Identifying predictors of postoperative blood-related health complications

JSC370H1 - Final Project

Ryan Shi

Contents

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 hospital-acquired infections or healthcare-associated 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 hospital-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 hospital-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 values are 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 Tools used

For data exploration, I created:

- A table showing the summary statistics of the numeric variables using `kable` and `kable_styling` from kableExtra;
- Six histograms of the outcomes as rates and counts using `geom_histogram` from ggplot2 and `plot_layout` from patchwork;
- A heatmap of pairwise correlations between the numeric variables using `melt` from reshape2 and `geom_tile` and `geom_text` from ggplot2;
- Three pairs of choropleth maps showing the means of selected numeric variables by state using `plot_geo` from plot_ly;
- Three boxplots of the outcomes grouped by ED volume using `geom_boxplot` from ggplot2 and `plot_layout` from patchwork;
- Three scatterplots of the log-transformed outcomes vs `flu_vac`, `covid_vac`, and `staph` grouped by ED volume using `geom_point` and `geom_smooth` from ggplot2 and `plot_layout` from patchwork.

For regression analysis, I performed Poisson and negative binomial regression using `glm` from the stats package and `glm.nb` from MASS.

For the machine learning models, I created bagging, random forest, and XGBoost models using `randomForest` from randomForest and `train` from caret. For the bagging models, I set the 'mtry' hyperparameter in `randomForest` to 5, which is the number of predictors for each outcome. For the XGBoost models, I performed grid search on various hyperparameters in `train`, such as 1, 3, 5, and 7 for 'max_depth' and 0.001, 0.01, 0.1, and 0.2 for 'eta'.

