Investigating Mortality Rates from Cardiovascular Pediatric Surgery in STS Public Reporting

Jiajun Song Yizi Zhang

Abstract

The purpose of this report is to provide parents with a decision rule for deciding which hospital to go to for a procedure given the specific condition of their children. To address this question, we examine the mortality rate for each procedure type in a hospital. Specifically, we fit a Bayesian joint model of mortality rates and hospital volumes to estimate the mortality rates by hospital and procedure. After ranking the hospital performances for the given procedures, we find that Texas Children's Hospital, UF Health Shands Children's Hospital and Helen DeVos Children's Hospital are the best hospitals for pediatric cardiovascular procedures. In addition to having good predictive output, our ranking system allows for us to quantify the uncertainty surrounding parameter and ranking estimates, allowing us to rigorously conclude that our findings are significant.

1 Introduction

As a national leader in health care transparency and accountability, the Society of Thoracic Surgeons (STS) believes that the public has a right to know the quality of surgical outcomes. In this report, we investigate mortality rates, both overall and stratified by procedure complexity, for the hospitals participating in the STS public reporting, leading to a ranking of hospitals in terms of their performance in conducting pediatric cardiovascular surgeries.

Ranking that is based solely on average mortality rate will be biased because of the existence of *case mix*. A hospital's *case mix* takes into account many different factors such as cardiac surgeon ratings and procedure types. For example, hospitals specialized in pediatric cardiology may treat more patients with more complicated and higher-risk conditions than other hospitals and, therefore, operate on patients with a lower chance of survival.

To model the hospital mortality rates, we start with a Binomial regression model with random effects that account for the variability between the hospitals and procedure types. However, we wanted to incorporate the effects of hospital volumes on the mortality rates more directly to adjust for *case-mix*, leading us to a Bayesian joint modeling framework where we model the hospital volumes as Poisson distributed. We further added interaction effects between hospital size and procedure types to allow the random effect of procedure types to vary among hospitals. The Bayesian joint modeling with interaction effects is the primary model we use to address the goals of our analysis.

To rank the hospital performance for a given procedure type, we compare hospitals based on the mortality risk, which is defined as the sum of the hospital effect and the interaction effect between the hospital size and the given procedure type. Meanwhile, we take the uncertainty during sampling into account by aggregating the ranking list from each iteration.

The reminder of this report is structured as follows. We briefly review the data used and our initial findings. We present the joint model and the various features it's designed to identify. We review our key findings in the results section and conclude with our answers to the key questions presented above and propose possible extensions. An appendix with further details of our analysis can be found after the references.

2 Materials and Methods

2.1 Data

We used the public reporting data collected from 82 participant hospitals in the STS Congenital Heart Surgery Database[1] for our analysis. The STS data presents hospital-specific results for procedures in each of the 5 STAT Mortality Category during the 2015-2018 reporting period. STAT Category 1 includes the least complex operations, which are associated with the

lowest risk of mortality. STAT Category 5 includes the most complex operations, which are associated with the highest risk of mortality. The hospital-specific results include the number of pediatric surgical procedures and the number of deaths resulting from those procedures.

Because the STS dataset does not include any additional profile information about the participant hospitals, we extracted hospital-level covariates from the U.S. News[2] and the American Hopsital Directory[3] to supplement the STS dataset. These hospital-level covariates include the number of beds, the NICU level, the cardiovascular surgery rating, the case-mix index, and the location (rural or urban) of the hospitals.

2.2 Exploratory Data Analysis

In this section, we provide a brief summary of our findings from the exploratory data analysis. A more detailed walk through of the EDA can be found in the appendix.

1. Per George et al.[4], we considered potential relationship between hospital mortality rates and volume. We found that low volume hospitals tend to have high mortality rates.

2. *Data sparsity* issue. Some low volume hospitals with zero mortality rates only perform a few procedures per year. The mortality rates from these hospitals are quite unstable, and therefore we should be careful about shrinkage in a hierarchical modeling setting.

3. Regardless of the hospital volumes, hospital mortality rates increase with the increasing level of STAT Mortality Category. See Figure 6 in the appendix.

4. Large hospitals operate more on complex and risky procedures while small hospitals treat more patients with less complex and risky conditions. Therefore, we wanted to include interaction effects between hospital sizes and procedure types.

5. We examined whether there is an association between hospital mortality rates and hospital-level covariates in both EDA and posterior inference. We found that the covariates are not predictive of mortality rates and decided not to include them in our primary model.

3 Model

To shed light on the issue of whether hospital mortality rates are related to volume, we consider a Bayesian joint modeling of mortality rates and hospital volumes for our STS data. Joint models provide more efficient estimates of the hospital-specific effects on both mortality rates and hospital volumes, and reduce bias in the estimates of the overall effect of hospital performance.

Let h ($h = 1, \ldots, H$) index the hospital, and j ($j = 1, \ldots, 5$) index the procedure type, $Y_{h,j}$ denotes the number of deaths and $n_{h,j}$ denotes the number of procedures for the j-th procedure done in hospital h. We assume $Y_{h,j}$ is generated from the following process:

$$Y_{h,j} \mid \alpha_h, \beta_{c_h,j}, \varepsilon_{h,j} \sim \text{Binomial}(n_{h,j}, p_{h,j}),\ \text{logit}(p_{h,j}) = \alpha_h + \beta_{c_h,j} + \varepsilon_{h,j}. \tag{1}$$

$$\text{V}_h \mid \phi_h, \lambda, \alpha_h \sim \text{Pois}(\phi_h \exp(\lambda \alpha_h)) \tag{2}$$

In the first half of the joint model, $p_{h,j}$ is the underlying mortality rate, which is determined by a hospital effect α_h and a cluster-procedure interaction effect $\beta_{c_h,j}$ with logit link.

