MRSA Forecast

Chris Lee

2024-09-15

Forecasting is beginning to be integrated into decision-making processes not only for business operation but also for infectious disease outbreak response.

The demo uses the data from [California Health and Human Services Open Data Portal](https://data.chhs.ca.gov/dataset/methicillin-resistant-staphylococcus-aureus-mrsa-bloodstream-infections-bsi-in-california-hospitals). All California general acute care hospitals are required to report Methicillin-resistant Staphylococcus aureus (MRSA) bloodstream infection (BSI) cases that occur following hospitalization.

MRSA as its name suggested is caused by strains of Staphylococcus aureus develop antibiotic resistance. The bacteria can develop antibiotic resistance naturally in the environment or through defense mechanisms to block or destroy antibiotic drugs. MRSA is contagious. It can spread through skin- to-skin contact or on surfaces [1](https://my.clevelandclinic.org/health/diseases/11633-methicillin-resistant-staphylococcus-aureus-mrsa).

Goal: In this post, our goal is to predict Standardized Infection Ratio (SIR) for each facility in 2023.

Overview of variables used in the post.

## Data Overview

```{r}  
#| label: Overview data  
#| fig-cap: " Distribution of the variables"  
#| code-fold: true  
#| fig-width: 8  
my\_glimpse(data = mrsa\_combine|> select(-Facility\_ID))[-7,]  
```

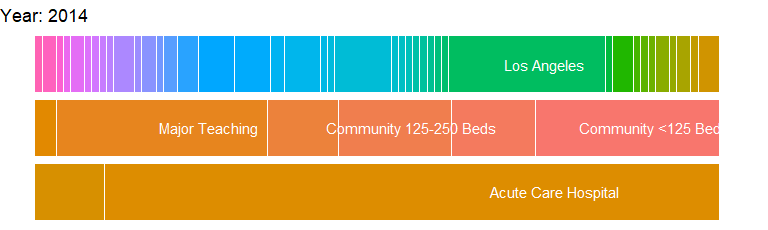
| vars | type\_ | range\_ |
| --- | --- | --- |
| Year | continuous | [2014, 2023] |
| Infections\_Reported | continuous | [NA, NA] |
| Patient\_Days | continuous | [NA, NA] |
| SIR | continuous | [NA, NA] |
| Hospital\_Category\_RiskAdjustment | discrete | Acute Care Hospital、Long-Term Acute Care Hospital、Rehabilitation Hospital or Unit、Critical Access Hospital、NArange5values |
| Hospital\_Type | discrete | Community <125 Beds、Major Teaching、Community 125-250 Beds、Long-Term Acute Care、Pediatric、Community >250 Beds、Free-Standing Rehabilitation、Critical Access、NArange9values |

