JSC370H1 - Final

Ryan Shi

Contents

1 Introduction

Healthcare-associated infections or healthcare-acquired infections (HAI) are common in hospitals and can increase morbidity and mortality in patients as well as induce health complications (Kanerva et al. 2008). Other factors that may influence the treatment of patients include staff vaccination rates (Hollmeyer et al. 2012) and emergency department (ED) volume (Brar et al. 2013). Identifying the specific infections and other factors that correlate with the greatest increase in health complications and morbidity 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 occurrence of complications 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

2.0.1 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.

2.0.2 Preparing the datasets

For dataset 1, I kept rows measuring the number of central line and staph infections, which are independent variables. This is because the other sources of infections recorded by the database implicate parts of the

abdomen such as the colon and urinary tract, rather than affecting the bloodstream directly. I reshaped the table such that each measure occupies 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 selected these columns along with the facility ID and state and renamed them as follows:

- `central_line`: percent of infections from central lines;
- `staph`: percent of infections from the staph bacteria.

For dataset 2, I kept rows measuring ED volumes and rates of vaccination, which are independent variables. Note that since the table already classifies the ED volumes into four levels, I was not able to obtain the exact counts of the volumes. I selected the columns with ED volumes and vaccination rates along with the facility ID and state and renamed them as follows:

- `ed_vol`: level of ED volume as *very high*, *high*, *medium*, or *low*;
- `covid_vac`: percent of staff vaccinated against COVID;
- `flu_vac`: percent of staff vaccinated against the flu.

For dataset 3, I kept rows measuring percents of postoperative hemorrhage/hematoma, serious blood clots, and bloodstream infection, which are the dependent variables. I reshaped the table, selected the rate columns along with the facility ID and state, and renamed them as follows:

- `hem`: percent of postoperative hemorrhage/hematoma;
- `clot`: percent of postoperative serious blood clots;
- `stream`: percent of postoperative bloodstream infection.

2.0.3 Missing values

I replaced all values of "Not Available" with NA, and I removed all rows which had no values recorded for all measures in each dataset. I then investigated the proportion of missing values, which are as follows:

- 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 2261 rows with 10 columns in wide format.

2.0.4 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 for some data exploration parts.

2.0.5 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 independent and dependent variables. (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 dependent variables 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 dependent variables vs `flu_vac`, `covid_vac`, and `staph`, grouping by ED volume. I transformed the dependent variables 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 independent and dependent variables in the heatmap. I used `geom_point` and `geom_smooth` from ggplot2 and `plot_layout` from the patchwork package;
- 3 tables showing summaries of ANOVA F-tests for the following formulae: `hem ~ central_line + staph + ed_vol + covid_vac + flu_vac`, `clot ~ central_line + staph + ed_vol + covid_vac + flu_vac`, and `stream ~ central_line + staph + ed_vol + covid_vac + flu_vac`. I used `aov` from the stats package and `kable_styling` from kableExtra;
- 3 tables showing summaries of linear 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 independent variables 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.

