Surgical Site Infection Risk Following Colon Procedures in California Hospitals (2024)

Rob Daniels

January 15, 2026

Executive summary/abstract

Surgical site infections (SSIs) following colon procedures are a significant source of patient morbidity, mortality, and healthcare costs. This project analyzes 2024 California acute-care hospital data to estimate facility-level SSI risk for colon surgeries, accounting for facility type and county-level variation. A hierarchical Bayesian binomial model with county-specific random effects and fixed effects for facility type was fit, producing partially pooled estimates that stabilize rates for low-volume facilities. Across 288 hospitals performing 29,835 procedures, 628 SSIs were reported, corresponding to a statewide pooled infection proportion of 2.1%.

Model results indicate that, compared with the reference group (Community, $<$ 125 Beds), all other facility types have higher log-odds of infection: posterior means of 0.30, 0.61, and 0.59 for Community 125–250 Beds, Community $>$ 250 Beds, and Major Teaching hospitals, respectively. The county-level random effect standard deviation, $\sigma \approx 0.46$, reflects moderate variation in baseline risk across counties. These findings illustrate the value of hierarchical modeling for sparse, heterogeneous healthcare data, providing stable and interpretable facility-level risk estimates while highlighting residual county-level variation. The results support evidence-based understanding of SSI risk and can inform targeted quality improvement initiatives.

Introduction

Surgical site infections (SSIs) are a common complication of surgery, accounting for approximately 20–31% of all hospital-acquired infections and contributing to increased patient morbidity, mortality, and healthcare costs [1]. After colorectal surgery, roughly 4.2% of patients develop SSIs, often resulting in prolonged hospitalization and higher treatment costs [3]. Because SSIs are considered partially preventable, facility-level infection rates are widely used for hospital quality monitoring, benchmarking, and public reporting.

Accurate estimation of SSI risk at the facility level is challenging because many hospitals perform relatively few colon procedures and therefore report zero or very small numbers of infections in a given year. In such settings, raw observed rates can be dominated by sampling variability, producing extreme values that do not reliably reflect underlying risk. These features motivate the use of statistical models that explicitly account for uncertainty and allow information to be shared across comparable facilities to obtain stable and interpretable risk estimates.

The goal of this project is to use 2024 California acute-care hospital data to estimate facility-level SSI risk for colon procedures while accounting for facility type and county-level context [2]. Specifically, the analysis seeks to answer the following questions:

1. What are the posterior estimates of SSI risk for each facility, and how do these compare with the observed rates when uncertainty and partial pooling are considered?
2. How much do SSI risks vary across counties after adjusting for facility type?
3. Which facility types are associated with higher infection probabilities, after accounting for county-level variation?
4. What is the posterior probability that a facility-specific SSI risk exceeds the statewide pooled infection proportion?

To address these questions, a hierarchical Bayesian binomial model with county-specific random effects and fixed effects for facility type was fit to the data. This specification allows facilities within the same county to share information while capturing meaningful differences across facilities and counties. The hierarchical approach produces partially pooled estimates with full posterior uncertainty, providing more stable and interpretable risk estimates than raw observed rates, particularly for low-volume hospitals.

Data

The data for this analysis were obtained from the California Department of Health and Human Services and consist of 2024 facility-level reports of colon surgery SSIs. These data are mandatory hospital self-reports submitted as part of statewide healthcare-associated infection

surveillance. The dataset was downloaded on December 28, 2025, and comprises 288 acute-care hospitals across 42 counties. It includes only adult patients and excludes pediatric hospitals. Key variables include the number of colon procedures with SSIs (numerator), the total number of colon procedures (denominator), county, and facility type, enabling estimation of facility-level infection risk and comparison across counties and facility types (Table 1).

Table 1: Facility types among California acute-care hospitals reporting colon surgery SSIs in 2024.

Facility type	n	%
Community, < 125 Beds	67	23.3%
Community, 125–250 Beds	58	20.1%
Community, > 250 Beds	42	14.6%
Major Teaching	121	42.0%
Total	288	100.0%

Across all facilities, 628 SSIs were reported for 29,835 colon procedures, resulting in a pooled infection proportion of 2.1%. Facility-level SSI rates vary substantially, particularly among low-volume hospitals, highlighting the need for partial pooling. These data are well suited to answering the research questions because they provide complete, facility-level counts with contextual information on facility type and county. This allows for estimation of SSI risk at the facility level, comparison across facility types, and quantification of county-level variation. By including all non-pediatric acute-care hospitals that reported colon surgery outcomes in 2024, the dataset provides a comprehensive snapshot appropriate for hierarchical modeling.