Distribution of the variables

## Data Visualization

Overall the distribution of County, Hospital\_Category\_RiskAdjustment, and Hospital\_Type didn’t have major changes from 2014 to 2023. However, two new clinic categories were found in the Hospital\_Type after 2016. This could be possible, either there were not many rehabilitation Hospitals or Units or they were not included in the study prior to 2016.

```{r}  
#| label: Overview categorical data  
#| fig-cap: " Distribution of the categorical variables"  
#| warning: false  
#| code-fold: true  
#| fig-width: 8  
#| fig-height: 10  
  
freq\_data <- function(df, year){  
 df <- mrsa\_combine |>   
 filter(Year == year)|>  
   
 mutate(`Hospital Type` = case\_when(Hospital\_Type == "Community, <125 Beds"~ "<125",  
 Hospital\_Type == "Community, 125-250 Beds"~ "125-250",  
 Hospital\_Type == "Community, >250 Beds"~ ">250",  
 .default = Hospital\_Type),  
 `Risk Adjustment` = Hospital\_Category\_RiskAdjustment)|>  
 select(Facility\_ID, `Hospital Type`, `Risk Adjustment`, County)  
   
 df <- as.data.frame(df)  
 results\_list <- vector("list", ncol(df) - 1)  
   
 # Loop through the columns starting from the 2nd column  
 for (i in 2:ncol(df)) {  
 # Store the result of cal\_freq(i) in the list  
 results\_list[[i - 1]] <- cal\_freq(i, df)  
 }  
 df <- do.call(rbind, results\_list)  
  
 # label bar graph  
 df$vars <- ifelse(df$Freqx > 0.15, as.character(df$Var1), "")  
 return(df)  
}  
  
d2014 <- freq\_data(df, 2014)  
d2014 <- freq\_data(df, 2014)  
d2016 <- freq\_data(df, 2016)  
d2019 <- freq\_data(df, 2019)  
d2023 <- freq\_data(df, 2023)  
d2014$Year <- 2014  
d2016$Year <- 2016  
d2019$Year <- 2019  
d2023$Year <- 2023  
  
p <-  
rbind(d2014,d2016,d2019, d2023)|>  
 #filter(Year == 2013)|>  
 #select(-Year)|>  
 ggplot(aes(x = round(Freqx,2), y = fct\_rev(namex), fill = Var1)) +   
 geom\_col(position = "fill", color = "white") +  
 scale\_x\_continuous(labels = label\_percent()) +  
 labs(y = NULL, x = NULL, fill = "var1")+  
 guides(fill="none") +  
 geom\_text(aes(label = vars), position = position\_stack(vjust = 0.75), color = "white")+  
 theme\_void()   
  
p+transition\_time(as.integer(Year))+   
 enter\_fade() +  
 exit\_fade()+labs(title = "Year: {frame\_time}")  
  
#anim\_save(filename="animation.gif", mygif)  
```



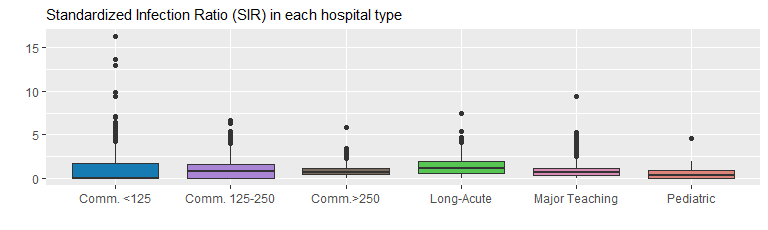
Distribution of the categorical variables

## Distribution of SIR values with respect to Hospital\_Type, 2014-2023

We can see that the average SIR is about the same in all hospital types except the community hospital <125\*. The SIR value in the community hospital <125 ranges from 0 to more than 15, it poses the biggest SIR variation among the hospital types. In addition, We cab use ANOVA analysis to check if the mean SIR was truly the same for the rest of hospital types.

\*Critical Access hospital/unit and free-Standing Rehabilitation are not included in the plot due to incomplete SIR data.

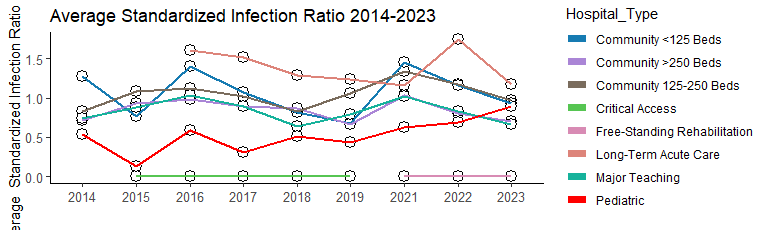
```{r}  
#| label: SIR distribution  
#| fig-cap: " Distribution of SIR values with respect to Hospital\_Type"  
#| code-fold: true  
#| fig-width: 8  
#| fig-height: 10  
#| warning: false  
  
color\_palette <- c("#167bb2","#aa85d6","#7A6C5D","#56c552","#d88bb4","#dd847a","#16b29b","red","#dcda98", "green")  
  
mrsa\_combine|>  
 filter(  
 Hospital\_Type != "Free-Standing Rehabilitation",  
 Hospital\_Type != "Critical Access")|>  
 mutate(Hospital\_Type = case\_when(Hospital\_Type == "Community <125 Beds"~ "Comm. <125",  
 Hospital\_Type == "Community 125-250 Beds"~ "Comm. 125-250",  
 Hospital\_Type == "Community >250 Beds"~ "Comm.>250",  
 Hospital\_Type == "Long-Term Acute Care"~ "Long-Acute",  
 .default = Hospital\_Type))|>  
   
 ggplot(aes(x= Hospital\_Type, y= SIR, fill=Hospital\_Type))+  
 geom\_boxplot()+  
 scale\_fill\_manual(values=color\_palette)+  
 theme(  
 legend.position="none",  
 plot.title = element\_text(size=11)  
 ) +  
 ggtitle(paste0("Standardized Infection Ratio (SIR) in each hospital type ")) +  
 xlab("")+ylab("")   
```



