Case Study 2

Yunran Chen, Yiwei Gong, Siqi Fu, Lingxi Song

Introduction

Since New York Times's exposed the scandal of UNC Hospital's pediatric cardiovascular surgical program, considerable attention has been drawn to the evaluation of hospitals' performance on complex surgical programs. A common evaluation system is based on the ratio of observed and expected mortality rate (O/E score), where higher ratio suggests worse performance of a hospital.

Investigating surgical programs is difficult for numerous elements may influence patient complications. A success of surgical procedure require efforts from surgeons, anesthesiologists, intensive care doctors, and support staff as well as the condition of the patient themselves. A mistake on any of these parts may lead to issues and even deaths. Unfortunately, these elements are unavailable and most of them can hardly be quantified, which brings high variance on the mortality rate across hospitals and even within the same hospital.

Building a hierarchical model to borrow information across hospitals can make effective use of the provided information. However, such information sharing needs to be conducted in a clever way due to (1)the volume-based performance of each hospital; (2) the case-mix pattern of each hospital. We aim to provide expected mortality rate well addressing the aforementioned problems and construct an interval estimator of the O/E score (the ratio of observed and expected mortality rate) to evaluate pediatric heart surgery programs, which may benefit patients' decision making and supervise the quality of each program.

We build a Bayesian hierarchical model based on the dataset from STS[1], which includes 75999 cases for 83 hospitals mortality data of pediatric heart surgery of neonates, infants and children from 2015-2018. It includes the number of observed death and total procedures by the STAT category of complexity. And it also provides expected mortality rate adjusted by procedural and patient-level factors for each STAT category of complexity within each hospital, which is based on STS CHSD mortality risk model[2].

Exploratory Data Analysis

Most hospitals have relative small volume and the volume is related to the mortality rate of each hospital. Since patients tend to choose programs with high volume, around 64% hospitals do less than 250 procedures per year, which is a relatively small number. Figure 1 shows the scatter plot of observed versus expected mortality rate.[3] As number of procedures increases, the mortality rate in category 5 decreases significantly, the variance of mortality rate decreases for all 5 categories. A small volume may have two effects: Firstly, surgical teams cannot get enough practice to keep their high skills. Less resources will be allocated to the program to ensure the patients' recovery. Therefore, a lower volume of surgeries is related to a higher mortality rate. Secondly, low volume decreases the power of statistical results, which will cover potential problems. A shrinkage effect is needed to allow us borrow information across hospitals but we need to take care of volume effects

[1] https://publicreporting.sts.org/chsd-exp

[2] https://publicreporting.sts.org/chsd-risk-model

[3] We marked the outliers of STAT Mortality Category 1-5, which are 1.5×IQR above the 75 percentile and below 25 percentile

on the mortality rate, which suggests the information should be shared only within the hospitals with similar volumes.

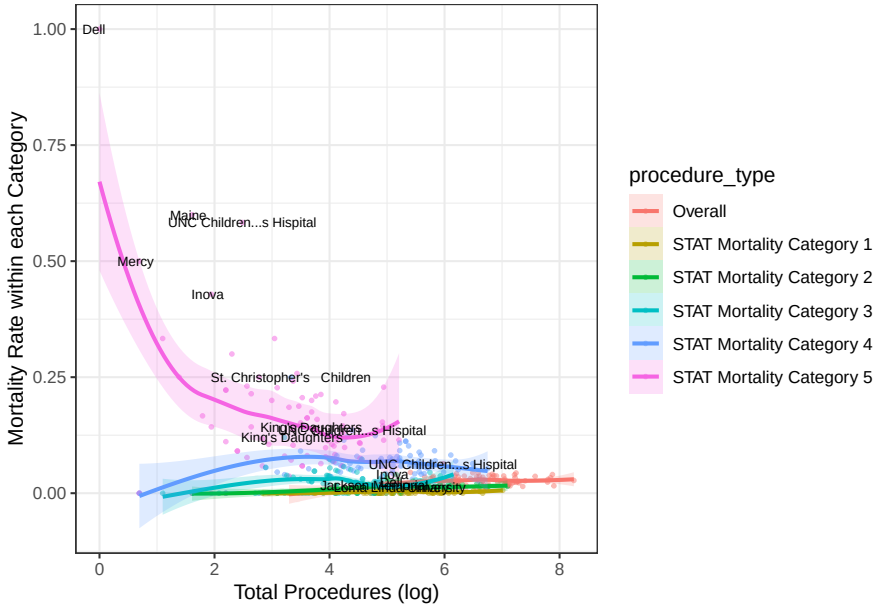

Figure 1: Category-specific Observed Mortality Rates versus Number of Total Procedures