3 Results

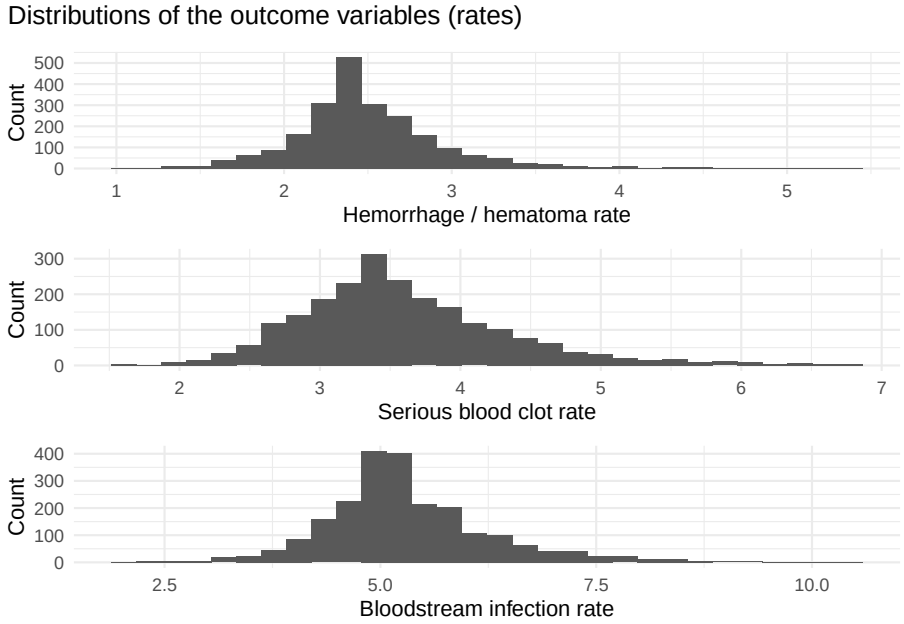
3.1 Exploratory data analysis

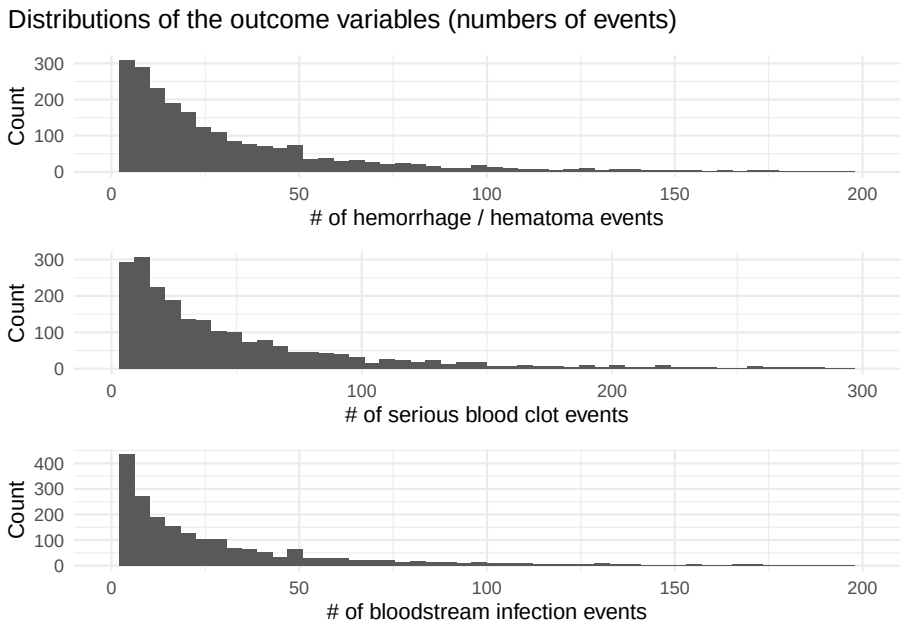
Table 2: Summary of the numeric variables (rates)

	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

First, I explore the distributions of the numeric variables. Table 2 shows that the rates of central line and staph infections are very skewed towards 0%, meaning these infections rarely occur in hospitals. This agrees with my previous finding that 37.8% and 44.2% of `central_line` and `staph` respectively are 0%. Both vaccination rates are skewed towards 100%, indicating that hospitals are generally successful in encouraging vaccination among their workers. Finally, the outcome variables are all skewed towards 0%, meaning post-operative blood-related complications are generally uncommon. Of these variables, the rate of postoperative bloodstream infections is the largest across all listed metrics, indicating that it is relatively more common than the other outcomes.

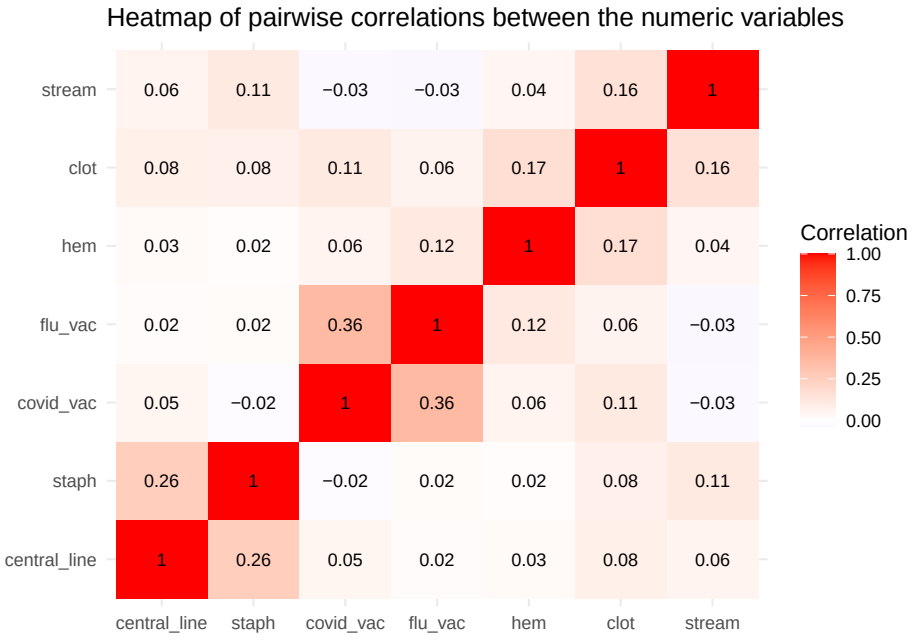
To further explore their distributions, I now look at the outcome variables both as rates and as counts of postoperative complications.