Table 2 presents the distribution of facility-level SSI rates. The "mean" is the unweighted mean of facility proportions, giving each facility equal weight regardless of procedure volume. This measure reflects the average facility-level rate, not the patient-level statewide rate.

Table 2: Summary statistics of facility-level observed SSI rates

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	1.06	1.76	2.62	16.7

Figure 1 shows the distribution of facility colon procedure volumes. Volumes are highly skewed: most facilities performed fewer than 200 procedures, while the highest-volume facility (Stanford Health Care) performed 623 procedures.

Figure 2 illustrates the relationship between facility volume and observed SSI rate. Many facilities report only 1–6 infections, resulting in discrete observed rates. On the log-scaled volume axis, these fractions form visible curved bands, particularly among facilities with fewer than 250 procedures, indicating high variability at low volumes and motivating the use of a hierarchical model.

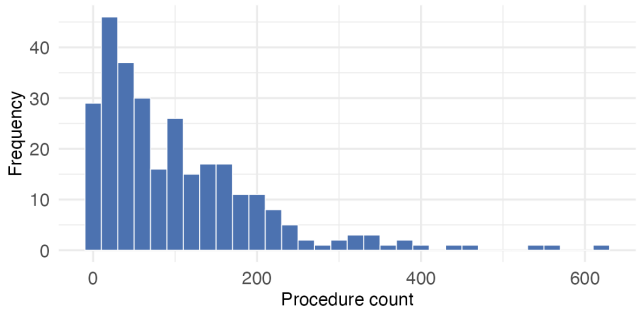

Figure 1: Distribution of facility-level colon procedure volumes in California acute-care hospitals, 2024.

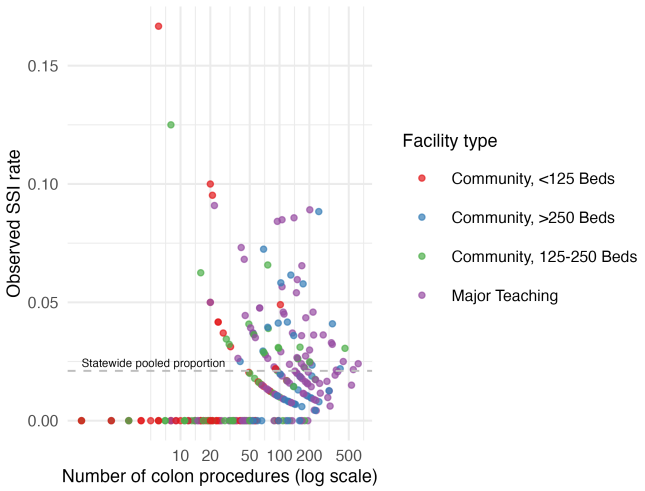

Figure 2: Observed SSI rates (in decimals) versus colon procedure volumes (log scale) for California acute-care hospitals in 2024, colored by facility type. The dashed line indicates the statewide pooled infection proportion.

Model

Facilities are grouped within counties, creating a hierarchical data structure in which outcomes from facilities in the same county may be partially influenced by shared county-level characteristics, such as patient population composition, referral patterns, and institutional practices. To account for this structure, a hierarchical Bayesian binomial model with a random intercept for county and fixed effects for facility type was fit to the data. This model was appropriate because the outcome consisted of counts of infections out of known procedure totals, and the hierarchical formulation allowed partial pooling across facilities within counties. This stabilized

estimates for low-volume facilities while capturing meaningful variation across counties.