Regarding the clustered procedure-hospital interactions, we decide the cluster membership based on the hospital volume. More specifically, let V_h denote the volume of the h-th hospital, i.e., number of total procedures during the time period of interest. If $\sum_{h'=1}^{H} I(V_h \geq V_{h'}) \leq \tau H$, where τ is a volume quantile threshold to be specified, hospital h belongs to the cluster of small hospitals ($c_h = 1$), whereas hospital h is categorized as a large hospital ($c_h = 2$) if $\sum_{h'=1}^{H} I(V_h \geq V_{h'}) \geq \tau H$. The primary reason we consider a clustered procedure-hospital

interaction effect instead of a direct interaction between procedure and hospital is to avoid parameter redundancy and identifiability issue.

In the second part of the joint model, we assume that the volume of hospitals V_h is related to baseline performance α_h due to *case-mix*. Therefore we apply Poisson regression to model this relationship, where ϕ_h controls for over-dispersion and λ determines the relationships between α_h and V_h. Intuitively, λ should be negative so that high mortality risk inferred from high α_h will induce low volume in hospital h. We show in the later sections that our model reflects this intuition with high significance.

For the implementation of this joint model, we propose a fully Bayesian approach with the following weakly informative priors:

$$\alpha_h \mid \sigma_\alpha^2 \sim \mathcal{N}(0,\, \sigma_\alpha^2),\ \beta_{c_h,j} \mid \sigma_\beta^2 \sim \mathcal{N}(0,\, \sigma_\beta^2),\ \varepsilon_{h,j} \mid \sigma_\varepsilon^2 \sim \mathcal{N}(0,\, \sigma_\varepsilon^2), \tag{3}$$

$$\phi_h \mid \sigma_\phi^2 \sim \mathcal{G}(\sigma_\phi^{2\,-1}, \sigma_\phi^{2\,-1}),\ \lambda \mid \sigma_\lambda^2 \sim \mathcal{N}(0,\, \sigma_\lambda^2), \tag{4}$$

$$\sigma_\alpha^2, \sigma_\beta^2, \sigma_\varepsilon^2, \sigma_\phi^2, \sigma_\lambda^2 \sim \mathcal{IG}(1,1). \tag{5}$$

The joint model is fit with the *rstan* package in R.

4 Results

4.1 Model Checking

1. Convergence diagnostics. In Appendix Figure 7, we provide the convergence diagnostics for parameters $\beta_{c_h,j}$ and λ. As shown in Figure 7, MCMC samples show good mixing and good convergence as well.

2. Posterior predictive check. In Appendix Figure 8 to Figure 12, we show the posterior predictive check for the mortality rate and the number of death cases. In addition, we examine the discrepancies between the distribution of mean and standard deviation of the number of death cases in the replicated data and in the observed data. Other than the variance of the mortality rate, the replicated data indicates that the model fit is a reasonable one. However, there is strong zero-inflation for procedure type 1 and procedure type 3 in terms of the mortality rate. Accounting for this pattern may improve the fitting of variation among mortality rate and we leave it to future directions due to limited time.

4.2 Parameter Inference

The posterior mean and 95% credible interval of λ are -3.27 (-4.6, -2.28). Since 0 is not included in the interval, we conclude in favor of the low volume hospitals having higher mortality rates. To explicitly compare the effects of hospital volumes on mortality rates, we plotted the posterior mean of α_h estimated by our joint model along with the corresponding volume V_h in Figure 1. The canonical model[4] (Geroge et al.) without any adjustment for hospital volumes is also displayed in Figure 1 as a comparison. The posterior means for our joint model are indicative of noticeably high mortality rates among low volume hospitals and relatively low rates for large hospitals.

To examine the validity of clustered interactions, we compare the 95% credible interval for $\alpha_h + \beta_{c_h,j}$ in two clusters and five procedure types in Figure 2. In the figure, $c_i; p_j$ stands for the overall mortality risk for hospitals within cluster i and of procedure type j, where we get by aggregating the posterior samples of $\alpha_{h:c_h=i} + \beta_{i,j}$. As shown in the figure, cluster two, i.e., hospitals with larger volumes will have lower mortality rates for more complex procedures such as procedure type 3, 4, and 5, whereas cluster one tends to perform better in procedure type 1, 2.