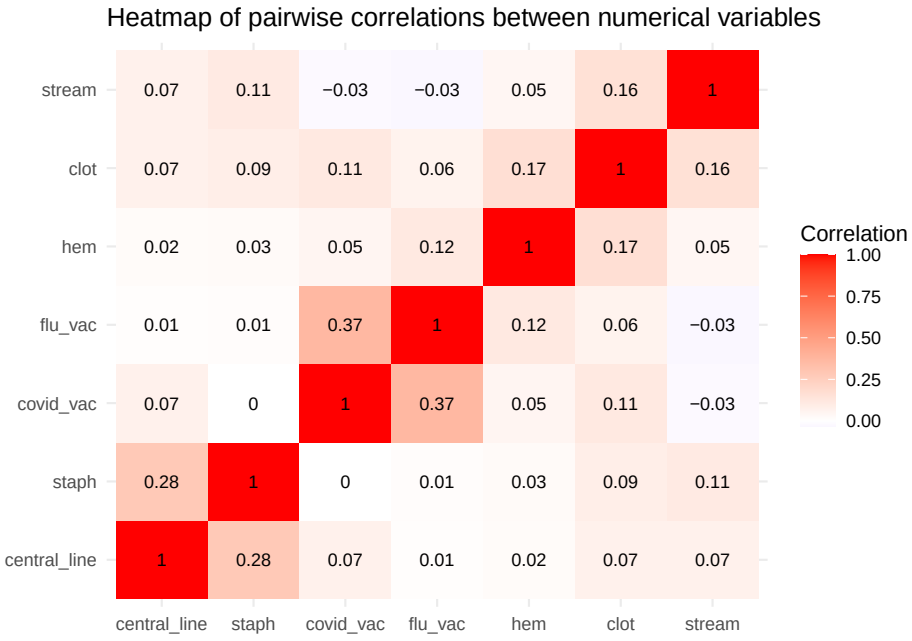
3 Preliminary results

Table 1: Summary of numerical variables

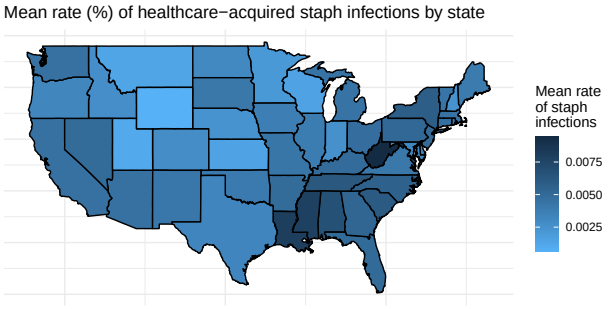
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Central line infection	0.00	0.00	0.0516	0.0606	0.0916	0.4139
Staph infection	0.00	0.00	0.0037	0.0044	0.0068	0.0315
COVID vaccinations	34.70	84.60	91.3000	89.1463	96.2000	100.0000
Flu vaccination	4.00	68.00	84.0000	78.7818	94.0000	100.0000
Hemorrhage/hematoma	1.10	2.26	2.4400	2.5033	2.7000	5.4300
Serious blood clots	1.61	3.11	3.4800	3.6016	3.9900	6.7800
Bloodstream infection	2.17	4.78	5.1500	5.3241	5.7700	10.5800

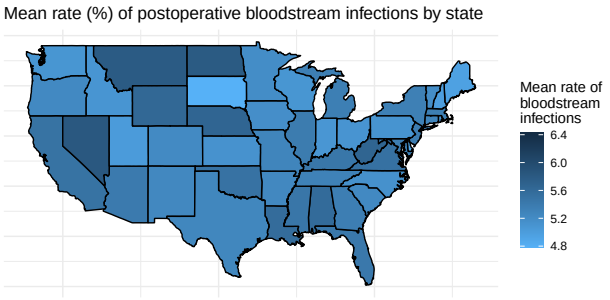
Table 1 shows that central line and staph infections are very skewed towards 0%, meaning they rarely occur in hospitals. Indeed, most values are 0% (as mentioned before), since the table shows it as both the minimum and 1st quartile value. Both vaccination rates are skewed towards 100%, indicating that hospitals are generally successful in encouraging vaccination among their workers. The last three response variables

