The Effects of Age and Hospital Stay on Risk of Infection

Allen Arriaga

10/10/25

Hospital-acquired infections are a critical concern for healthcare systems worldwide, posing threats to patient safety and increasing healthcare costs. Understanding the factors that contribute to infection risk is extremely important for improving hospital practices and patient outcomes. In this analysis, I applied multiple linear regression modeling to data on hospital performance and patient characteristics to investigate how hospital stay length, patient age, and diagnostic testing rates influence infection risk.

The main question of this investigation is: "How do average patient age and length of stay affect the risk of hospital-acquired infection?"

1 Data Description and Processing

The dataset used in this analysis was collected from hospitals across the United States and includes a variety of operational, demographic, and clinical characteristics. Each observation represents data from one hospital during the study period.

This dataset contains several types of variables related to hospital resources, patient demographics, and infection control practices.

Predictors:

- **Stay:** Average patient stay (days)
- **Age:** Average patient age (years)
- **Culture:** Cultures per 100 patients without infection symptoms
- **Xray:** X-rays per 100 patients without pneumonia symptoms
- **Beds:** Average number of hospital beds
- **MedSchool:** 1 = hospital has a medical school, 2 = no
- **Region:** Geographic region (NE, NC, S, W)
- **Census:** Average daily patient count
- **Nurses:** Average full-time equivalent nurses
- **Facilities:** Percent of 35 possible hospital services provided

The dataset was checked for missing or inconsistent values before analysis. Of the original 116 observations, 3 records containing missing data were removed, leaving 113 complete observations for the final dataset.

The dependent variable in this dataset is InfctRsk, which measures the average estimated probability of acquiring infection in hospital (percent)

2 Data Exploration

The dataset contains observations on hospitals from different regions. As we believe infection rates may vary depending on regional factors such as population density and healthcare resources, we examine the overall distribution of infection risk across all regions.

Table 1: Number of Hospitals by Region

Var1	Freq
NE	28
NC	32
S	37
W	16

I also included the distribution of the dependent variable, Infection Risk, As well as its correlation to the Region variable.

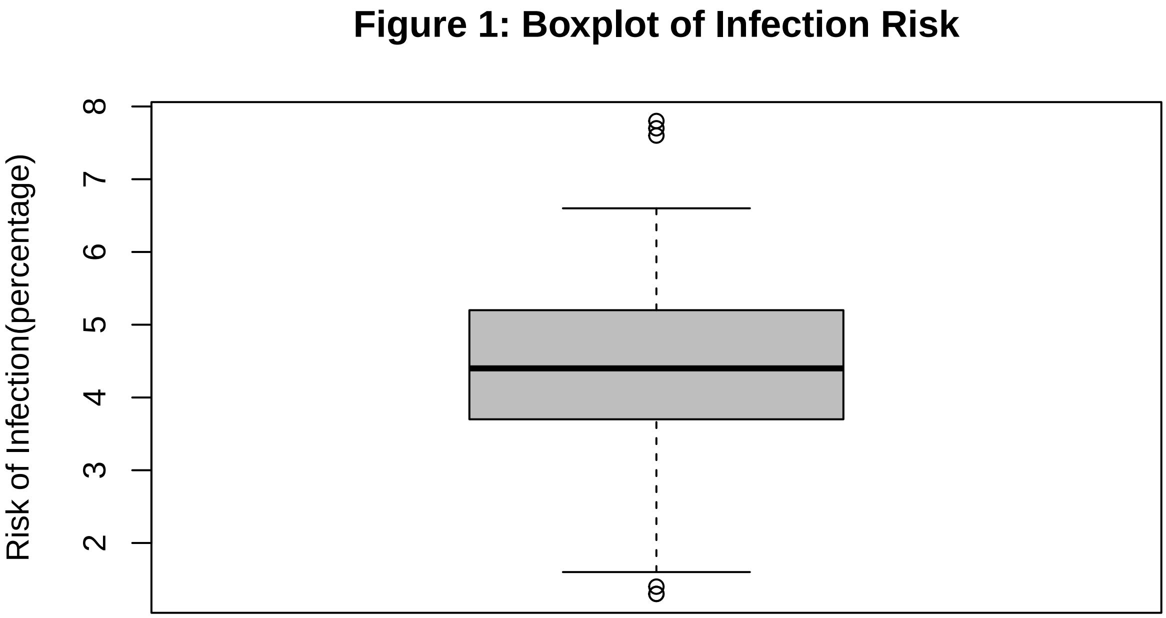

Figure 1: Infection Risk

Table 2: Summary Statistics for Infection Risk(percentage)