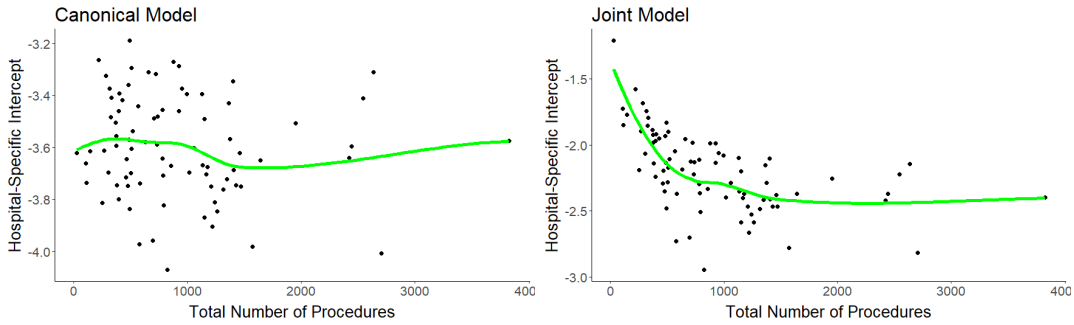

Figure 1: Posterior mean of α_h v.s. hospital volume V_h estimated by the canonical model and the joint model.

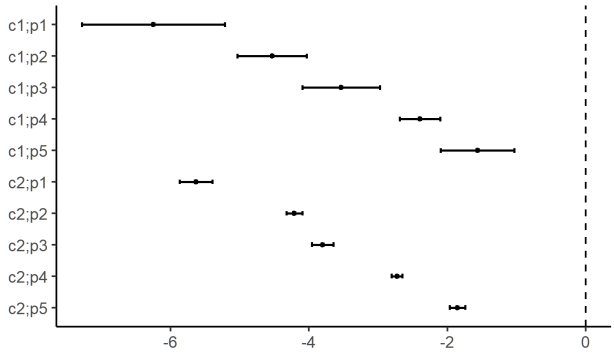

Figure 2: The 95% credible interval for the posterior sample. $c_i; p_j$ stands for the mean of the posterior samples of $\alpha_{h:c_h=i} + \beta_{i,j}$.

4.3 Ranking Hospital Performance

To satisfy the overarching goal of providing patients with a decision rule about which hospital to go to for a certain procedure type, we define the quantity below with parameters from our primary model

$$m_{h,j} = \alpha_h + \beta_{c_h,j} \tag{6}$$

Here, $m_{h,j}$ represents the mortality risk for taking a surgery of procedure type j at hospital h. Hence for procedure type j, we would recommend hospitals with smallest $m_{h,j}$.

To rank the hospital performance with Monte Carlo samples of $m_{h,j}$, we take into account the uncertainty during the sampling, by aggregating the ranking lists. As shown in Figure 3, many hospitals have close ranking, but the head and tail are noticeably different. Because of this finding, we rank the hospitals based on their probability of ranking among the top 10% during the sampling. Specifically, for given procedure type j, let $m_{h,j}^{(l)}$ denote the mortality risks for hospital h in the lth iteration, and we define

$$r_{h,j}^{(l)} = \sum_{h'=1}^{H} I(m_{h,j}^{(l)} \leq m_{h',j}^{(l)}),\ s_{h,j} = \frac{\sum_{l=1}^{L} I(r_{h,j}^{(l)} \leq 0.1H)}{L} \tag{7}$$

to represent the rank within each iteration and the ultimate performance score for hospital h applying a procedure type j surgery. Hospitals with the highest performance score in each procedure are listed in Table 1. The predicted hospital rankings in 1 suggest that if a child needs surgery for a level 1 or 2 condition, the parents should go to Helen DeVos Children's Hospital. While if the child suffers from more complicated and higher-risk condition, UF Health Shands Children's Hospital or Texas Children's Hospital are the best options.

Best for Type 1	Helen DeVos Children's Hospital
Best for Type 2	Helen DeVos Children's Hospital
Best for Type 3	UF Health Shands Children's Hospital
Best for Type 4	Texas Children's Hospital
Best for Type 5	UF Health Shands Children's Hospital

Table 1: Best hospital for different STAT procedure types.

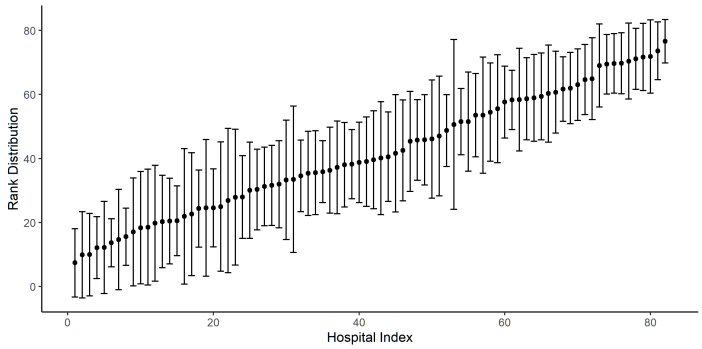

Figure 3: The mean rank and one standard deviation of all hospitals of procedure type 1.

5 Discussion

As presented above, the purpose of this analysis tries to answer the question of which hospital to go to for a given pediatric cardiovascular procedure. Both the literature and the data suggested that patients treated at low volume hospitals stand a much lower chance of survival. To account for this dependency, our method jointly modeled the hospital mortality rates and hospital volumes. From our model fits, we estimate that small hospitals indeed has a higher operation mortality rates, especially for more complex operations. These estimates are statistically significant based on the posterior distributions.

