

Identifying predictors of postoperative blood-related health complications

JSC370H1 - Final Project

Ryan Shi

Contents

1	Introduction	2
2	Methods	2
2.1	Variables	2
2.2	Acquiring the datasets	3
2.3	Preparing the datasets	3
2.4	Missing values	3
2.5	Outliers and implausible values	4
2.6	Data exploration tools	4
3	Results	5
3.1	Exploratory data analysis	5
3.2	Regression analysis	9
3.3	Machine learning models	10
4	Summary	10
5	Works cited	11

1 Introduction

Health complications following surgery, or postoperative health complications, often pose problems to treatment efficacy and contribute to higher healthcare costs (Dencker et al. 2021). Potential factors for them include healthcare-associated infections or hospital-acquired infections (HAI), which are common in hospitals and can increase morbidity and mortality in patients (Kanerva et al. 2008). Other factors that may influence treatment include staff vaccination rates (Hollmeyer et al. 2012) and emergency department (ED) volume (Brar et al. 2013). Identifying the specific infections and factors that correlate with the greatest increase in postoperative health complications will inform initiatives of hospitals to improve their quality and efficiency of care.

For 5424 hospitals registered with Medicare in the U.S., the Centers for Medicare & Medicaid Services have collected data regarding HAI, ED volume, staff vaccination rates, and complications and mortality rates.

- The Healthcare Associated Infections - Hospital dataset (dataset 1) records the number of cases in each hospital for six common sources of infections: central venous catheters (central lines), urinary tract catheters, surgical site infection from colon surgery, surgical site infection from abdominal hysterectomy, Methicillin-resistant *Staphylococcus aureus* (*S. aureus* or staph) bacteria, and *Clostridium difficile* (*C. diff*) bacteria. The values are cumulative from 04/01/2022 to 03/31/2023.
- The Timely and Effective Care - Hospital dataset (dataset 2) includes various metrics, such as the average time patients spent in ED and rates of septic shock. The metrics of interest here are the ED volume and staff vaccination rates for COVID and the flu, all per hospital. The values are cumulative from 01/01/2022 to 12/31/2022.
- The Complications and Deaths - Hospital dataset (dataset 3) records various rates of complications and rates of death per hospital, such as for post-surgery wound dehiscence, post-surgery respiratory failure, and collapsed lung. The values are cumulative from 07/01/2020 to 06/30/2022.

The question to be explored is, “Can healthcare-acquired blood infections, ED volumes, and/or staff vaccination rates reliably predict postoperative blood-related complication rates in Medicare hospitals?” My current hypothesis is that a subset of the variables can reliably predict complication rates after surgery. In addition, I hypothesize that all but vaccination rates positively correlate with postoperative blood-related complications. This is since infections would increase the risk of complications occurring and higher ED volumes would lead to more sources of contamination, while increased vaccination might reduce transmissions of infections from staff to patients. I also aim to investigate how these relations vary across states.

2 Methods

2.1 Variables

For sources of healthcare-acquired blood infections, which is a predictor in my question, I chose to look at central lines and the staph bacteria since they affect the bloodstream directly. I did not look at the other sources of infections recorded by the datasets since they do not affect the bloodstream directly, instead implicating parts of the abdomen such as the colon and urinary tract.

For postoperative blood-related complication rates, which is the outcome in my question, I chose to look at hemorrhage/hematoma events (loss of blood from damaged blood vessels), serious blood clots, and bloodstream infections since they are common complications and affect blood cells and vessels. I did not look at the other complications recorded by the datasets since they do not relate to blood, instead implicating the respiratory system and epithelial tissues.

Thus, I will use the following variables for my analysis:

Variable	Definition	Dataset
Predictors		
central_line	Percent of infections from central lines	Dataset 1
staph	Percent of infections from the staph bacteria	Dataset 1
ed_vol	Level of ED volume (very high, high, medium, or low)	Dataset 2
covid_vac	Percent of hospital staff vaccinated against COVID	Dataset 2
flu_vac	Percent of staff vaccinated against the flu	Dataset 2
Outcomes		
hem	Percent of postoperative hemorrhage/hematoma events	Dataset 3
clot	Percent of postoperative serious blood clots	Dataset 3
stream	Percent of postoperative bloodstream infections	Dataset 3

2.2 Acquiring the datasets

I acquired the datasets by extracting their links to their CSV files from the network activity of their pages and directly loading those files. This was because I did not manage to extract the data from the JSON file which the API provided. Dataset 1 has 173232 rows, dataset 2 has 115498 rows, and dataset 3 has 91428 rows. All three datasets has 20 columns and are in long format, where each hospitals has multiple rows for different measures (e.g., rates of infections). For each of them, I excluded uninformative or redundant columns such as facility name, address, phone number, and measure name, while I kept columns such as facility ID, state, measure ID, and the value for the measure. I did not encounter import issues with the datasets, except that I had to convert some columns from character to numeric form.