From the above histograms, the rates appear to be normally distributed with slightly longer right tails, while the counts of events appear to be Poisson distributed with no obvious upper bounds. This suggests that compared to linear regression, Poisson or negative binomial regression might be more suitable to model postoperative complications, which will be performed later on.

Next, I explore the pairwise correlations between the numeric 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 coefficients being between 0 and 0.1. Among all pairs of predictors and outcomes, the pairs `flu_vac` and `hem`, `covid_vac` and `clot`, and `staph` and `stream` have the greatest coefficients, with values 0.12, 0.11, and 0.11 respectively. This suggests

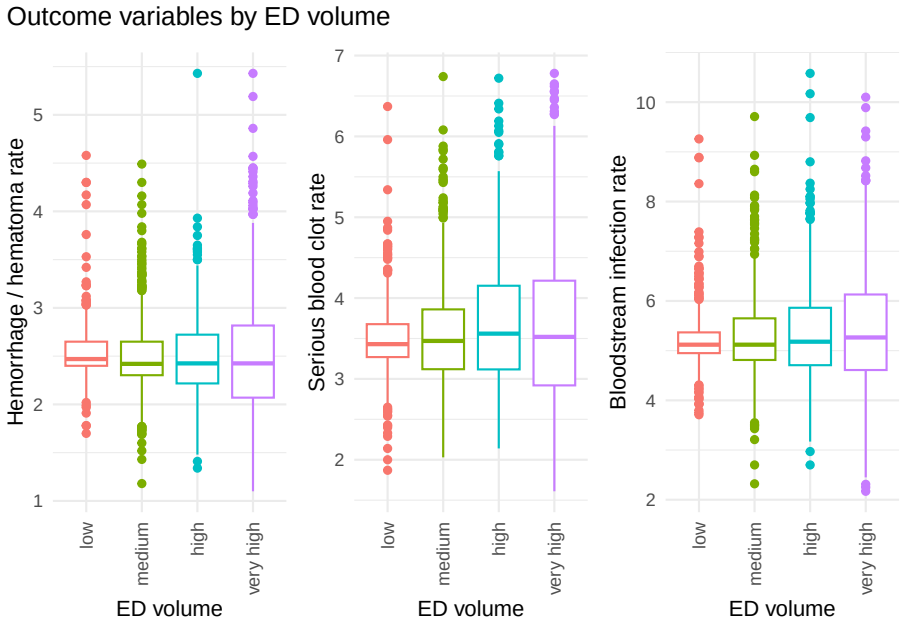
that `flu_vac`, `covid_vac`, and `staph` are likely to be the most significant predictors of these response variables. Interestingly, with a coefficient of 0.37, there appears to be collinearity between both rates of vaccination, suggesting that hospital staff tend to be vaccinated against multiple pathogens at the same time. As well, with a coefficient of 0.28, there appears to be collinearity between `central_line` and `staph`, suggesting that these healthcare-acquired infections likely occur together.

Since `flu_vac` and `hem`, `covid_vac` and `clot`, and `staph` and `stream` appear to be correlated in the heatmap, I now explore their relationships across states using the three pairs of cloropleth maps (shown here):