We ranked the hospital performance for each procedure type after aggregating the ranking list from each posterior samples to account for uncertainty. We calibrate the model by comparing its predictions to the general advice people would rely on based on the U.S. News hospital rankings. Specifically, we searched for the "Pediatric Cardiology & Heart Surgery" ranking on the U.S. News. We found that Texas Children's Hospital is ranked the number one among all hospitals for best pediatric cardiovascular surgery, while UF Health Shands Children's Hospital and Helen DeVos Children's Hospital are ranked the number 12 and 34 respectively. The predicted hospital rankings by our model are fairly consistent with the general advice. A future extension of our work would be to check our predictions against the general advice or the observed data with ranking-based accuracy metrics.

The key limitations we face in our analysis were data sparsity. We found that the estimated mortality rates of the small hospitals are shrunk to resemble the national average. Although we tried to fix the shrinkage issue by fitting a joint model of mortality rates and volumes, the mortality rates at low volume hospitals are still quite underestimated. Ideally, we could find better informative priors to adjust for the shrinkage issue to extend our work.

Another aspect that would be interesting to examine is to collect more informative hospital attributes or individual patient data within each hospital. This would make our model more robust and these additional case-mix adjustment will give us better estimates of mortality rates.

References

1. "Congenital Heart Surgery Public Reporting". *STS Public Reporting.* URL: https://publicreporting.sts.org/chsd-exp. Accessed: 3/24/2021.

2. "Best Hospitals by Specialty National Rankings". *U.S. News.* URL: https://health.usnews.com/best-hospitals/rankings. Accessed: 3/19/2021.

3. "Free Hospital Profiles". *American Hospital Directory.* URL: https://www.ahd.com/. Accessed: 3/19/2021.

4. George, Edward I., et al. "Mortality rate estimation and standardization for public reporting: Medicare's hospital compare." *Journal of the American Statistical Association* 112.519 (2017): 933-947.

5. O'Malley, A. James, et al. "Case-mix adjustment of the CAHPS® Hospital Survey." *Health services research* 40.6p2 (2005): 2162-2181.

6. Zaslavsky, Alan M. "Issues in case-mix adjustment of measures of the quality of health plans." *Proceedings, Government and Social Statistics Sections* (1998).

Appendix

Exploratory Analysis of Mortality Rates

Our first examination was in looking at the distribution of observed mortality rates across hospitals and procedure types. From Figure 4, we see that zero mortality rates are most often observed for low number of procedures. This creates an additional challenge because the low mortality rates may arise from the low volumes at these hospitals instead of better cardiovascular surgical procedures. If more patients in need of complex operations are treated at the low-volume hospitals, the mortality rates are likely to be high.

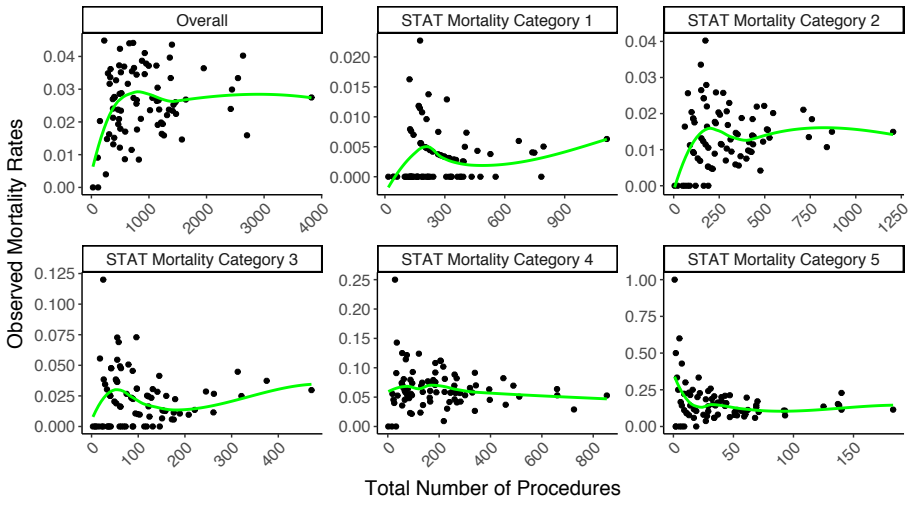

Figure 4: Observed mortality rates by hospital and procedure type.

Focusing on the observations with non-zero mortality rates, low volume hospitals tend to have higher mortality rates than large hospitals. In addition, we observe the highest mortality rates for STAT Mortality category 5 and the lowest for category 1.

The Influence of Hospital Attributes

We wanted to select a set of variables for case-mix adjustment from the five available hospital attributes listed in Table 2. Following the practice of O'Malley et al.[5], our criterion for selection of case-mix adjustors is the *impact factor*[6], which is the product of two measures: predictive power (the strength of the relationship between the candidate adjustor and the outcome variable) and heterogeneity factor (the amount of variation among hospitals in the adjustor variable). The predictive power in Table 2 is computed as the decrease in deviance due to the addition of the given variable. Variation in Table 2 is obtained by first stratifying the hospitals based on the given variable and then compute the F-value of the one-way ANOVA test among the varying strata. We required a minimum impact factor of 1 for a variable to be included. Therefore, we only chose "Number of Beds" and "Cardiovascular Surgery Rating" to include in our joint model. The estimated posterior means and 95 % credible intervals of "Number of Beds" and "Cardiovascular Surgery Rating" are $-0.1(-0.26, 0.07)$ and $0(0, 0)$. Therefore, we chose not to include them in our final model.