The mortality rate of different procedure complexity is clearly different as shown in Figure 2. Procedures with a higher level of complexity have a higher mortality rate and longer tail (especially for category 5), which suggest introducing a fix effect for procedure category and setting a relative heavy-tail prior on these coefficients.

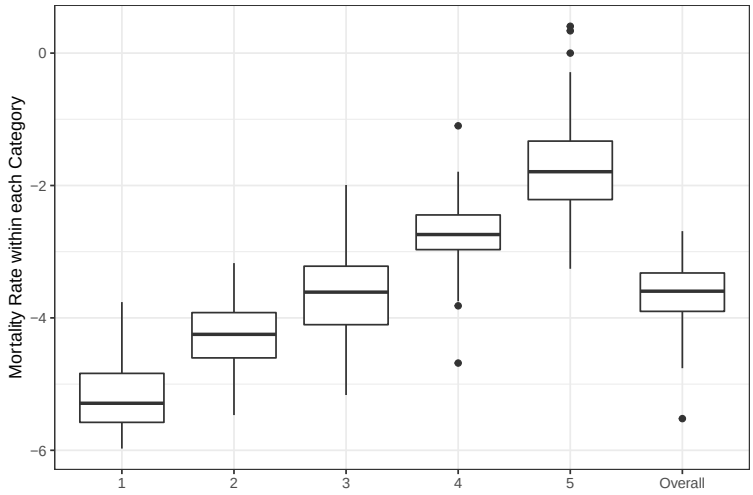

Figure 2: Ordinal Effects of Procedure Complexity

There exists case mix within each hospital. Figure 3 shows the proportion of categories of each hospital versus the total volume of the hospital. The hospital with high reputation tends to attract more complex cases. We see a significant decline of category 1 procedures and increase of category

2 and 4 category procedures. Therefore, the shrinkage should be based on the volume under each category instead of total volume.