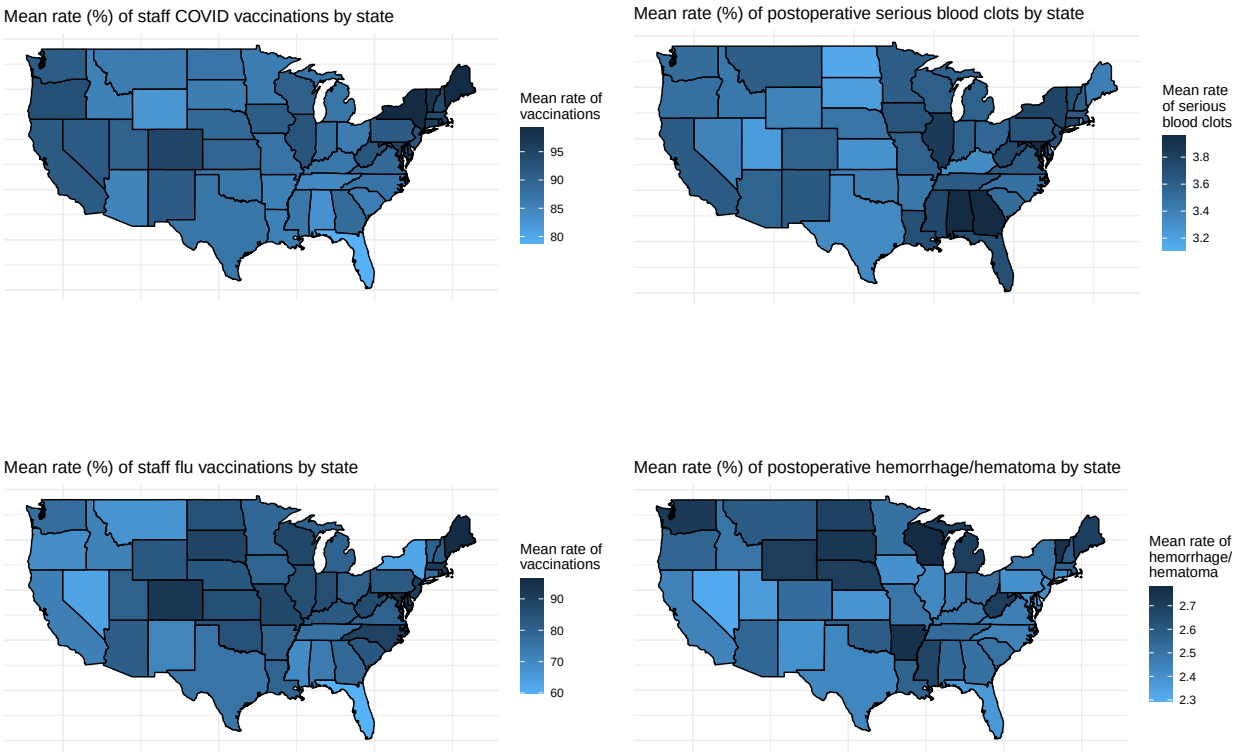
are all skewed towards 0%, meaning they rarely occur after surgeries. Of the response variables, the rate of postoperative bloodstream infection has the greatest values for all listed metrics, indicating that it is somewhat more common.



The heatmap above shows that overall, there is little correlation between the independent variables (`central_line`, `staph`, `covid_vac`, `flu_vac`) and the dependent variables (`hem`, `clot`, `stream`), with most correlations being between 0 and 0.1. Among all pairs of independent and dependent variables, 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.