Hospital Attributes	Predictive Power	Variation	Impact Factor
NICU Level	0.103	0.083	0.008
Number of Beds	6.888	0.811	5.586
Cardiovascular Surgery Rating	10.717	1.189	12.743
Case-Mix Index	0.0295	0.917	0.027
Urban/Rural	0.243	0.928	0.225

Table 2: Computed impact factors of the hospital attributes for case-mix adjustment.

Predicted Mortality Rates

We plotted the predicted mortality rates by hospital and procedure type with our primary model in Figure 5. Note that Figure 5 roughly captures the trend in Figure 4, although the mortality rates of procedure 3, 4 and 5 in low volume hospitals are underestimated by our model due to shrinkage issues. Improving the shrinkage here is a future direction we are interested to work on.

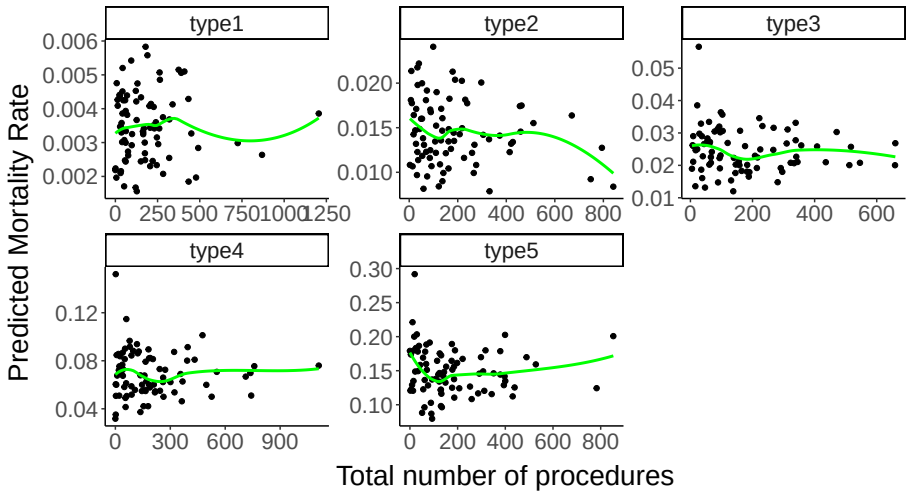

Figure 5: Predicted mortality rates by hospital and procedure type.

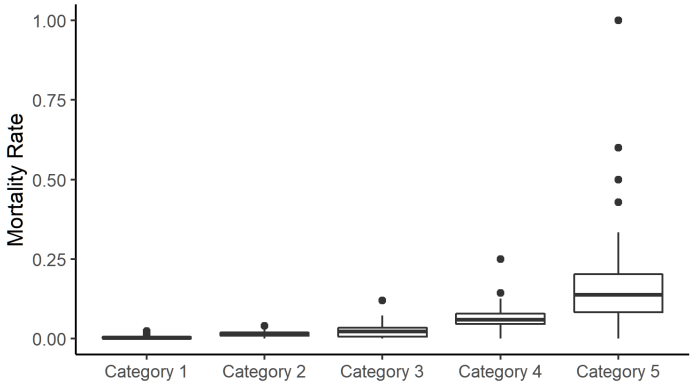

Figure 6: Mortality rate within each procedure category. Hospital mortality rates increase with the increasing level of STAT Mortality Category.

Sensitivity Analysis

We determine the robustness of our method by examining the extent to which results are affected by changes in the volume threshold τ and, consequently, the cluster membership discussed in Section 3. Table 3 shows that there are not noticeable discrepancies in posterior parameter estimates for the varying level of τ. Therefore, our estimates of λ and $\beta_{c_h,j}$ are quite stable and we are confident with our findings based on the two parameters.

Volume Threshold τ	λ	$\beta_{1,1}$	$\beta_{1,2}$	$\beta_{1,3}$	$\beta_{1,4}$	$\beta_{1,5}$	$\beta_{2,1}$	$\beta_{2,2}$	$\beta_{2,3}$	$\beta_{2,4}$	$\beta_{2,5}$
0.1	-3.44	-4.66	-2.98	-1.52	-0.80	-0.18	-3.53	-2.10	-1.67	-0.58	0.29
0.2	-3.20	-4.39	-2.68	-1.68	-0.54	0.29	-3.35	-1.93	-1.51	-0.44	0.43
0.3	-3.05	-3.49	-2.30	-1.52	-0.48	0.23	-3.27	-1.80	-1.39	-0.31	0.58
0.4	-2.71	-3.13	-1.92	-1.23	-0.31	0.62	-2.98	-1.49	-1.09	0.02	0.87
0.5	-2.85	-3.15	-2.00	-1.29	-0.33	0.5	-3.13	-1.58	-1.22	-0.10	0.78

Table 3: Effects of volume threshold τ on posterior means of model parameters.

We also assessed the robustness of our ranking system by checking whether the hospital rankings change with different values of τ. Table 4 displays the best hospital for the given procedure type predicted by our joint model. Although the best hospitals vary slightly among different τ's, they are primarily limited to the "Texas Children's Hospital", "UF Health Shands Children's Hospital" and "Helen DeVos Children's Hospital". This suggests that the estimated rankings by our model are fairly consistent.

Fitted Results

Table 5 shows the selected parameters credible intervals and convergence diagnostics statistics. The result is based on 3 chains, 5000 iterations and 1000 warm-up per chain.

