 37.5% of autoimmune disease patients have SSI, while only 0.7% non-autoimmune disease patients suffer from SSI.

The chi-squared test also gives a p-value less than 0.05, indicating that autoimmune disease and SSI are not independent, so we therefore include autoimmune disease as a predictor in our model.

	non-infected	infected
No Autoimmune	12546	92
Autoimmune	5	3

Table 8: Count of Autoimmune Diseases VS SSI

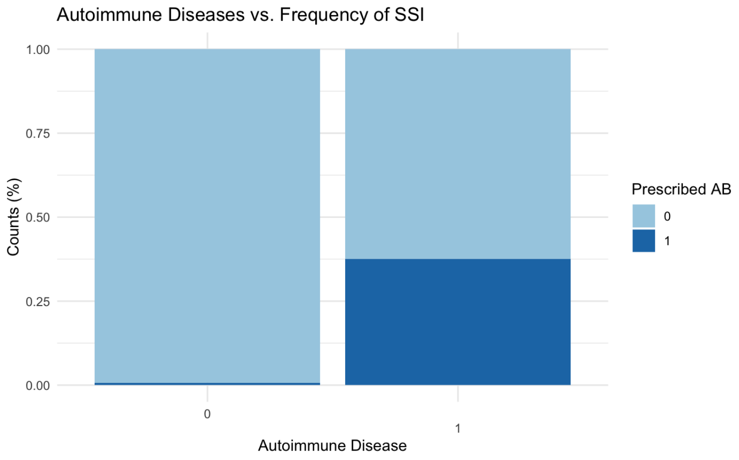


Figure 11: Bar plot of Autoimmune Diseases against SSI

Pearson's Chi-squared test with Yates' continuity correction
data: Autoimmune.disease and antibiotics_after_procedure_less_thirty
X-squared = 99.87, df = 1, p-value < 2.2e-16

4.11 Predictor: Comorbidity: Immunocompromised

Figure 12 shows that patients with immunocompromised history in records have obviously higher proportion of SSI than non-immunocompromised patients. About 7.4% of immunocompromised patients have SSI, while only 0.7% non-immunocompromised patients suffer from SSI.

The chi-squared test also gives a p-value less than 0.05, indicating that immunocompromised and SSI are not independent, so we therefore include immunocompromised disease as a predictor in our model.

	non-infected	infected
No immunocompromised	12464	88
Immunocompromised	87	7

Table 9: Count of Immunocompromised VS SSI

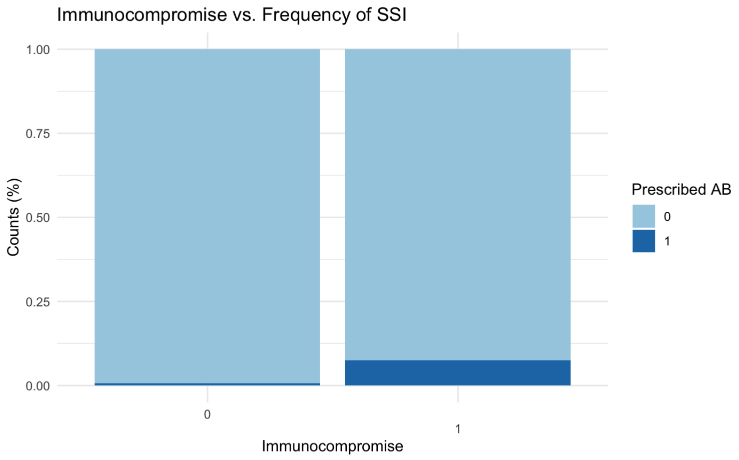


Figure 12: Bar plot of Immunocompromised against SSI

Pearson's Chi-squared test with Yates' continuity correction
data: Immunocompromised and antibiotics_after_procedure_less_thirty
X-squared = 48.256, df = 1, p-value = 3.741e-12

4.12 Predictor: Tobacco Use

Figure 13 shows that patients using tobacco in records have obviously higher proportion of SSI than non-tobacco user. About 25% of tobacco users have SSI, while only 0.7% non-tobacco users suffer from SSI.