- For the first pair of maps, `flu_vac` and `hem` appear to have a slight negative relationship across states. Several states with lower rates of flu vaccinations among hospital staff, such as Arkansas and Wiscousin, have higher rates of postoperative hemorrhage / hematoma. Conversely, several states with higher rates of flu vaccinations among hospital staff, such as Maine and Maryland, have lower rates of postoperative hemorrhage / hematoma.
- For the second pair of maps, `covid_vac` and `clot` appear to have a negative relationship across states. Notably, states in New England have high rates of COVID vaccination among hospital staff and low rates of postoperative serious blood clots, while multiple southern states have low rates of COVID vaccination among hospital staff and high rates of postoperative serious blood clots.
- For the third pair of maps, `staph` and `stream` appear to have a slight negative relationship across states. Several states with high mean rates of hospital-acquired staph infections, such as West Virginia and Louisiana, have low mean rates of postoperative bloodstream infections. However, the relationship between the variables is weak since rates of postoperative bloodstream infections appear to be roughly uniform across states.

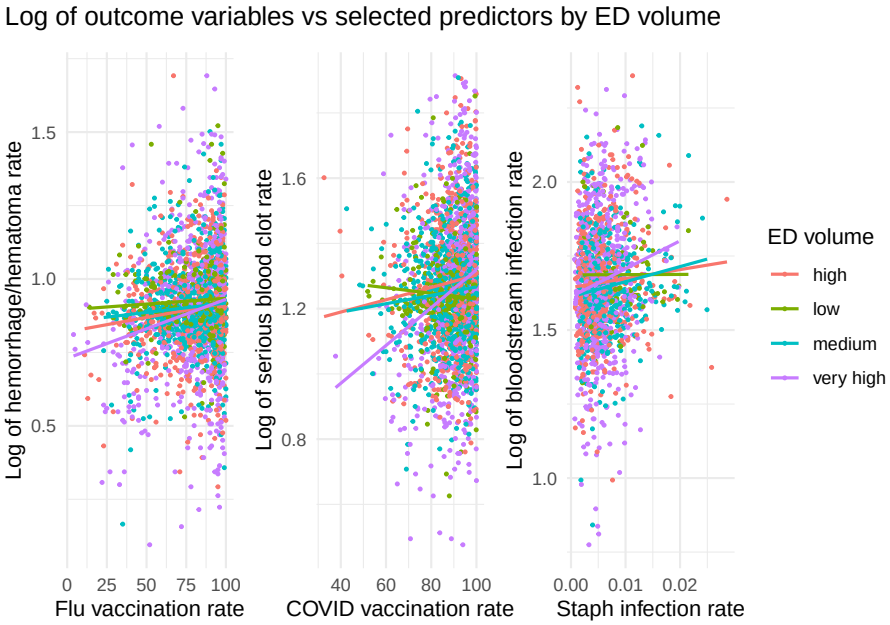
Overall, the pairs of variables appear to have negative relationships across states, which is intriguing since in the heatmap, they are positively correlated without grouping by state. This is more intuitive for the first two pairs of variables, since hospitals with a high proportion of vaccinated workers are likely more careful when handling potential sources of health hazards. Thus, states with these hospitals are likely to have fewer postoperative complications overall.

Next, I explore the relationship between ED volume and the numeric variables.



For all three outcomes, the above boxplots show that the distributions become wider and more dispersed as ED volume increases. This makes sense since hospitals with higher ED volumes might have more potential sources of infection. As well, during periods of high ED volumes, hospital workers might be more variable in their attention to health hazards since they likely to be busier and more overworked.

Since I suspect heteroskedasticity in the outcomes when grouping by ED volume, I apply a logarithmic transformation to the outcomes for the following scatterplots. As before, I choose to explore the relationships between `flu_vac` and `hem`, `covid_vac` and `clot`, and `staph` and `stream` since they have the largest correlations of any pair of predictors and outcomes in the heatmap.



In the above scatterplots, the relationships for all three pairs of variables are positive for all ED volumes, which agrees with the positive correlations in the heatmap. Moreover, these relationships are the most significant when ED volume is 'very high', since the regression lines for the group are the steepest. Nevertheless, these relationships are weak overall since the transformed outcomes still appear to be overly dispersed, which means that the transformations likely do not yield significant changes in the relationships between the variables. Thus, the outcomes will not be transformed for the subsequent analyses.