Convergence Diagnostic

We show the traceplot of λ and β from the primary model. As shown in Figure 7, MCMC samples show good mixing and good convergence as well.

	0.1	0.2	0.3
Best for Type 1	Helen DeVos Children's Hospital	Helen DeVos Children's Hospital	UF Health Shands Children's Hospital
Best for Type 2	Helen DeVos Children's Hospital	Helen DeVos Children's Hospital	Helen DeVos Children's Hospital
Best for Type 3	Texas Children's Hospital	UF Health Shands Children's Hospital	Texas Children's Hospital
Best for Type 4	UF Health Shands Children's Hospital	Texas Children's Hospital	Texas Children's Hospital
Best for Type 5	Penn State Children's Hospital	UF Health Shands Children's Hospital	Helen DeVos Children's Hospital
	0.4	0.5	
Best for Type 1	Texas Children's Hospital	Texas Children's Hospital	
Best for Type 2	Helen DeVos Children's Hospital	Helen DeVos Children's Hospital	
Best for Type 3	Texas Children's Hospital	Texas Children's Hospital	
Best for Type 4	Helen DeVos Children's Hospital	Texas Children's Hospital	
Best for Type 5	Texas Children's Hospital	Texas Children's Hospital	

Table 4: Effects of volume threshold τ on hospital ranking for different procedures.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
lambda	-3.20	0.04	0.58	-4.57	-3.51	-3.11	-2.79	-2.32	170.03	1.01
beta[1,1]	-4.39	0.03	0.63	-5.69	-4.80	-4.37	-3.96	-3.23	508.02	1.00
beta[1,2]	-2.68	0.02	0.40	-3.47	-2.95	-2.68	-2.40	-1.89	263.20	1.01
beta[1,3]	-1.68	0.02	0.42	-2.52	-1.97	-1.68	-1.39	-0.87	300.38	1.01
beta[1,4]	-0.54	0.02	0.35	-1.20	-0.78	-0.55	-0.31	0.15	198.52	1.01
beta[1,5]	0.29	0.02	0.41	-0.51	0.00	0.29	0.57	1.09	287.17	1.01
beta[2,1]	-3.35	0.03	0.39	-4.11	-3.61	-3.36	-3.09	-2.56	174.92	1.01
beta[2,2]	-1.93	0.03	0.38	-2.65	-2.18	-1.93	-1.68	-1.17	163.04	1.01
beta[2,3]	-1.51	0.03	0.38	-2.24	-1.77	-1.52	-1.26	-0.76	165.56	1.01
beta[2,4]	-0.44	0.03	0.38	-1.15	-0.70	-0.45	-0.19	0.31	160.57	1.01
beta[2,5]	0.43	0.03	0.38	-0.28	0.18	0.42	0.68	1.19	163.55	1.01
varBeta	2.30	0.02	0.62	1.38	1.86	2.20	2.63	3.79	961.46	1.00
varLambda	4.52	0.09	6.66	1.18	2.08	3.04	4.83	16.57	4971.00	1.00

Table 5: Selected parameters credible intervals and convergence diagnostics statistics. The result is based on 3 chains, 5000 iterations and 1000 warm-up per chain.

Posterior Predictive Check

We show the posterior predictive check for the mortality rate and the number of death cases. We also examine the distribution of mean and standard deviation for the mortality rate and the number of death cases. Other than the variance of the mortality rate, the replicated data indicates that the model fit is a reasonable one. There is strong zero-inflation for procedure type 1 and procedure type 3 in terms of the mortality rate. We consider that accounting for this pattern may improve the fitting of variation among mortality rate and leave it to future directions.

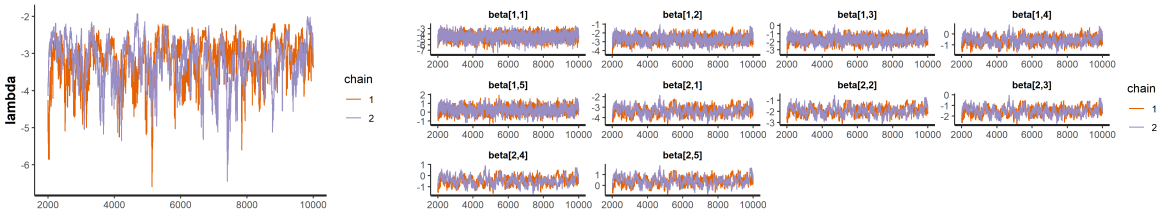

Figure 7: Traceplot of λ and β in the primary model. MCMC samples show good mixing and good convergence as well.