The hierarchical specification is:

Likelihood:

$$y_i \mid a_{c[i]}, \phi_i \sim \text{Binomial}(n_i, \phi_i), \quad i = 1, \ldots, 288$$

Linear predictor:

$$\text{logit}(\phi_i) = a_{c[i]} + \beta_1 \, \text{beds}_{125\text{–}250,i} + \beta_2 \, \text{beds}_{>250,i} + \beta_3 \, \text{major\_teach}_i$$

County effects:

$$a_c \mid \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2), \quad c = 1, \ldots, 42$$

Weakly informative priors were assigned to the regression coefficients and county-level hyperparameters:

$$\beta_k \sim \mathcal{N}(0, 2.5^2), \ \mu \sim \mathcal{N}(0, 2.5^2), \ \sigma \sim t_3^+(0, 2.5)$$

These choices follow Gelman et al. (*Bayesian Data Analysis*, Section 5.7) and provide stability for variance estimates while allowing the observed data to dominate the posterior [4].

The model was implemented using JAGS in R with three chains and 100,000 posterior iterations following burn-in. Residual diagnostics did not indicate substantial lack of fit. Standard convergence diagnostics indicated that all chains converged to the same stationary distribution.

Autocorrelation patterns were consistent with expectations for hierarchical models. The regression coefficients (β_1–β_3) and the overall county-level mean (μ) exhibited moderate autocorrelation that persisted until approximately 150 lags. The standard deviation of the county-level effects (σ) showed low autocorrelation that dissipated by roughly 20 lags. County-level random effects displayed a range of autocorrelation patterns: some counties showed moderate autocorrelation persisting up to 150–200 lags, while others exhibited low autocorrelation after approximately 10 lags, reflecting differences in the number of facilities and the amount of information available across counties. Effective sample sizes were sufficient for reliable posterior summaries, with ESS $> 3{,}000$ for regression coefficients, μ, and σ, and county-level effects ranging from approximately 3,000 to more than 20,000.

Model fit was further summarized by a mean deviance of 899.4 and an effective complexity penalty of 24.2. The penalty was lower than the nominal number of parameters because partial pooling caused county-level effects to share information, effectively reducing the number of independent parameters. Together, these diagnostics support stable estimation and valid posterior inference for both facility- and county-level infection risks.

For comparison, a non-hierarchical binomial logistic regression model without county-level random effects was also fit. This alternative specification estimated only global regression coefficients and implicitly assumed a common baseline infection risk across counties. Consequently, it could not capture geographic heterogeneity in baseline risk and provided limited support for facility-level inference in low-volume settings. These limitations supported the use of the hierarchical specification, which more closely reflects the multilevel structure of the data and the substantive sources of variation in SSI risk.

Results

The R model summary output is shown below. Only three county coefficients are displayed due to space limitations. These correspond to the largest, middle, and smallest values.

```
Iterations = 11001:111000
Thinning interval = 1
Number of chains = 3
Sample size per chain = 1e+05

1. Empirical mean and standard deviation for
   each variable, plus standard error of the
   mean:

         Mean     SD  Naive SE Time-series SE
a[4]  -3.7369 0.2561 0.0004675      0.0035666
a[6]  -4.3654 0.4789 0.0008743      0.0032860
a[16] -5.0277 0.4407 0.0008046      0.0040004
b[1]   0.2989 0.2348 0.0004287      0.0036989
b[2]   0.6113 0.2254 0.0004115      0.0038306
b[3]   0.5934 0.2125 0.0003880      0.0038373
mu    -4.4482 0.2171 0.0003965      0.0036734
sigma  0.4633 0.1023 0.0001868      0.0006561

2. Quantiles for each variable:

         2.5%     25%     50%     75%   97.5%
a[4]  -4.2527 -3.9064 -3.7328 -3.5622 -3.2479
a[6]  -5.3311 -4.6765 -4.3603 -4.0491 -3.4318
a[16] -5.9675 -5.3045 -5.0034 -4.7226 -4.2328
b[1]  -0.1530  0.1396  0.2956  0.4551  0.7683
b[2]   0.1822  0.4574  0.6076  0.7601  1.0669
b[3]   0.1898  0.4481  0.5894  0.7336  1.0252
mu    -4.8906 -4.5907 -4.4429 -4.2995 -4.0391
sigma  0.2912  0.3911  0.4533  0.5244  0.6922
```

Posterior estimates indicate that all non-reference facility types had higher log-odds of infection than the reference group (Community, $<$ 125 Beds). Posterior means for the regression coefficients on the log-odds scale were 0.30, 0.61, and 0.59 for Community 125–250 Beds, Community $>$ 250 Beds, and Major Teaching facilities, respectively. The 95% credible intervals for all three

groups lay largely above zero, indicating a higher infection probability relative to the reference.