3.2 Regression analysis

I will use generalized linear models (e.g., Poisson regression) since the outcomes are expressed in rates, which are calculated using counts of postoperative complications out of the total number of operations for each hospital. Thus, the response variables in the models will be the counts, while the offset terms will be the total number of operations for each hospital.

First, I perform Poisson regression on `hem_count ~ central_line + staph + ed_vol + covid_vac + flu_vac + offset(log(total))`, where `hem_count` is the number of postoperative hemorrhage / hematoma events for each hospital and `log(total)` is the offset. However, this model is a poor fit to the data since its standardized residuals (not shown) appear to be overdispersed. In addition, `flu_vac` is the only significant predictor in the model, with a corresponding P-value of 5.49e-16 < 0.001. Thus, instead of Poisson regression, I perform negative binomial regression on `hem_count ~ flu_vac + offset(log(total))`.

Table 3: Negative binomial regression summary for *hem* vs *flu_vac*

	Estimate	Std. error	z-value	P-value
(Intercept)	-3.8198012	0.0258873	-147.555010	0e+00
flu_vac	0.0017121	0.0003170	5.400499	1e-07

In the above table, the summary shows that for every increase of 1% in the rate of flu vaccinations among staff in a hospital, the logarithm of the expected number of postoperative hemorrhage / hematoma events increases by 0.0017. In other words, the expected number of postoperative hemorrhage / hematoma events increases by a factor of exp(0.0017) = 1.0017. The P-value of the predictor is 1e-07 < 0.001, indicating significance. The AIC of the current model (13534) is lower than the AIC of the corresponding Poisson model (15357), indicating the current model better fits the data.

Next, I perform Poisson regression on `clot_count ~ central_line + staph + ed_vol + covid_vac + flu_vac + offset(log(total))`, where `clot_count` is the number of postoperative serious blood clot events for each hospital and `log(total)` is the offset. However, as before, the standardized residuals of the model appear to be overdispersed. In addition, `covid_vac` is the only significant predictor in the model, with a corresponding P-value of 2e-16 < 0.001. Thus, instead of Poisson regression, I perform negative binomial regression on `clot_count ~ covid_vac + offset(log(total))`.

Table 4: Negative binomial regression summary for *clot* vs *covid_vac*

	Estimate	Std. error	z-value	P-value
(Intercept)	-3.6201166	0.0524825	-68.977564	0
covid_vac	0.0033493	0.0005843	5.731728	0

In the above table, the summary shows that for every increase of 1% in the rate of COVID vaccinations among staff in a hospital, the logarithm of the expected number of postoperative serious blood clot events increases by 0.0033. In other words, the expected number of postoperative serious blood clot events increases by a factor of exp(0.0033) = 1.0033. The P-value of the predictor is 0 < 0.001, indicating significance. The AIC of the current model (15427) is lower than the AIC of the corresponding Poisson model (19051), indicating the current model better fits the data.

Finally, for the same reasons as before, I perform negative binomial regression on `stream_count ~ staph + offset(log(total))`, where `stream_count` is the number of postoperative bloodstream infection events for each hospital, `staph` is the only significant predictor in the corresponding Poisson model (P-value = 8.27e-06 < 0.001), and `log(total)` is the offset.

Table 5: Negative binomial regression summary for *stream* vs *staph*

	Estimate	Std. error	z-value	P-value
(Intercept)	-2.973257	0.0095291	-312.018112	0e+00
staph	8.069511	1.5284267	5.279619	1e-07