Distribution of SIR values with respect to Hospital\_Type

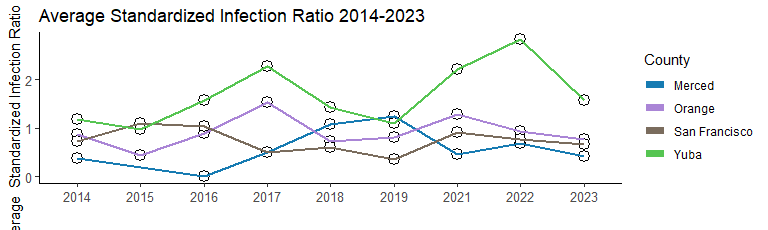
Next, we will be analyzing Hospital\_Type, County with respect to a numerical SIR. The SIR ratio was gradually increased in the Pediatric hospitals, unlike in the Long-Term Acute Care hospital, the ratio was slowly decreased except in 2022. In 2021, Community <125 Beds, Community 125-250 Beds, “Community >250 Beds, and Major Teaching, had the highest SIR ratio.

```{r}  
#| title: SIR trend  
#| code-fold: true  
#| warning: false  
#| fig-width: 8  
#| fig-height: 8  
#| fig-column: page-left  
  
mrsa\_combine|>  
 group\_by(Year, Hospital\_Type)|>  
 summarise(avg\_SIR = mean(na.omit(SIR)))|>  
 mutate(Risk\_Category = Hospital\_Type)|>  
 filter(Risk\_Category != "NA",   
 Year != 2013)|>  
 ggplot(aes(x= as.character(Year), y= avg\_SIR, color=Hospital\_Type))+  
   
 geom\_point(shape=21, color="black", fill="white", size=4) +  
 # geom\_point(size = 4, color = "#0e263560", shape = "circle open") +  
 geom\_line(aes(group=Risk\_Category),linewidth = 1)+  
 geom\_line(linewidth = 2.5) +  
 theme\_bw()+  
 theme(#axis.text.x=element\_blank(),  
 axis.text=element\_text(size=10),  
 legend.position = "right",  
 axis.line = element\_line(colour = "black"),  
 panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 panel.border = element\_blank(),  
 panel.background = element\_blank()  
 )+  
 scale\_color\_manual(values=color\_palette) +  
 labs(y="Average Standardized Infection Ratio (SIR)", x="",   
 title="Average Standardized Infection Ratio 2014-2023")  
```