The chi-squared test also gives a p-value less than 0.05, indicating that tobacco usage and SSI are not independent, so we therefore include tobacco use disease as a predictor in our model.

	non-infected	infected
No tobacco use	12545	93
Tobacco use	6	2

Table 10: Count of Tobacco Use VS SSI

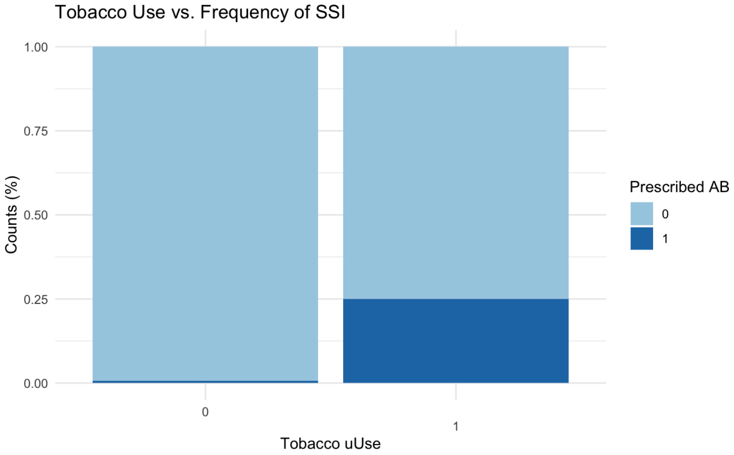


Figure 13: Bar plot of Tobacco Use against SSI

Pearson's Chi-squared test with Yates' continuity correction
data: Tobacco.use and antibiotics_after_procedure_less_thirty
X-squared = 34.782, df = 1, p-value = 3.688e-09

5 Firth logistic model

5.1 Introduction to Firth logistic model

Logistic regression model is the go-to choice for classification. We want to see how our variables are able to predict SSI or not. Moreover, since the number of infections and non-infections is imbalanced, we need to use statistical model that could handle this, and Firth logistic model is the regression model that could be used to deal with imbalanced dataset. Note the base level for procedure site (category2) is eyelid.

5.2 Model Results

Table 11: Firth Logistic Regression Model Results

Variable	Estimate	Lower 95%	Upper 95%
category2Genitourinary/Anus	0.275	0.0002	0.2018
category2Ears/Nose/Mouth/Lips	3.6778	0.3638	498.1516
category2Chest/abdomen	3.8891	0.5114	449.2261
category2Lower Extremity	0.1823	0.0168	24.8471
category2Head/Face/neck	1.0748	0.1269	140.5691
category2Upper extremity	2.5138	0.3427	320.5328
category2Other/unspecified	0.3111	0.0241	43.5326
Diabetes	3.1100	1.2891	7.2808
Immunocompromise	7.3078	2.3126	20.3457
Tabacco.use	2.9326	0.4942	13.2438
Vascular.disease	3.4726	0.8501	13.0851
multiple_encounters	0.5265	0.2906	0.9038
mohs	0.0362	0.0208	0.0618
elix_vw_score	1.0120	0.9974	1.0262
preventive_antibiotics	2.6431	1.6162	4.2597

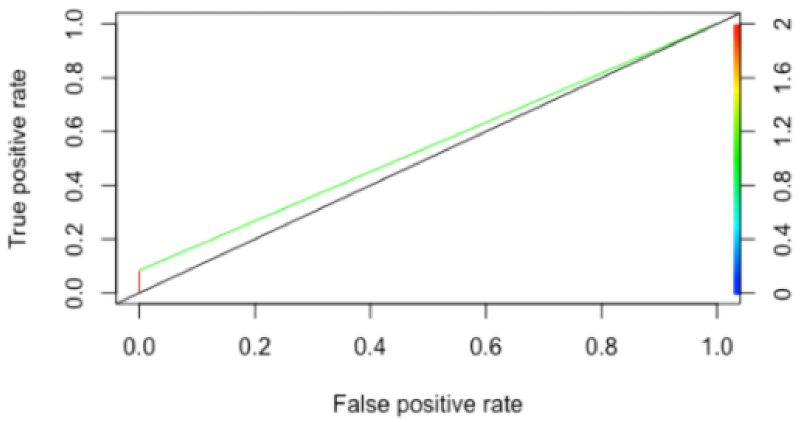


Figure 14: ROC curve for firth logistic regression model

Table 12: Confusion matrix for firth regression

Prediction	Not Infected	Infected
Not Infected	12550	87
Infected	1	8

5.3 Interpretation of results

1. According to figure 14 We found that our model does not do a good job of isolating the features that are important for predicting a positive infection rate. Although, in the confusion

matrix in table 12, the precision is high, we believe that this is due to lack of data points instead of good choice of model

2. According to table 11 all of the operation sites are statistically significant since the 95% interval did not contain 0.