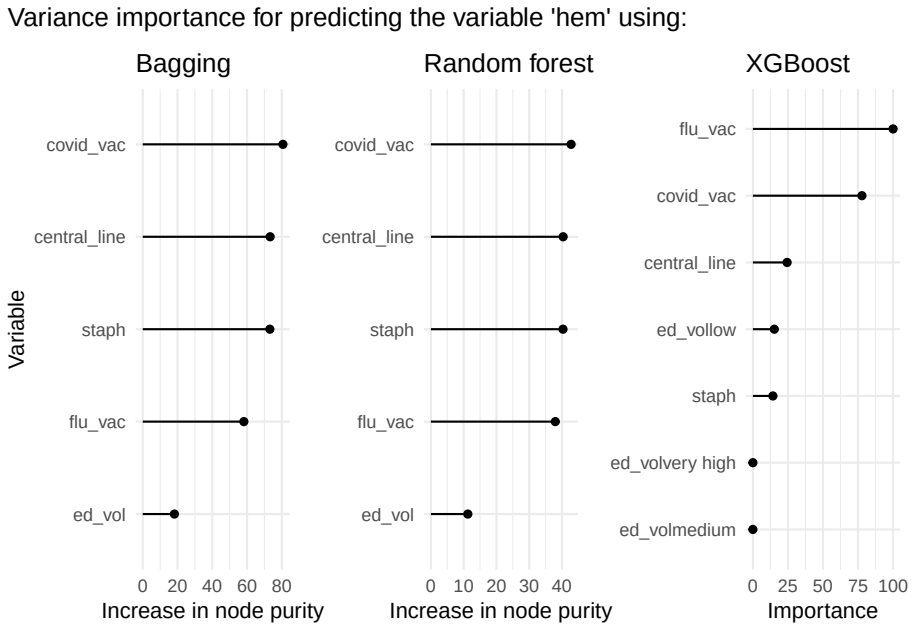
In the above table, the summary shows that for every increase of 1% in the rate of hospital-acquired staph infections in a hospital, the logarithm of the expected number of postoperative bloodstream infections increases by 8.07. In other words, the expected number of postoperative bloodstream infections increases by a factor of exp(8.07) = 3195. This regression coefficient is very large, likely because the values for `staph` in the

data are very close to 0%. The P-value of the predictor is 1e-07 < 0.001, indicating significance. The AIC of the current model (12569) is lower than the AIC of the corresponding Poisson model (14139), indicating the current model better fits the data.

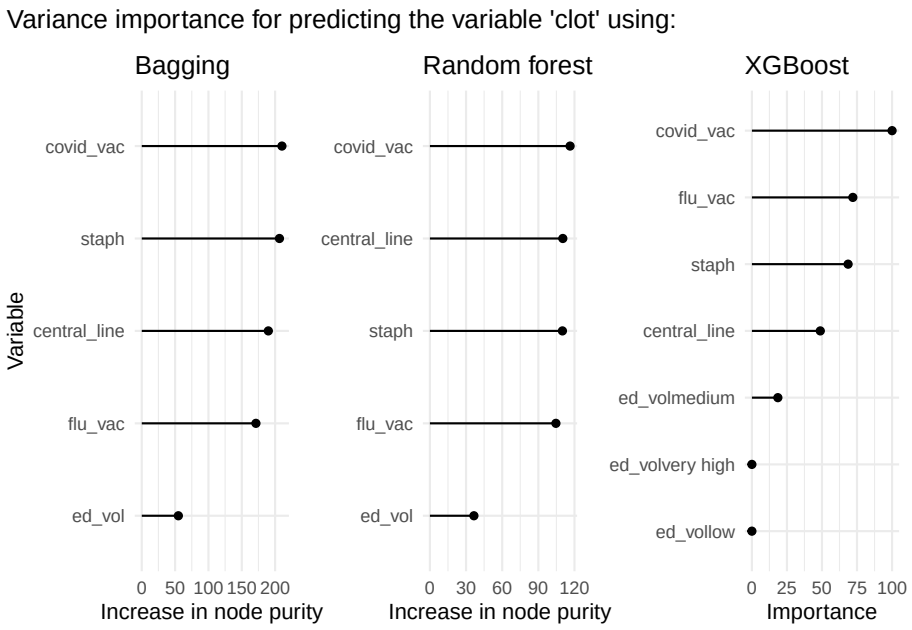
Overall, `flu_vac`, `covid_vac`, and `staph` are relatively significant predictors for `hem`, `clot`, and `stream` respectively, which agrees with the results of my exploratory data analysis.