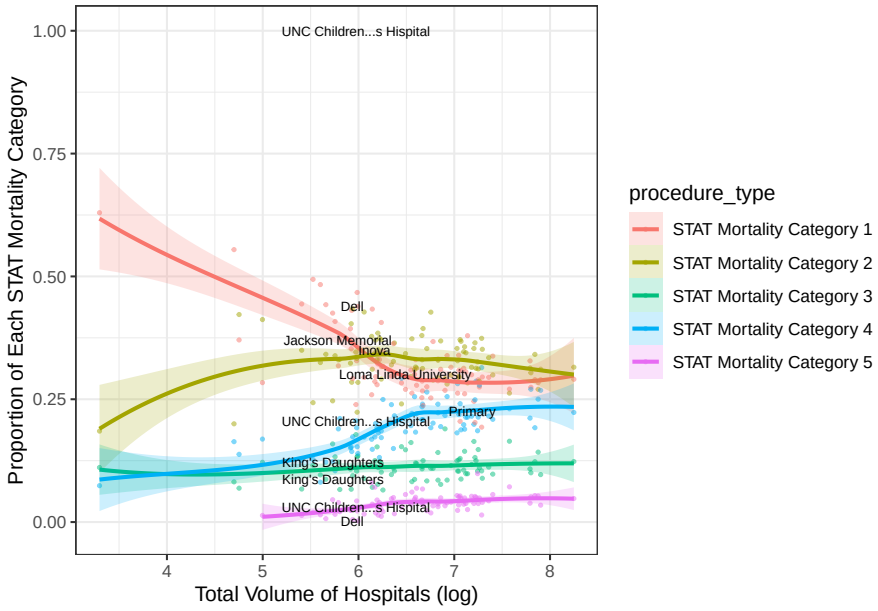

Figure 3: Case Mix: number of total procedures vs proportion of STAT categories

Model

In order to address the aforementioned challenges, we apply a Bayesian hierarchical model, introducing fix effects of procedure type with different priors for distinct categories and considering a shrinkage based on the logarithm of volume within each type. For the shrinkage effects across hospitals, we assume hospitals with similar number of type-specific volume have similar mortality and thus can share information together. Therefore, we consider a random intercept and random effect of `log(total_procedure)` in our model. We introduce fix effects of procedure types with heavy-tail t-distribution. Specifically, we consider t distribution with 10 and 3 degree of freedom . See Figure 4 for comparison across different shrinkage. We also compare models under different priors, which suggests the resulting estimator is not sensitive to prior settings (See Appendix for more details).

$$logit(P(Y_{hi}=1)) = \beta_0 + \beta_1 I(Type2)_{hi} + \beta_2 I(Type3)_{hi} + \beta_3 I(Type4)_{hi} + \beta_4 I(Type5)_{hi} + \beta_5 \log(TypeVol)_h + b_{0h} + b_{1h}\log(TypeVol)_{hi}$$

where i refer to procedure i, h refer to the hospital h. $Y_{hi}=1$ indicate a failed procedure i (a death) in hospital h, $I(Typek)_{hi}$ indicates STAT Mortality Category k of each procedure i in hospital h, and $log(TypeVol)_h$ represents the logarithm of type-specific volume. We assume $\beta_0 \sim T_3(0,10) \perp \beta_1 \sim T_3(0,10) \perp \beta_2 \sim T_3(0,10) \perp \beta_3 \sim T_3(0,10) \perp \beta_4 \sim T_3(0,10) \perp \beta_5 \sim T_3(0,10)$, and consider $\begin{pmatrix} b_{0h} \\ b_{1h} \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{11} & \tau_{12} \\ \tau_{12} & \tau_{22} \end{pmatrix})$(i.i.d), where correlated matrix comes from $LKJ(1)$ and variances come from half-cachy prior $(0,5)$.

Our model shrinks towards the right direction. Figure 4 shows model comparison between different shrinkage across hospitals, where red, blue and green indicate the type-specific expected mortality rate, x-axis represents the type-specific volume. We can see all the models shrink the observed mortality rates towards the expected mortality rates and hospitals with a lower type-specific volume were shrunk more. However, the ordinary shrinkage model, total volume-based shrinkage model and the model provided by STS seem to shrink too much since estimated mortality rates tend to increase as type-specific volume increases when the type-specific volume is relatively small, which is contradict to the intuition and the related scientific studies. In contrast, based on our model, the estimated mortality rates tend to decrease as type-specific volume increase and remain stable if type-specific hit a relative large value. And our model performs slightly better in terms of waic (1569.509), compared to ordinary shrinkage model (1580) and total volume-based shrinkage model (1578).