3. The coefficients of this model show the importance of each categorical feature. By ranking the coefficients of the Firth logistic regression model, we find that immunocompromise, Vascular.disease, Diabetes, and Tobacco.use are the four most important features. immunocompromise is naturally a good indicator because infection happens when immune systems could not prevail the bacteria. The Vascular.disease and Diabetes are two variables we could not change before the surgery. However, we could suggest patients not smoking for three months before the surgery, so that the possibility of infection of surgery would decrease dramatically

6 Random Intercept model

6.1 Introduction to Random Intercept model

Because the number of positive cases in our data set was alarmingly low and did not yield significant results in the Firth logistics regression model, we decided to build a second model, taking patients who have had multiple encounters into account. The random intercept model preserves all patients, enlarging our data set from over 12,000 observation to over 24,000. It assumes a different intercept for the random effects – that which is caused by different patients in the model. More specifically, the model takes the following form, and the base for operation site is Chest/Abdomen.

$$S\vec{S}I = \beta_{\overrightarrow{patients}} + \boldsymbol{\beta_{Other\ Variables}}\boldsymbol{X} + \vec{\epsilon} \tag{1}$$

where the $\beta_{\overrightarrow{patients}}$ is the random intercept.

6.2 Model Results

Table 13: Random intercept result

Variable	Estimate	Std	z_value	p_value
(Intercept)	-6.435969			
category2Ears/Nose/Mouth/Lips/Pharynx	3.308074	0.488823	6.767	0.0000000000131
category2Eyelid	2.188364	1.133580	1.930	0.053546
category2Genitourinary/Anus	3.356364	0.754785	4.447	0.0000087167120
category2Head/face/neck (unspecified)	2.728026	0.450494	6.056	0.0000000013987
category2Lower extremity	1.773079	0.522572	3.393	0.000691
category2Other/unspecified	-0.293143	0.828709	-0.354	0.723538
category2Upper extremity	0.248419	0.496832	0.500	0.617071
mohs	-3.606491	0.256372	-14.067	0
elix_vw_score	0.011278	0.006141	1.837	0.066261
Immunocompromise	0.785173	0.765569	1.026	0.305077
Autoimmune	1.646292	0.930078	1.770	0.076718
Vascular	-11.665921	268.915742	-0.043	0.965398
Diabetes	-11.632465	219.568853	-0.053	0.957749
preventive_antibiotics	1.365316	0.224829	6.073	0.0000000012578