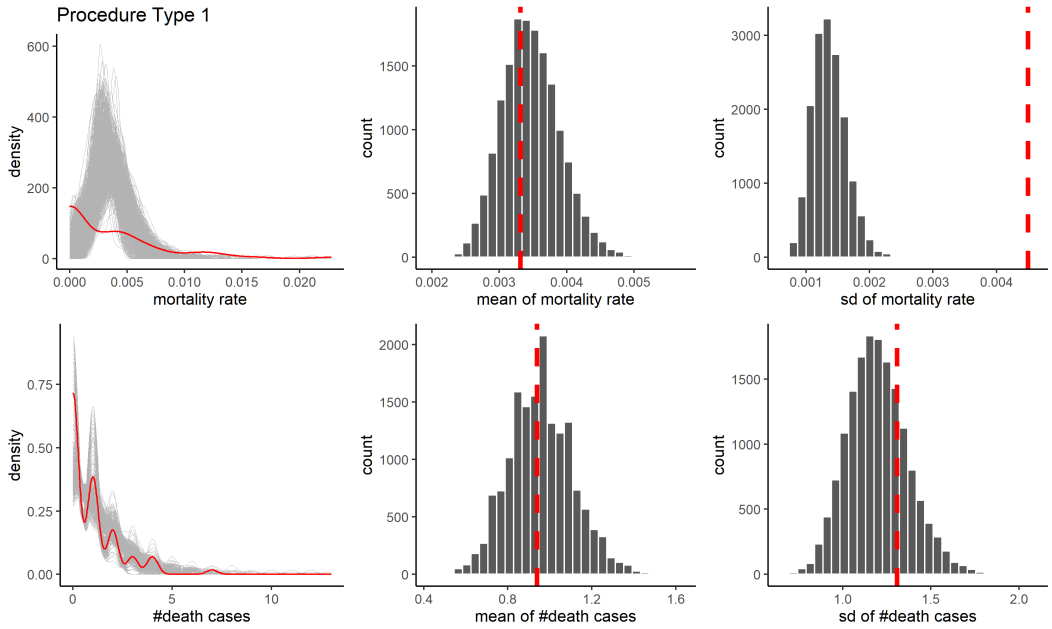

Figure 8: Posterior predictive check for the mortality rate and number of death cases for procedure type 1.

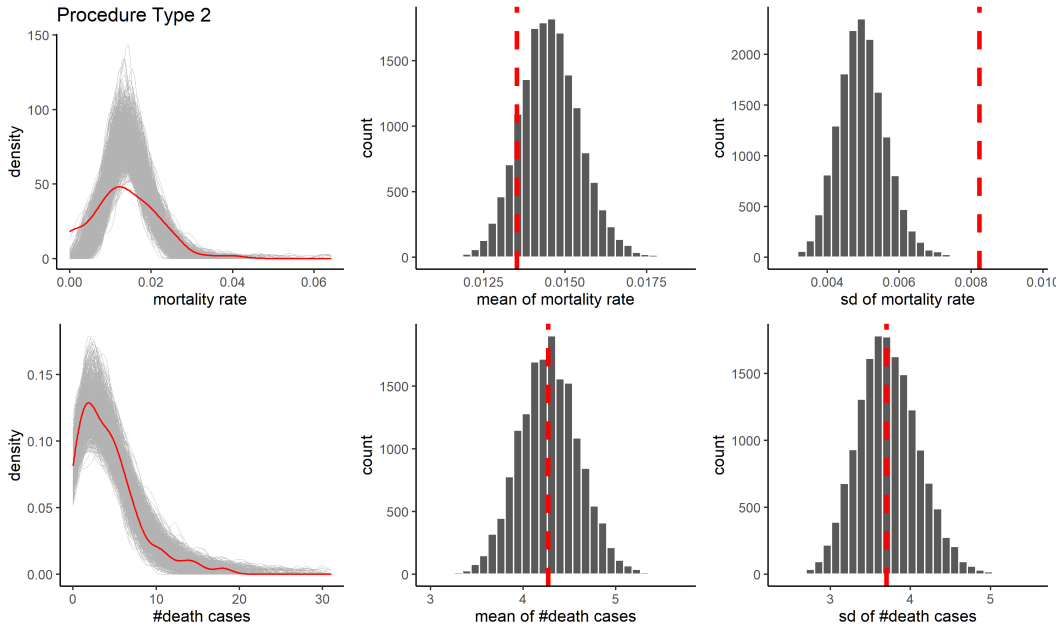

Figure 9: Posterior predictive check for the mortality rate and number of death cases for procedure type 2.

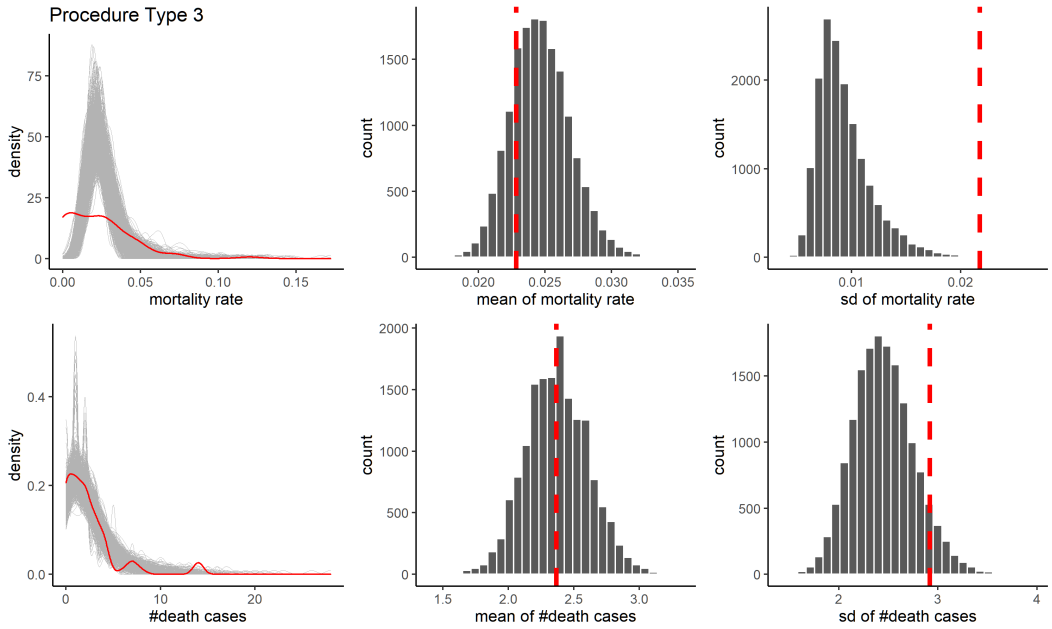

Figure 10: Posterior predictive check for the mortality rate and number of death cases for procedure type 3.