2.3 Preparing the datasets

For dataset 1, I kept rows measuring the number of central line and staph infections, which are predictors. I reshaped the table such that each measure occupied a separate column, and I calculated the percents of infection using the counts of infection and the number of operations involving the source of infection per hospital, which are provided in the dataset. I renamed the predictors as `central_line` and `staph` respectively, and selected their columns along with the facility ID and state.

For dataset 2, I kept rows measuring ED volumes and percents of vaccination among staff, which are predictors. Since the dataset already classified the ED volumes into four levels, I was not able to obtain the exact counts of the volumes. I reshaped the table, renamed the predictors as `ed_vol`, `covid_vac`, and `flu_vac` respectively, and selected their columns along with the facility ID and state.

For dataset 3, I kept rows measuring percents of postoperative hemorrhage/hematoma, serious blood clots, and bloodstream infection, which are the outcomes. I also calculated the counts corresponding to these percents using the total number of operations for each hospital, which will be used in the regression analysis. I reshaped the table, renamed the outcomes as `hem`, `clot`, and `stream` respectively, and selected their columns along with the facility ID and state.

2.4 Missing values

Before joining the datasets, I replaced all values of “Not Available” with NA, and I removed all rows which had no values recorded for any measures. I then investigated the proportion of missing values for each dataset:

- In dataset 1, the rates of central line and staph infections are 9.26% and 5.21% missing respectively.
- In dataset 2, the ED volumes and rates of vaccinations are 11.3%, 15.5%, and 4.45% missing respectively.

- In dataset 3, the rates of postoperative hemorrhage, blood clots, and bloodstream infection are 1.70%, 0.00%, and 13.2% respectively.

Since I did not deem these proportions of missing values to be significant, I joined the datasets and removed all rows with missing values, leaving 2248 rows with 16 columns in wide format.

2.5 Outliers and implausible values

Looking at the distribution of each variable, I did not find outliers or implausible values for `staph`, `ed_vol`, `covid_vac`, and `flu_vac`. For the other variables:

- `central_line`: Most values are between 0% and 35%. I removed 3 outliers with values greater than 50%.
- `hem`: Most values are between 1% and 4%. I removed 1 outlier with value greater than 6%.
- `clot`: Most values are between 1% and 6%. I removed 1 outlier with value greater than 7%.
- `stream`: Most values are between 1% and 10%. I removed 1 outlier with value greater than 12%.

For `central_line` and `staph`, 37.8% and 44.2% of their points have a value of 0%. These points may skew my overall results, but since they make up a large proportion of the values, I did not remove them entirely. Instead, I only removed them temporarily in some exploratory graphs for clarity.

2.6 Data exploration tools

For data exploration, I created:

- A table of the numerical variables' summary statistics using `kable` and `kable_styling` from `kableExtra`;
- A heatmap of pairwise correlations between the numerical variables. I used `melt` from the `reshape2` package on the correlation matrix, then plotted and colored the values using `geom_tile` and `geom_text` from `ggplot2`;
- 3 pairs of choropleth maps showing the means of `staph` and `stream`, `covid_vac` vs `clot`, and `flu_vac` vs `hem`. I selected these pairs since according to the previous heatmap, they have the largest correlations of any pair of predictors and outcomes. (This is explained in the next section.) I first obtained a dataset with the coordinates of state boundaries from the `maps` package, to which I had to add abbreviations for each state. I then merged it with my main dataset and plotted it using `geom_polygon` from `ggplot2`;
- 3 boxplots showing the distributions of the outcomes grouped by ED volume. I used `geom_boxplot` from `ggplot2` and `plot_layout` from the `patchwork` package;
- 3 scatterplots showing the relations between the log-transformed outcomes vs `flu_vac`, `covid_vac`, and `staph`, grouping by ED volume. I transformed the outcomes since I suspected heteroskedasticity in their values. As with the choropleth maps, I selected these pairs of variables since they have the largest correlations of any pair of predictors and outcomes in the heatmap. I used `geom_point` and `geom_smooth` from `ggplot2` and `plot_layout` from the `patchwork` package;
- Summaries of ANOVA F-tests (not shown) for all predictors and outcomes. I used `aov` from the `stats` package and `kable_styling` from `kableExtra`;
- 3 tables showing summaries of regression models for the following formulae: `hem ~ flu_vac`, `clot ~ central_line + staph + covid_vac`, and `stream ~ central_line + staph`. For each dependent variable, I selected these predictors since they have the most significant P-values in the ANOVA analysis with the same dependent variable. I used `lm` from the `stats` package and `kable_styling` from `kableExtra`.