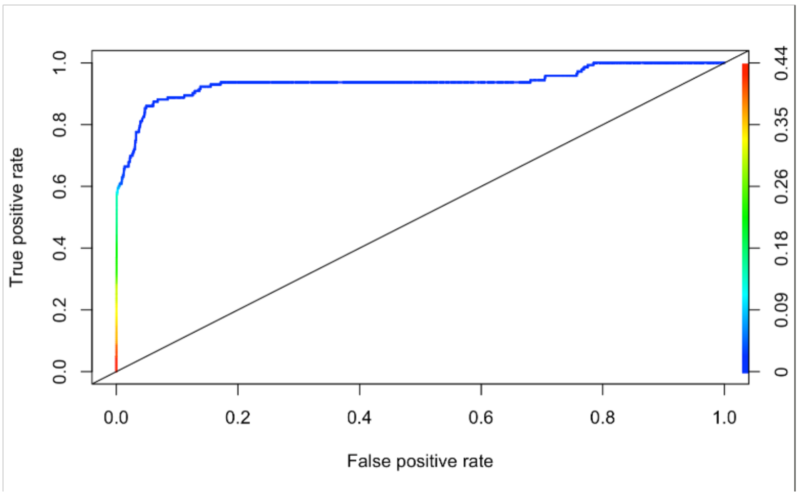


Figure 15: ROC curve

Table 14: Confusion matrix for predicted results of random intercept model

Prediction	Not Infected	Infected
Not Infected	85	54
Infected	58	24200

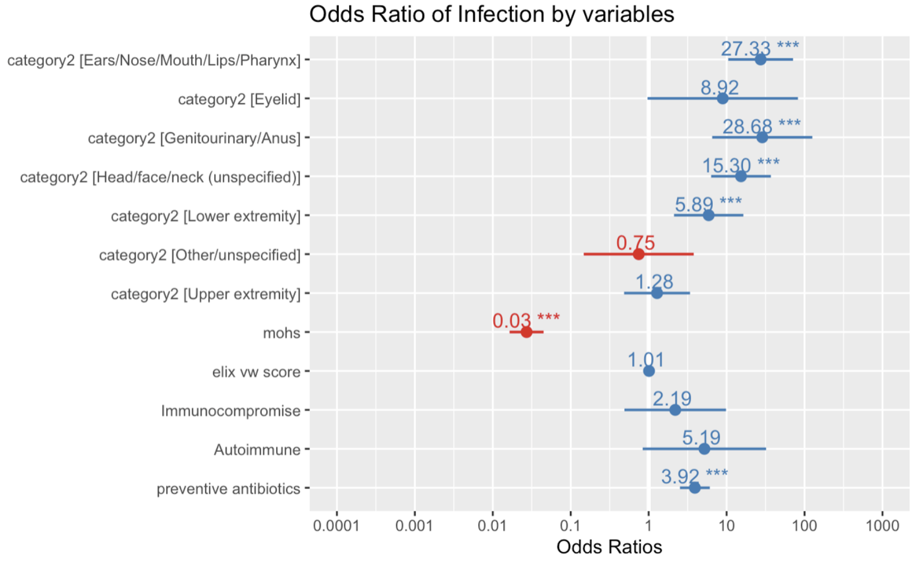


Figure 16: Odds Ratio for random intercept model

6.3 Interpretation of results

1. According to the p-value in table 13, all operating sites are significantly more risky, except for *Eyelid, Other/unspecified, upper extremity.*

2. From figure 16, we conclude that if the doctor prescribes the preventative antibiotics, then the patients have higher risk of getting infected. However, this is not to say that the preventative antibiotics lead to higher infection rates, because doctors could be prescribing the medicine because the patients were originally more vulnerable to post-op infections. Additionally, the odds ratio plot tells us that if the patient had MOHS procedure, they would have lower risks of getting an infection. However, we do not have information on the size of the MOHS procedure, which handicaps our interpretation of the result.

3. Based on figure 14, we have 0.9954 accuracy. However, even though we add more positive cases into our dataset, we still have very imbalanced dataset. Therefore, the accuracy score did not reflect the model performance. Instead, we used the precision (defined as $\frac{TP}{TP+FP}$) to see whether the model would give accurate prediction of the infection. The Precision score tells us how accurate our true positive predictions were. The score for this model is 0.61, which is moderate but not practically significant enough to be used to predict whether a patients would be infected post-surgery. The recall rate, which shows how much of the true positive cases we captured, is also at 0.59, which is also not practically significant.

4. The ROC curve, displayed in figure 15 shows us that the model does have strong predictive power and is an improvement from the Firth logistic regression model.

7 Conclusion

1. Comparing the model results of random intercept model with Firth logistic regression model, we find that random intercept model has a better ROC curve and generally performed better than Firth regression model. We reason that the improved performance is due to the enlarged data set and the introduction of random intercept for each patient.

2. In both Firth logistic regression model and Random intercept model, preventive_antibiotics and MOHS variables are statistically significant, so we conclude that they are good indicators for infection. Although we could not quantitatively predict whether the patient would be infected with those two variables, we could draw the conclusion about the causal relation between those two variables and infection. More specifically, if a patient gets the antibiotics before surgery, this patient would have higher infection rate. If a patient has a MOHS surgery, this patient would have higher probability of getting infection.

3. In both Random intercept model and the Firth logistic, some of the operation sites are statistically significant. This results matches the medical observations because we know that some of tissues are naturally have higher probability of infection. Therefore, we will be able to rank infection risk of different operation site and give higher dose of anti-biotics when the operation location is more likely to be infected.

7.1 Recommendation

1. The unbalanced data would make many statistics model invalid, so we suggest to record more infection data in the dataset.

2. Based on Dr Jason's recommendation, we should extract the comorbidity related data from *Elixhauser.csv.* If a patient is not listed in the "Elixhauser.csc" file, it means that they did not have any of the listed comorbidities, so should be denoted as 0 for all the comorbidity variables. In this way, the comorbidity variable statistics will match with other population statistics. Here is the more detailed way of encoding the variables:

   - Diabetes: elix_dm=1 OR elix_dmcx=1
   - Vascular disease: elix_perivasc=1
   - Autoimmune disease: elix_arth=1
   - Immunocompromise: elix_aids=1

3. Based on Dr Jason's recommendation, tobacco use variable can be further improved: use the tobacco_user column from *Social_history.csv.* Keep the last record for each patient if we are using the last encounter dataset. Then we would encode entries of "yes", "quit", or "passive" as 1, and encode "never" to be 0

4. Dr Jason has an updated list of ICD-9 codes for skin lesion which could help us eliminate less data (having less NA)

5. Use feature engineering technique to generate more variables which will match medical interpretation, and incorporate those variables in previous model

6. use more advanced model like neural network classifier to see if the predictive power increased.

7.2 Limitations

1. Due to limited size of positive infection cases, the model tend to put cases into non-infection to boost the accuracy score, which in turn lowers the predictive power of the model

2. The data set contains lots of NA data and we had to filter out those NA after merging data set. Thus, some of useful information were filtered out in the data clearning process

3. After filtering, the data set is extremely imbalanced, so statistical model would acquire insights of those non-infection people instead of infection people.