Statistic	Value
Min	1.300000
1st Quartile	3.700000
Median	4.400000
Mean	4.354867
3rd Quartile	5.200000
Max	7.800000

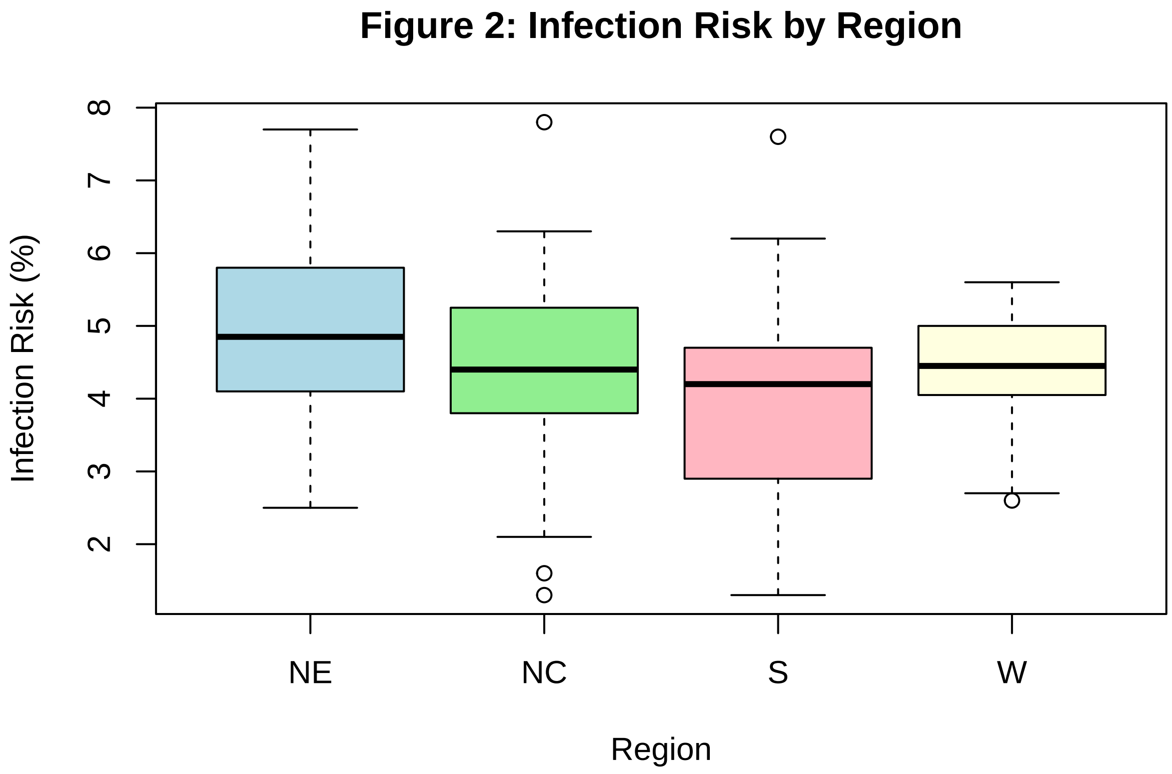

Figure 2: Infection Risk by Region

Table 3: Summary of Infection Risk by Region

Region	Min	Q1	Median	Mean	Q3	Max
NE	2.5	4.200	4.85	4.860714	5.750	7.7
NC	1.3	3.850	4.40	4.393750	5.225	7.8
S	1.3	2.900	4.20	3.927027	4.700	7.6
W	2.6	4.075	4.45	4.381250	4.850	5.6

The summary statistics indicate that infection risk varies slightly by region. Hospitals in the Northeast (NE) show the highest mean infection risk (4.86%), while hospitals in the South (S) have the lowest (3.93%).

The spread is somewhat wider in the South and North Central regions, suggesting more variability among hospitals in those areas.

3 Modeling the Effect of Age and Stay on Risk of Infection

The variables Age and Stay measure average patient age (in years) and average length of hospital stay (in days), respectively. Both variables are expected to influence infection risk: older patients may have weaker immune systems, and longer hospital stays increase exposure time to potential sources of infection. By examining these variables together, we can assess how hospital factors interact to affect the probability of acquiring an infection during hospitalization. This analysis helps identify which factors may contribute most to infection risk and guide improvements in hospital management and patient care practices.

3.1 Hypothesis

I hypothesize that a higher age and longer hospital stay will result in an increase probability of acquiring infection in hospital

3.2 Model 1:

The dataset includes hospitals from four major U.S. regions: Northeast, North Central, South, and West. Regional differences in healthcare resources, population density, and hospital practices may influence the likelihood of hospital-acquired infections.