3 Results

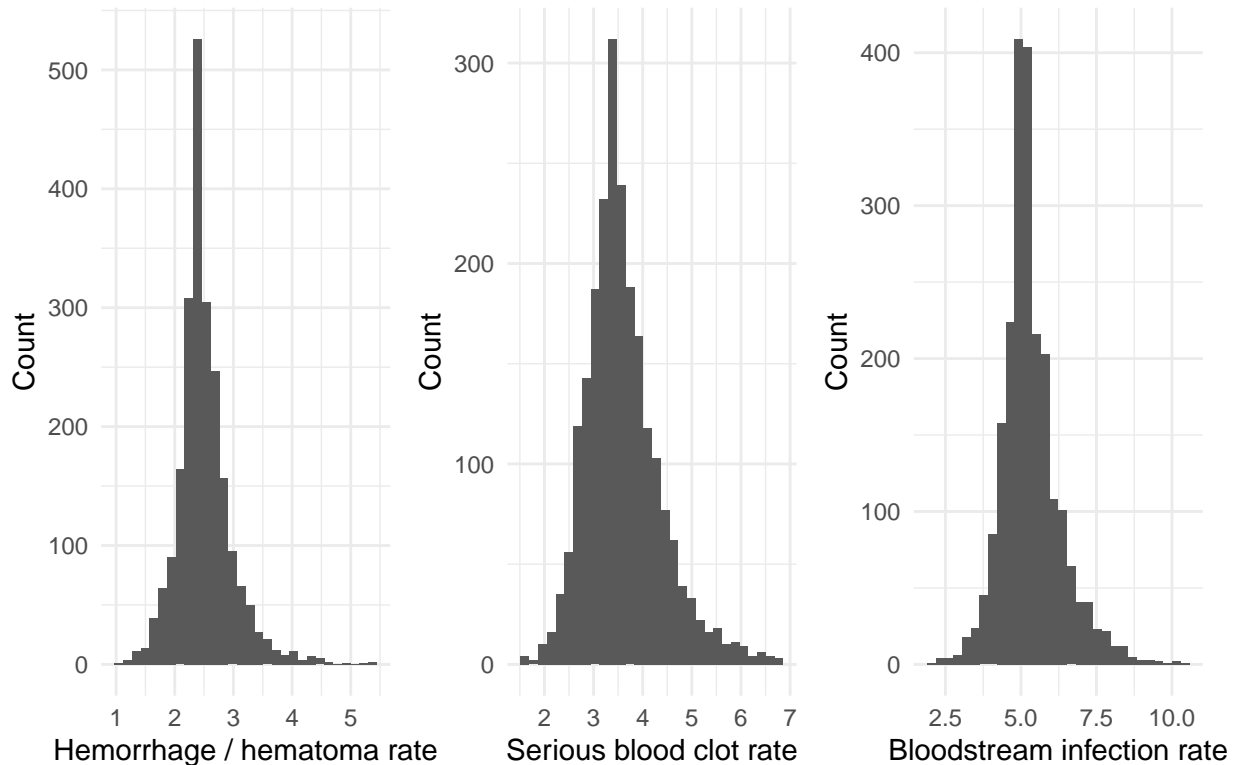
3.1 Exploratory data analysis

Table 2: Summary of the numerical variables (rates)