3.3 Machine learning models

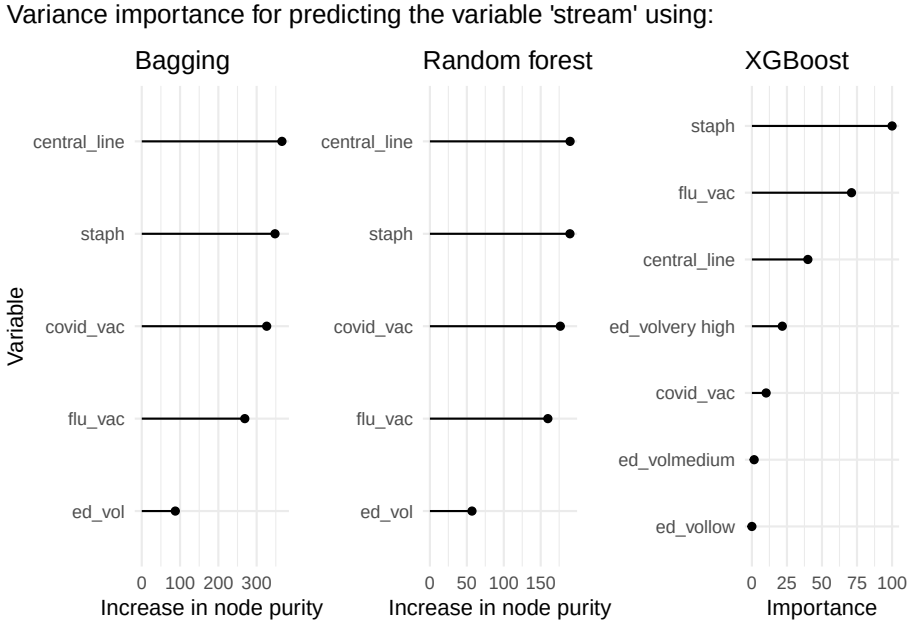
To answer my question using machine learning, I now create bagging, random forest, and XGBoost models for each outcome and explore their variable importance metrics. Specifically, I aim to determine whether XGBoost has better predictive accuracy compared to the other less complex models. To train the models and calculate their test MSEs, I split the data into 70% train and 30% test sets.



The above plot shows that `covid_vac` is the most important variable for predicting `hem` when using bagging and random forest. In fact, `covid_vac`, `central_line`, `staph`, and `flu_vac` all have very similar levels of importance in the random forest model. Importantly, `flu_vac` is the most important variable in the XGBoost model, which agrees with my previous finding that it is a relatively significant predictor for `hem`. Despite this, `flu_vac` has lower importance in the bagging and random forest models. Finally, `ED_vol` is the least important variable for all models, which agrees with the results of my exploratory data analysis.