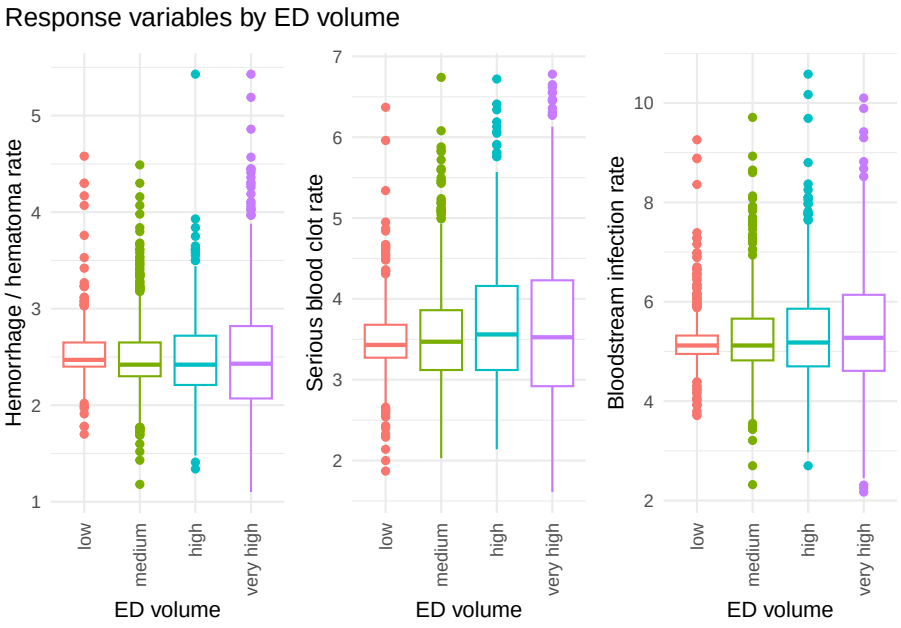




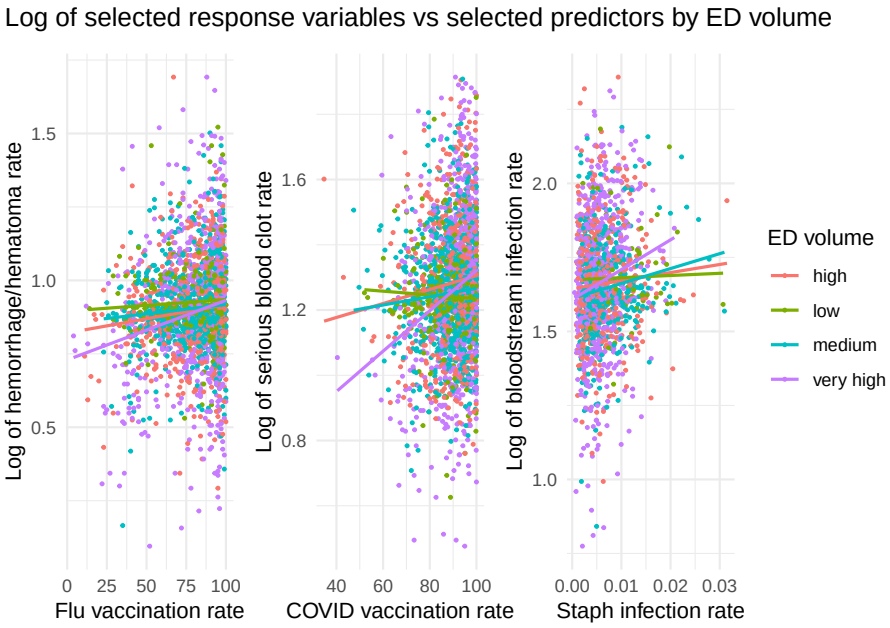


For the three pairs of maps above, they show that the corresponding pairs of variables may have negative relationships when comparing between states. This somewhat contradicts the positive correlations of these variables shown in the heatmap. For the first pair, states with high mean rates of staph infections, 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.



For all three response variables, 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.



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 dependent variables, 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 dependent variables are not transformed for the linear models, which we now turn our attention to.

Table 2: ANOVA test for $hem \sim central\_line + staph + ed\_vol + covid\_vac + flu\_vac$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
central_line	1	0.2670160	0.2670160	1.257467	0.2622509
staph	1	0.2168162	0.2168162	1.021059	0.3123767
ed_vol	3	1.1126790	0.3708930	1.746658	0.1553736
covid_vac	1	1.2812907	1.2812907	6.034024	0.0141079
flu_vac	1	5.0681161	5.0681161	23.867443	0.0000011
Residuals	2247	477.1377038	0.2123443	NA	NA

The ANOVA test in table 2 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)	2.2700659	0.0431976	52.550783	0
flu_vac	0.0029611	0.0005343	5.542045	0

In table 3, 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.

Table 4: ANOVA test for $clot \sim central\_line + staph + ed\_vol + covid\_vac + flu\_vac$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
central_line	1	7.3552720	7.3552720	12.87805	0.0003396
staph	1	6.7102226	6.7102226	11.74866	0.0006198
ed_vol	3	6.0908286	2.0302762	3.55473	0.0138322
covid_vac	1	15.2559870	15.2559870	26.71110	0.0000003
flu_vac	1	0.9204049	0.9204049	1.61150	0.2044125
Residuals	2247	1283.3692093	0.5711478	NA	NA

The ANOVA test in table 4 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 5: Regression summary for *clot* vs *central_line*, *staph*, and *covid_vac*

	Estimate	Std. error	t-value	P-value
(Intercept)	2.6932787	0.1549329	17.383522	0.0000000
central_line	0.5704941	0.2773376	2.057039	0.0397975
staph	13.0289436	3.6820737	3.538480	0.0004106
covid_vac	0.0091537	0.0017250	5.306489	0.0000001

In table 5, 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.

Table 6: ANOVA test for $stream \sim central\_line + staph + ed\_vol + covid\_vac + flu\_vac$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
central_line	1	11.105265	11.1052655	11.664549	0.0006483
staph	1	18.656022	18.6560224	19.595576	0.0000100
ed_vol	3	5.477241	1.8257472	1.917695	0.1245888
covid_vac	1	2.059464	2.0594640	2.163183	0.1414917
flu_vac	1	1.361406	1.3614063	1.429969	0.2318956
Residuals	2247	2139.262539	0.9520528	NA	NA

The ANOVA test in table 6 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 7: Regression summary for *clot* vs *central_line* and *staph*

	Estimate	Std. error	t-value	P-value
(Intercept)	5.1870552	0.0328323	157.986452	0.0000000
central_line	0.7249712	0.3567936	2.031906	0.0422803
staph	21.0074310	4.7502078	4.422424	0.0000102

In table 7, 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.

4 Summary

Overall, the analysis showed that the relations between the independent and dependent variables 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.

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.