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Predictors						
Central line infection	0.00	0.0000	0.0500	0.0592	0.0900	0.4594
Staph infection	0.00	0.0000	0.0035	0.0043	0.0065	0.0286
COVID vaccination	32.60	84.1000	91.0000	88.7775	96.2000	100.0000
Flu vaccination	4.00	68.0000	84.0000	78.7110	94.0000	100.0000
Outcomes						
Hemorrhage/hematoma	1.10	2.2600	2.4400	2.5029	2.7000	5.4300
Serious blood clots	1.61	3.1100	3.4800	3.5987	3.9900	6.7800
Bloodstream infection	2.17	4.7725	5.1500	5.3224	5.7700	10.5800

Table 1 shows that central line and staph infections are very skewed towards 0%, meaning they rarely occur in hospitals. Indeed, most values are 0% (as mentioned before), since the table shows it as both the minimum and 1st quartile value. Both vaccination rates are skewed towards 100%, indicating that hospitals are generally successful in encouraging vaccination among their workers. The last three response variables are all skewed towards 0%, meaning they rarely occur after surgeries. Of the response variables, the rate of postoperative bloodstream infection has the greatest values for all listed metrics, indicating that it is somewhat more common.

Distributions of the outcome variables

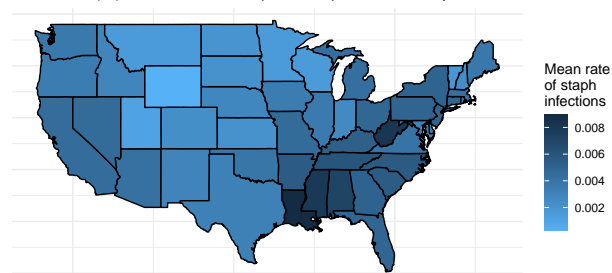


Heatmap of pairwise correlations between numerical variables

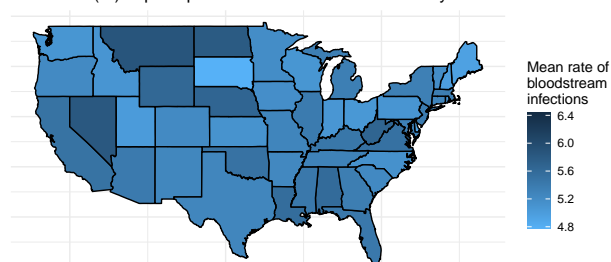


The heatmap above shows that overall, there is little correlation between the predictors (**central_line**, **staph**, **covid_vac**, **flu_vac**) and the outcomes (**hem**, **clot**, **stream**), with most correlations being between 0 and 0.1. Among all pairs of predictors and outcomes, the pairs **staph** and **stream**, **covid_vac** and **clot**, and **flu_vac** and **hem** have the greatest correlations, with values 0.11, 0.11, and 0.12 respectively. This suggests that **staph**, **covid_vac**, and **flu_vac** are likely to be the most significant predictors of these response variables. Interestingly, there appears to be collinearity between both rates of vaccination (correlation = 0.37), suggesting that hospitals tend to encourage multiple vaccinations together for their workers. As well, there appears to be collinearity between **central_line** and **staph** (correlation = 0.28), suggesting that these sources of healthcare-acquired infections are likely to occur together.

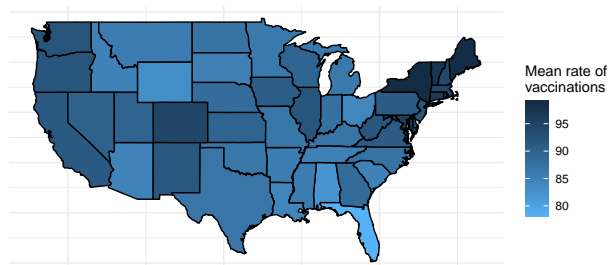
Mean rate (%) of healthcare-acquired staph infections by state



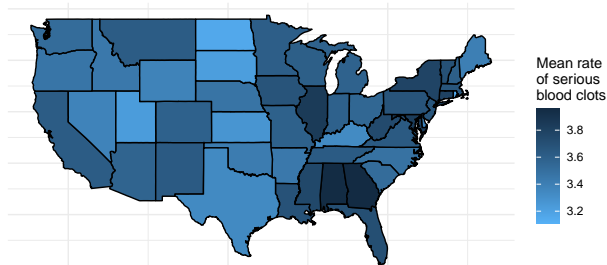
Mean rate (%) of postoperative bloodstream infections by state