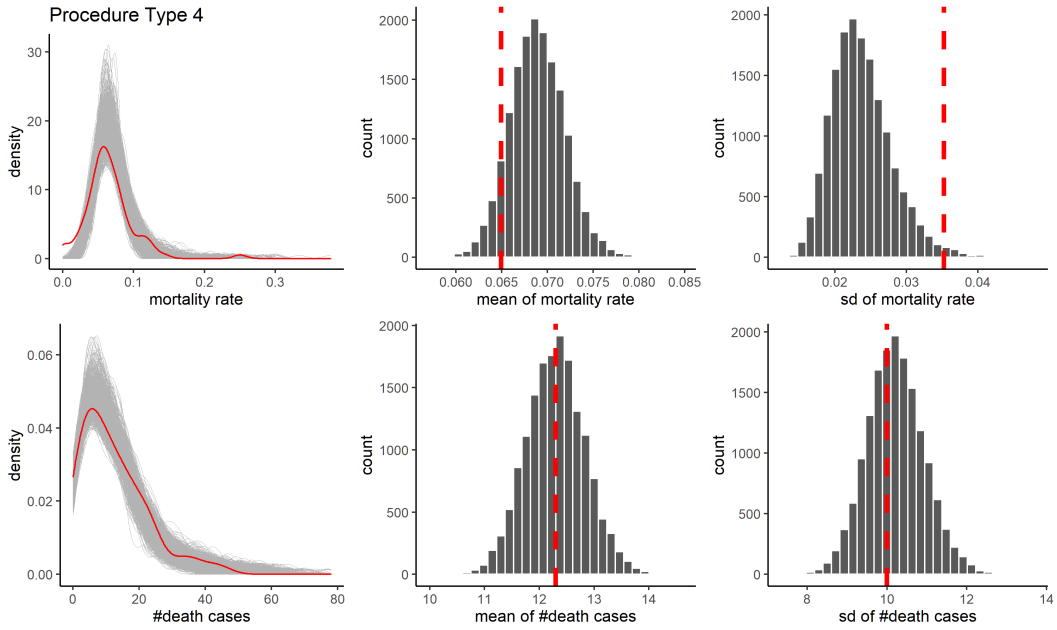

Figure 11: Posterior predictive check for the mortality rate and number of death cases for procedure type 4.

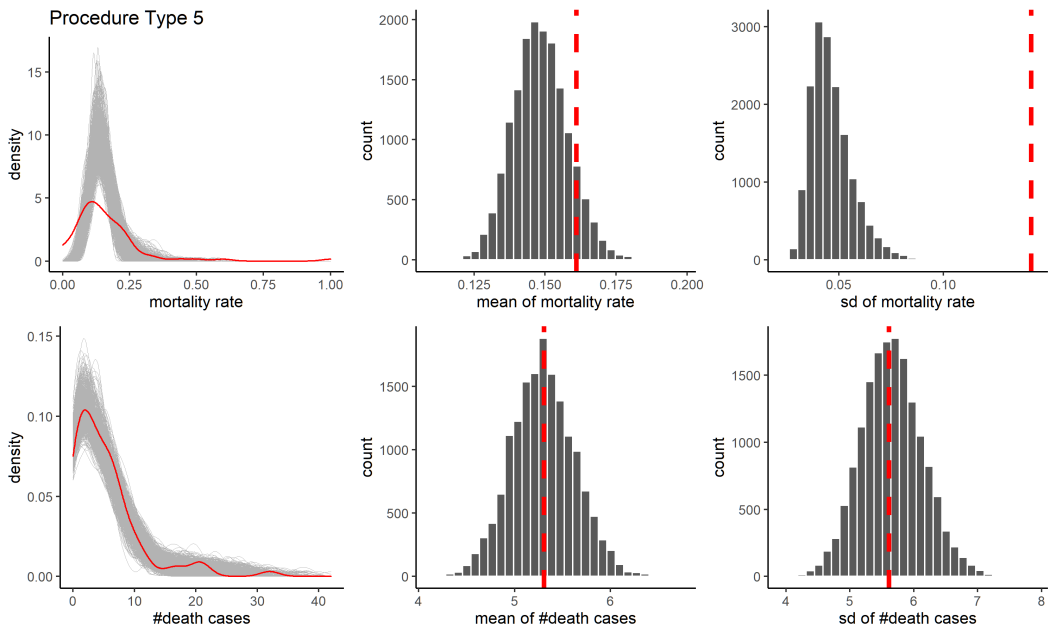

Figure 12: Posterior predictive check for the mortality rate and number of death cases for procedure type 5.