I fit a first linear model that included Stay and Region as predictors, with Infection Risk (InfctRsk) as the outcome variable. This model allows us to examine whether the average length of patient stay and geographic region together help explain variation in hospital infection risk.

Table 4: Predicting Infection Risk with Stay and Region

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2859471	0.7409280	0.3859310	0.7003069
Stay	0.4125527	0.0640350	6.4426162	0.0000000
RegionNC	0.1128748	0.3033626	0.3720789	0.7105632
RegionS	-0.1508367	0.3056157	-0.4935503	0.6226260
RegionW	0.7479536	0.3992388	1.8734491	0.0637099

The adjusted R-squared for this model was 0.3029, indicating that Stay and Region together explain about 30.3% of the variation in Infection Risk.

The results indicate that average length of hospital stay is the strongest predictor of infection risk. Hospitals where patients stay longer tend to have higher infection risk. Regional differences (particularly for the Western region) may play a smaller or less consistent role.

3.3 Model 2:

I fit a second linear model that included Age and Stay as predictors, with Infection Risk (InfctRsk) as the outcome variable. This model allows us to assess how both patient age and hospital stay length together contribute to the probability of acquiring an infection during hospitalization.

Table 5: Predicting Infection Risk with Stay and Age

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.2659094	1.3211506	1.715103	0.0891421
Stay	0.3879161	0.0571960	6.782230	0.0000000
Age	-0.0310675	0.0245041	-1.267850	0.2075285

The adjusted R-squared for this model was 0.2820, indicating that Length of Stay and Age together explain about 28.2% of the variation in Infection Risk.

4 Predicting with Models 1 and 2

I evaluated the ability of the first and second models to predict Infection Risk on test data. We used 10-fold cross validation using 80% of the data for training and 20% for testing on each fold.

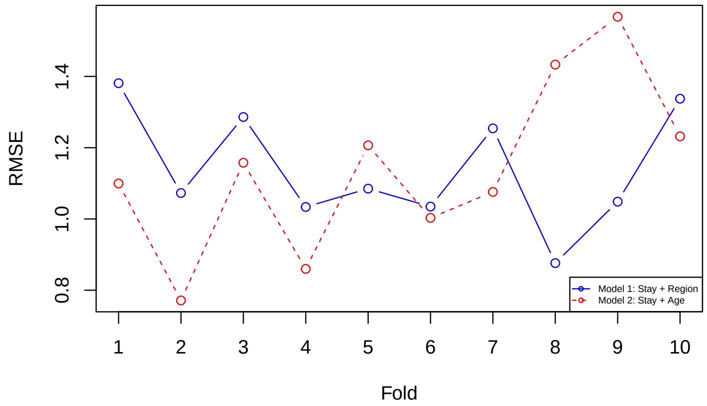

Figure 3: RMSE Over 10 Folds.

Table 6: Mean RMSE for Infection Risk Models 1 and 2

Label	Mean_RMSE
Model 1: Stay + Region	1.14
Model 2: Stay + Age	1.14

5 Summary and Conclusions

The modeling results show that both Model 1 (Stay + Region) and Model 2 (Stay + Age) explain a moderate portion of the variation in hospital infection risk, with adjusted R^2 values of 0.3029 and 0.2820 In both models, average length of stay was the strongest and most statistically significant predictor of infection risk ($p < 0.001$), supporting the idea that patients who stay in the hospital longer face higher exposure and risk of infection.

Regional differences in Model 1 showed weaker effects, with only the Western region approaching significance (p = 0.064), suggesting that while regional variation may exist, it is not a dominant factor once average hospital stay is taken into account. In Model 2, patient age was not statistically significant (p = 0.208), hinting that age alone does not have a large influence on infection risk once length of stay is accounted for.

Cross-validation results further supported these findings. Both models achieved an average RMSE of 1.14, suggesting similar accuracy wih predictions. However, given that Model 1 explained slightly more variation and included a region factor that may capture institutional or environmental effects, it performed marginally better overall.

These results partially support the hypothesis, as infection risk clearly increases with longer hospital stays, but the effect of patient age was not significant. This suggests that infection control efforts might be more effective if focused on managing hospital duration and exposure instead of just patient demographics alone.

If infection risk, patient age, or hospital stay length were not recorded accurately or were measured differently across hospitals, the model's results could be misleading. For example, if some hospitals underreport infections or define them differently, the data would be less reliable and the model's conclusions less accurate.

Overall, both models provide valuable insights into hospital infection risk factors, highlighting length of stay as the most influential predictor and suggesting that targeted interventions to reduce hospitalization duration could effectively lower infection rates.