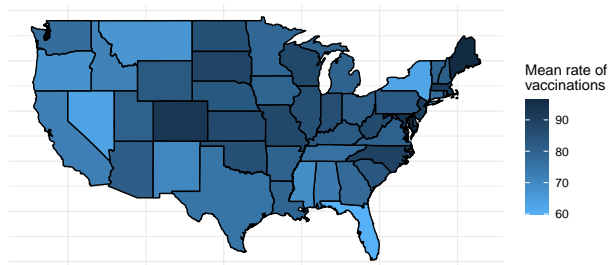
Mean rate (%) of staff COVID vaccinations by state



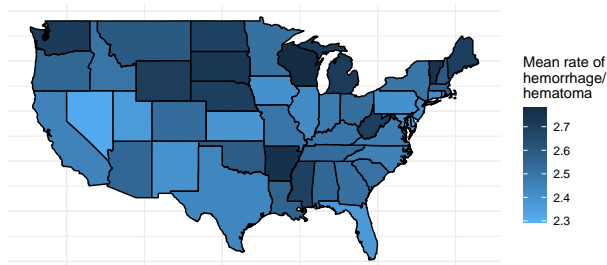
Mean rate (%) of postoperative serious blood clots by state



Mean rate (%) of staff flu vaccinations by state



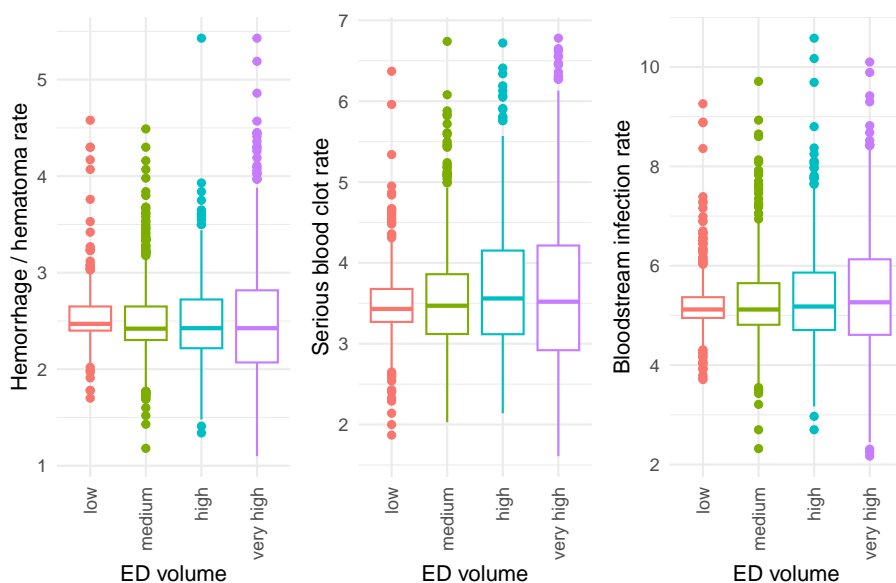
Mean rate (%) of postoperative hemorrhage/hematoma by state



For the three pairs of maps above, they show that the corresponding pairs of variables may have negative relationships when comparing between states. This somewhat contradicts the positive correlations of these variables shown in the heatmap. For the first pair, states with high mean rates of staph infections, such as West Virginia and Louisiana, have comparatively lower mean rates of postoperative bloodstream infections. Similarly, for the second pair, states with higher mean rates of staff COVID vaccinations, such as Georgia and Alabama, have comparatively lower mean rates of postoperative serious blood clotting. Finally, for the third pair this trend is the most pronounced, where multiple northern states have relatively low rates of staff flu vaccinations and relatively high rates of postoperative hemorrhage/hematoma. This inverse trend is more intuitive for the latter two pairs of variables, since hospitals with a high proportion of vaccinated workers are likely to be more careful when handling potential sources of health hazards.