The overall county-level intercept had a posterior mean of $\mu = -4.45$, representing the baseline log-odds for the reference facility type. This logit can be transformed to the probability scale using the inverse-logit function:

$$\begin{aligned}\text{logit}^{-1}(\mu) &= \frac{1}{1+\exp(-\mu)} \\ &\approx \frac{1}{1+\exp(4.45)} \\ &\approx 0.0116\ (1.16\%)\end{aligned}$$

which corresponds to the mean statewide estimate of SSI risk for a Community, < 125 Bed facility.

The posterior standard deviation of the county-level random effects, $\sigma \approx 0.46$, indicated moderate variation in baseline risk across counties, suggesting that county-specific factors contributed meaningfully to infection risk.

In addition to estimating facility- and county-level effects, posterior samples from the hierarchical Bayesian model were used to make probabilistic statements about individual hospitals relative to the statewide pooled infection proportion. For example, the posterior distribution of Riverside Community Hospital's facility-specific infection probability, $\phi_{\text{Riverside}}$, was extracted and compared to the overall statewide pooled proportion. The proportion of posterior draws in which $\phi_{\text{Riverside}}$ exceeded the statewide rate was computed, yielding a median posterior SSI rate of approximately 2.08%, with a 95% credible interval of 1.46% to 2.90%, and a 47.6% posterior probability of exceeding the statewide pooled infection proportion of 2.10%. These results indicate no strong evidence that Riverside Community Hospital's infection rate was higher than the state average. This demonstrates the advantage of the Bayesian framework, which allows direct probability statements about individual facilities, an inference not possible with traditional frequentist models, and highlights the utility of posterior distributions for facility-specific risk comparisons.

Conclusions

The hierarchical Bayesian model provides a coherent framework for estimating surgical site infection risk while accounting for both facility type and county-level variation. By partially pooling information across facilities within counties, the model stabilizes estimates for smaller-volume hospitals, reduces the noise inherent in raw observed rates, and yields more interpretable posterior distributions of facility-specific risk. The analysis highlights that larger community hospitals and major teaching facilities tend to have higher infection probabilities relative to smaller community hospitals, and that baseline risk differs meaningfully across counties, reflecting regional heterogeneity.

Posterior distributions enable probabilistic comparisons of individual facility risk relative to the statewide pooled infection proportion, demonstrating the added inferential flexibility of the Bayesian approach. For most facilities, these comparisons reveal the degree of certainty regarding elevated or lower-than-average risk, emphasizing the importance of incorporating both facility- and county-level effects in risk assessment.

Limitations of this study include reliance on facility-reported data, which may be incomplete or misclassified, and the absence of patient-level covariates that could further explain variation in infection risk. The model also assumes a common effect of facility type across counties, which may simplify local differences. Future work could extend the framework to incorporate additional predictors, explore alternative hierarchical structures, and conduct more extensive predictive validation.

Overall, the hierarchical approach demonstrates that multilevel modeling with partial pooling can generate robust, interpretable estimates of SSI risk, capturing both facility-specific patterns and regional variability while allowing direct probabilistic statements about relative risk. This approach provides a principled foundation for understanding variability in infection rates across hospitals and regions.

References

[1] CDC. *Chapter 9: Surgical Site Infection (SSI) Event*. National Healthcare Safety Network (NHSN), Centers for Disease Control and Prevention, January 2026. URL https://www.cdc.gov/nhsn/pdfs/pscmanual/9pscssicurrent.pdf. Patient Safety Component Protocol, NHSN.

[2] CDPH. Surgical site infections (ssis) for operative procedures in california hospitals. California Health and Human Services Open Data Portal, 2025. URL https://data.chhs.ca.gov/dataset/surgical-site-infections-ssis-for-28-operative-procedures-in-california-hospitals. Dataset for SSI in adult patients, 2024.

[3] O. Gantz, P. Zagadailov, and A. M. Merchant. The cost of surgical site infections after colorectal surgery in the united states from 2001 to 2012: A longitudinal analysis. *American Surgeon*, 85(2):142–149, February 2019. doi: 10.1177/000313481908500219. URL https://doi.org/10.1177/000313481908500219.

[4] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3 edition, 2013. doi: 10.1201/b16018.