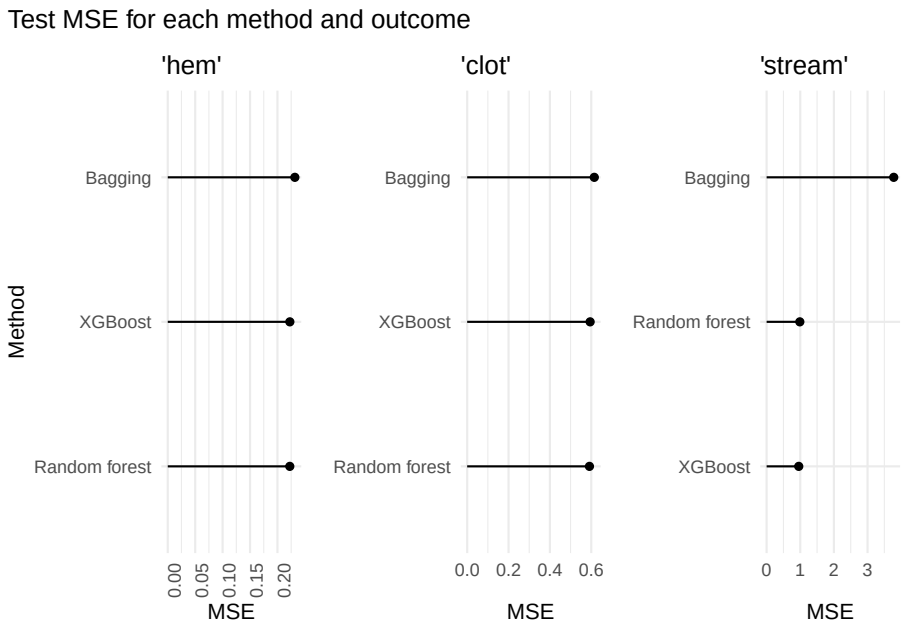


The above plot shows that `covid_vac` is the most important variable for predicting `clot` in all models, which agrees with my previous finding that it is a relatively significant predictor for `clot`. Again, `covid_vac`, `central_line`, `staph`, and `flu_vac` all have very similar levels of importance in the random forest model, while `ED_vol` is the least important variable for all models.



The above plot shows that `central_line` is the most important variable and `staph` is the second most important variable for predicting `stream` when using bagging and random forest. Interestingly, `central_line` and `staph` have very similar levels of importance in the random forest model. Importantly, `staph` is the most important variable in the XGBoost model, which agrees with my previous finding that it is a relatively significant predictor for `stream`. Again, `ED_vol` is the least important variable for all models.

Finally, I now look at the test MSEs of the models.



The above plot shows that the random forest model has the smallest MSE when predicting `hem` and `clot` on the test set, while the XGBoost model has the smallest MSE when predicting `stream` on the test set. In fact, all models have very similar MSEs when predicting `hem` and `clot`. Thus, XGBoost does not perform as well as I thought compared to the other approaches. However, it does perform significantly better than bagging when predicting `stream`, indicating an improvement in accuracy for this outcome.

Overall, random forest performs the best for predicting `hem` and `clot`, while XGBoost performs the best for predicting `stream`. In particular, the most important variables for the XGBoost models agree with the results of my regression analysis, which reinforces their significance.

4 Summary

In summary, I found that rates of vaccinations among hospital staff and hospital-acquired staph infections are able to predict postoperative blood-related complication rates to some extent. Specifically, `flu_vac`, `covid_vac`, and `staph` are the most significant predictors for `hem`, `clot`, and `stream` respectively, or in other words:

- The rate of flu vaccinations among hospital staff is likely related to the rate of postoperative hemorrhage / hematoma events;
- The rate of COVID vaccinations among hospital staff is likely related to the rate of postoperative serious blood clots; and
- The rate of hospital-acquired staph infections is likely related to the rate of postoperative bloodstream infections. This relationship might be relatively strong since it has the largest corresponding regression coefficient.

These relationships are supported by my exploratory data analysis, regression analysis, and the variable importance metrics in the machine learning models. Interestingly, contrary to my initial hypothesis, the rates of vaccinations among hospital staff positively correlate with the outcomes. In addition, this correlation direction appear to be reversed when grouping by state, which makes sense for the rates of vaccinations.

Despite this, my exploratory data analysis indicates that these relationships are likely generally weak. Moreover, I did not find any significant relationships between ED volume and the rest of the variables, even after applying a logarithmic transformation to the outcomes.

There are likely several limitations to my analysis, 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 adversely affected my analysis. To resolve this issue, it would be helpful to be able to look at the trends in the variables across time, such as by day or month. This can be investigated with other datasets to extend this analysis in the future.

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.