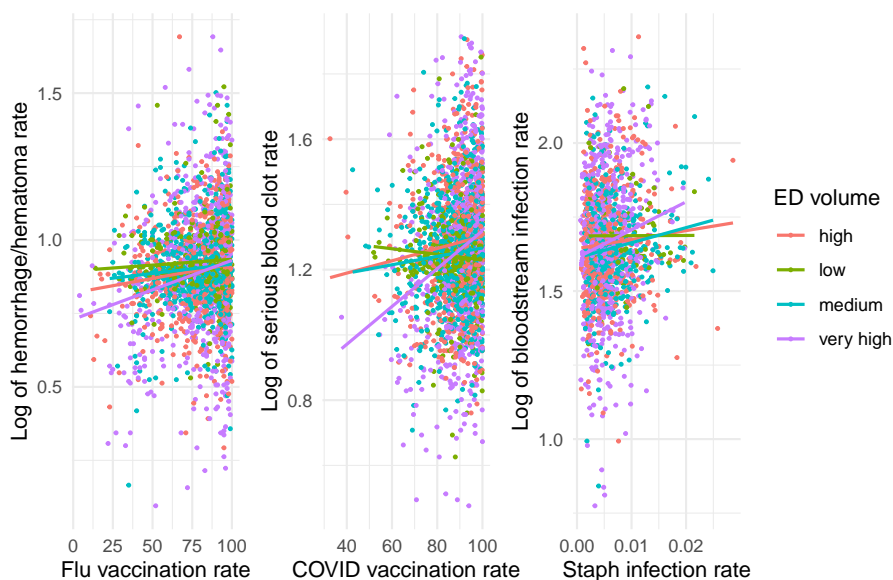
Due to the inconclusive results of the previous visualizations, the subsequent visualizations now investigate if ED volume plays any role in the main question.

Outcome variables by ED volume



For all three outcomes, the above boxplots show that the distributions become wider and more dispersed as ED volume increases. This suggests that hospitals with higher ED volumes may be more variable in their attention to health hazards, which makes sense given that their workers are also likely more overworked.

Log of selected response variables vs selected predictors by ED volume



According to these scatterplots, the relations for all three pairs of variables are more significant in the 'very high' category for ED volume. In other words, among all the regression lines, the lines for 'very high' are the most different from a horizontal line, which would indicate no correlation. Nevertheless, the data points still appear to be overly dispersed after the log-transformation of the outcomes, suggesting that:

- The relations between variables are minor even in the 'very high' category for ED volume.

- The transformations do not yield significant changes in the relations between the variables.

As such, the outcomes are not transformed for the linear models, which we now turn our attention to.

3.2 Regression analysis

The ANOVA test for `hem ~ central_line + staph + ed_vol + covid_vac + flu_vac` shows the most significant predictor as `flu_vac`, which has a P-value of $1.1e-6 < 0.01$. As such, it will be the only predictor for the model for `hem`.

Table 3: Regression summary for *hem* vs *flu_vac*

	Estimate	Std. error	t-value	P-value
(Intercept)	2.269596	0.0432237	52.508124	0
flu_vac	0.002964	0.0005350	5.540107	0

In table 2, the summary for the model `hem ~ flu_vac` shows that for every increase of 1% in the rate of staff vaccinated against the flu, the rate of postoperative hemorrhage/hematoma increases by 0.00296%. The corresponding P-value is $0 < 0.01$, indicating that the predictor is significant. However, the adjusted R-squared value (not shown) is 0.01301, indicating that the model explains very little of the variation in the data.

The ANOVA test for `clot ~ central_line + staph + ed_vol + covid_vac + flu_vac` shows the most significant predictors as `central_line`, `staph`, and `covid_vac`, which have P-values of $3.3e-4 < 0.01$, $6.2e-4 < 0.01$, and $3e-7 < 0.01$ respectively. As such, they will be the predictors for the model for `clot`.

Table 4: Regression summary for *clot* vs *central_line*, *staph*, and *covid_vac*

	Estimate	Std. error	t-value	P-value
(Intercept)	2.7351365	0.1479541	18.486391	0.0000000
central_line	0.7316521	0.2739148	2.671094	0.0076152
staph	11.5611923	3.8219703	3.024930	0.0025152
covid_vac	0.0086839	0.0016459	5.275987	0.0000001

In table 3, the summary for the model `clot ~ central_line + staph + covid_vac` shows that:

- When `staph` and `covid_vac` are held constant, for every increase of 1% in the rate of staph infections, the rate of postoperative serious blood clotting increases by 0.57%. The corresponding P-value is $0.04 > 0.01$, indicating that the predictor is moderately significant.
- When `central_line` and `covid_vac` are held constant, for every increase of 1% in the rate of staph infections, the rate of postoperative serious blood clotting increases by 13.0%. The corresponding P-value is $4.1e-4 < 0.01$, indicating that the predictor is significant.
- When `central_line` and `staph` are held constant, for every increase of 1% in the rate of staff vaccinated against COVID, the rate of postoperative serious blood clotting increases by 0.0092%. The corresponding P-value is $1e-7 < 0.01$, indicating that the predictor is significant.