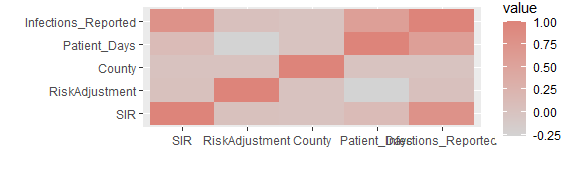
The trends of SIR ratio look similar in the more densely populated counties after 2020.

```{r}  
#| title: SIR trend  
#| code-fold: true  
#| warning: false  
#| fig-width: 8  
#| fig-height: 8  
  
mrsa\_combine|>  
 filter(County == "Orange" | County == "San Francisco"| County=="Merced"| County=="Yuba")|>  
 group\_by(Year, County)|>  
 summarise(avg\_SIR = mean(na.omit(SIR)))|>  
 ggplot(aes(x= as.character(Year), y= avg\_SIR, color=County))+  
   
 geom\_point(shape=21, color="black", fill="white", size=4) +  
 # geom\_point(size = 4, color = "#0e263560", shape = "circle open") +  
 geom\_line(aes(group=County),linewidth = 1)+  
 geom\_line(linewidth = 2.5) +  
 theme\_bw()+  
 theme(#axis.text.x=element\_blank(),  
 axis.text=element\_text(size=10),  
 legend.position = "right",  
 axis.line = element\_line(colour = "black"),  
 panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 panel.border = element\_blank(),  
 panel.background = element\_blank()  
 )+  
 scale\_color\_manual(values=color\_palette) +  
 labs(y="Average Standardized Infection Ratio (SIR)", x="",   
 title="Average Standardized Infection Ratio 2014-2023")  
```



Based on the scatter plots in the previous [post](https://clfee.github.io/home/posts/2024-09-01-Graphs-for-Communication), it seems there is a linear pattern between Patient\_Days and Infections\_Reported variables. We can further compare results between variables using correlation analysis.

```{r}  
#| label: Correlation  
#| fig-cap: " Correlation of variables"  
#| code-fold: true  
#| warning: false  
#| fig-width: 6  
#| fig-height: 6  
  
# spearman correlation  
  
dt <- mrsa\_combine|>  
 mutate(Hospital\_Type= ifelse(is.na(Hospital\_Type), "mixed", Hospital\_Type),  
 RiskAdjustment = ifelse(is.na(Hospital\_Category\_RiskAdjustment), "mixed",  
 Hospital\_Category\_RiskAdjustment))|>  
 mutate(Hospital\_Type= as.numeric(as.factor(Hospital\_Type)),  
 RiskAdjustment = as.numeric(as.factor(RiskAdjustment)),  
 County = as.numeric(as.factor(County)))|>  
 select(SIR, RiskAdjustment, County,Patient\_Days,Infections\_Reported)  
  
cor\_ <- round(cor(na.omit(dt), method = 'spearman'),2)  
cor\_matrix <- melt(cor\_)  
colnames(cor\_matrix ) <- c("X1","X2","value")  
ggplot(cor\_matrix, aes(X1, X2, fill= value)) +   
 geom\_tile() +  
 scale\_fill\_gradient(low="lightgrey", high="#dd847a")+  
 xlab("")+ylab("")  
```



Correlation of variables

The value of the correlation coefficient always ranges between 1 and -1, and we will treat it as a general indicator of the strength of the relationship between variables. Since the data is not normally distributed, we will use Spearman’s rho (Spearman’s rank correlation coefficient). It’s a rank correlation coefficient because it uses the rankings of data from each variable (e.g., from lowest to highest) rather than the raw data itself.

We can see that the SIR has a positive strong correlation with Infections\_Reported (correlation coefficients value 0.78), and it has a much weaker correlation with Patient\_Days(correlation coefficients value 0.12).The results also show a positive strong correlation between Patient\_Days and Infections\_Reported (correlation coefficients value 0.58). Weak Correlation is found between Hospital Risk Adjustment and Patient\_Days(0.27). No correlations is found between Hospital Risk Adjustment and Infections\_Reported(0.04), and between Hospital Risk Adjustment and SIR(0.03). County seems to have no correlation with other variables.