Of these predictors, `staph` appears to be the most significant one. However, the adjusted R-squared value (not shown) is 0.02018, which is somewhat higher than the value for the previous model summary, but still extremely low. This also indicates that the model explains very little of the variation in the data.

The ANOVA test `stream ~ central_line + staph + ed_vol + covid_vac + flu_vac` shows the most significant predictors as `central_line` and `staph`, which have P-values of $6.5e-4 < 0.01$ and $1e-5 < 0.01$ respectively. As such, they will be the only predictors for the model for `clot`.

Table 5: Regression summary for *clot* vs *central_line* and *staph*

	Estimate	Std. error	t-value	P-value
(Intercept)	5.1900802	0.0328751	157.872687	0.0000000
central_line	0.5541988	0.3533574	1.568380	0.1169337
staph	23.3249746	4.9354929	4.725967	0.0000024

In table 4, the summary for the model `stream ~ central_line + staph` shows that:

- When `staph` is held constant, for every increase of 1% in the rate of central line infections, the rate of postoperative blood infection increases by 0.72%. The corresponding P-value is $0.04 > 0.01$, indicating that the predictor is moderately significant.
- When `central_line` is held constant, for every increase of 1% in the rate of central line infections, the rate of postoperative blood infection increases by 21%. The corresponding P-value is $1e-5 < 0.01$, indicating that the predictor is significant.

Of these two predictors, `staph` appears to be the more significant one. However, the adjusted R-squared value is 0.01279 (not shown), which is extremely low. Again, the model explains very little of the variation in the data.

3.3 Machine learning models

4 Summary

Overall, the analysis showed that the relations between the predictors and outcomes of interest are mostly weak. This outcome continued to hold after accounting for states, transforming the response variables, and performing linear regression. In particular, the heatmap indicated generally low correlations between the variables, while the non-linear patterns in the scatterplots and adjusted R-squared values of the regression summaries suggest that the data is poorly described by linear models. In addition, the apparent negative relations between variables at the state level somewhat contradict the positive correlations in the heatmap, furthering confounding the analysis. Therefore, my hypothesis for the main question has little support currently, and further analysis is required to answer the question more conclusively.

There are many potential reasons for this less-than-ideal result, including unaddressed confounding factors. Importantly, an issue with the datasets used is that each dataset has different time periods of data collection. This means the values of interest are aggregated across different periods and may not be exactly comparable between datasets. Here, I have been assuming that the values for each hospital are similar between different time periods. However, this is a rather naive assumption and may have been a major limitation of my analysis.

Despite the insignificant findings, for two of the fitted models, the rate of healthcare-acquired staph infections had relatively large estimated regression coefficients (13% and 19%), meaning that it is likely to have a larger effect on the outcomes compared to other predictors and thus worthy of further investigation. Moreover, methods other than linear regression may prove more suitable for the data, and they will be explored in the continuation of this analysis.

5 Works cited

Brar S, McAlister FA, Youngson E, Rowe BH. 2013. Do Outcomes for Patients With Heart Failure Vary by Emergency Department Volume? *Circulation: Heart Failure*. 6(6): 1147–1154.

Dencker EE, Bonde A, Troelsen A, Varadarajan KM, Sillesen M. 2021. Postoperative complications: an observational study of trends in the United States from 2012 to 2018. *BMC Surgery*. 21(1): 1-10.

Hollmeyer H, Hayden F, Mounts A, Buchholz U. 2012. Review: interventions to increase influenza vaccination among healthcare workers in hospitals. *Influenza and Other Respiratory Viruses*. 7(4): 604–621.

Kanerva M, Ollgren J, Virtanen MJ, Lyytikäinen O. 2008. Risk factors for death in a cohort of patients with and without healthcare-associated infections in Finnish acute care hospitals. *Journal of Hospital Infection*. 70(4): 353-360.