## Base forecast model (Xgboost)

```{r}  
#| title: Forecast base model  
#| code-fold: true  
#| warning: false  
#| fig-width: 8  
#| fig-height: 10  
  
# Step 1: Data Preparation  
dt\_all <- mrsa\_combine|>  
 mutate(Hospital\_Type = case\_when(Hospital\_Type == "Community, <125 Beds"~ 1,  
 Hospital\_Type == "Community, 125-250 Beds"~ 2,  
 Hospital\_Type == "Community, >250 Beds"~ 3,  
 Hospital\_Type == "Major Teaching"~ 4,  
 .default = 0),  
 Hospital\_Type = as.factor(Hospital\_Type),  
 Facility\_ID = as.factor((Facility\_ID)))|>  
 select(-c(Hospital\_Category\_RiskAdjustment,County))  
  
# Step 2: Fill missing values  
# Use KNN to fill missing values  
dt\_no2023 <- kNN(dt\_all|>filter(Year != 2023), variable = "SIR", k = 10)  
dt\_all <- rbind(dt\_no2023, dt\_all|> filter(Year == 2023)|> mutate(SIR\_imp = FALSE))|>   
 group\_by(Facility\_ID)|>   
 arrange(Facility\_ID, (as.numeric(Year)))|>  
 mutate(lag\_IR = dplyr::lag(Infections\_Reported, n = 1,default = NA))|>  
 mutate(lag\_PD = dplyr::lag(Patient\_Days, n = 1, default = NA))|>  
 mutate(lag\_SIR = dplyr::lag(SIR, n = 1, default = NA),  
 impute\_SIR = as.factor(ifelse(SIR\_imp == TRUE , 1, 0)))  
   
dt <- na.omit(dt\_all)|>select(-c(Infections\_Reported,Patient\_Days,SIR\_imp))  
  
# Step 3: Data Splittingv   
# Split the data into training and test sets   
train\_data <- dt|> filter(Year != 2023)  
test\_data <- dt|> filter(Year == 2023)  
  
# Step 4: Prepare data for XGBoost  
train\_labels <- train\_data$SIR  
test\_labels <- test\_data$SIR  
  
# Convert factors to numeric for xgboost  
dtrain <- model.matrix(~. -1, data = train\_data |> select(-Year, -SIR))  
dtrain <- xgb.DMatrix(data = dtrain, label = train\_labels)  
dtest <- model.matrix(~. -1, data = test\_data |> select(-Year, -SIR))  
dtest <- xgb.DMatrix(data = dtest, label = test\_labels)  
  
# Step 5: Train XGBoost Model  
params <- list(objective = "reg:squarederror", eval\_metric = "rmse",booster= "gblinear")  
xgb\_model <- xgb.train(params = params, data = dtrain, nrounds = 300, verbose = 0, watchlist = list(train = dtrain, test = dtest), early\_stopping\_rounds = 50)  
  
  
# Step 6: Forecast and Evaluate  
preds <- predict(xgb\_model, dtest)  
  
# Evaluate model performance using RMSE  
rmse <- sqrt(mean((test\_labels - preds)^2))  
cat("RMSE:", rmse, "\n")  
  
# Step 7: Plot the Results  
test\_data$predicted\_SIR <- preds  
test\_data$diff\_act\_forecast <- abs(test\_data$predicted\_SIR-test\_data$SIR)  
test\_data <- full\_join(test\_data|>select(-Hospital\_Type), unique(map\_risk\_list),by= "Facility\_ID")  
na.omit(test\_data) |>  
 filter(Hospital\_Type == "Long-Term Acute Care")|>  
 #filter(Hospital\_Type==4, County == "Los Angeles" )|>  
 ggplot(aes(x= Facility\_ID))+  
 geom\_point(aes(y = SIR, color = "Actual"), fill = NA) +  
 geom\_point(aes(y = predicted\_SIR, color = "Forecasted"), fill = NA) +  
 theme\_minimal()+  
 theme(  
 legend.position="bottom",  
 plot.title = element\_text(size=11),  
 axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1)  
 ) +  
 labs(title = "Year 2023 Forecast vs Actuals", y = "SIR", x = "", caption ="Produced by CF Lee" )  
```

RMSE: 1.085491