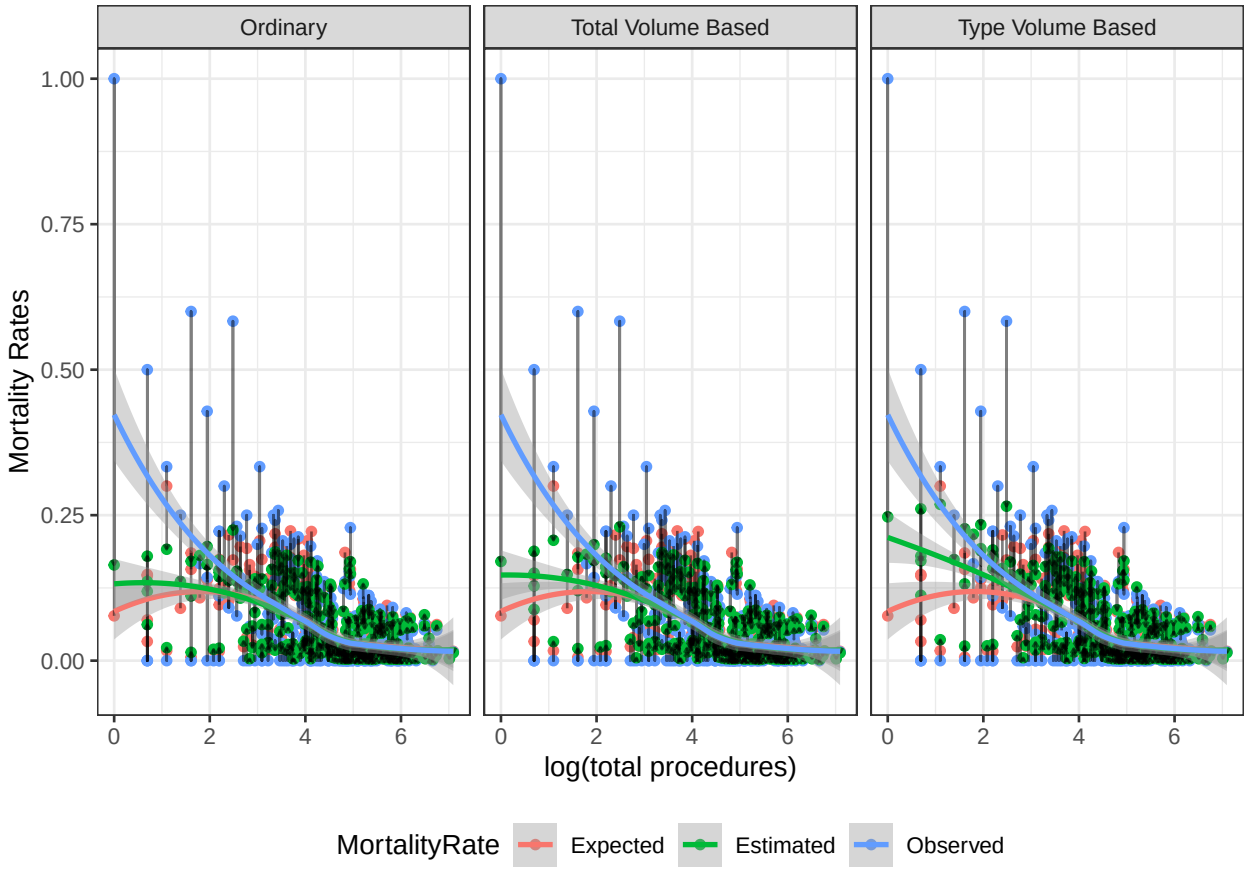

Figure 4: Comparison of Different Shrinkage Effects

Estimation and Inference

Table 1 shows estimations of our model parameters. For a typical hospital with only 1 type 1 procedure, the odds of mortality rate for type 1 procedures is about 0.4% to 1.4%. Holding the type-specific volume constant, for a typical hospital, the odds of the mortality of type 2 procedure is 3.4 to 5.5 times of that of type 1 procedure. Similarly, the odds of the mortality of type 3,4,5 procedure is 4.4 to 7.6, 14.7 to 23.4 compared to that of type 1 procedure. For a typical hospital, increasing the number of a certain type of procedure by 10% will decrease the odds of mortality

rate by 0.5% to 2.4% holding the number of procedures at the same risk level as constant. The variation across hospital are large and has a long right tail suggesting a large number of hospitals have high mortality rate. As shown in Figure 5, the variation of the potential effect of type-specific volume on type-specific mortality rate concentrates around 0. The two-modal shape may suggest a small proportion of hospitals learn from experience but most hospitals do not.

Referred from the credible interval, both the effects of procedure type and type-specific volume are significantly non-zero; a clear heterogeneity across hospital exists. We also fit a model with interaction term between procedure type and type-specific volume, which shows our model performs better based on since the logarithm of Bayes factor (0.85) is greater than 0.5, suggesting a substantial improvement and confidence inverals of all interaction terms include 0.

Table 1: Point Estimate and CI for Random and Fixed Effect

	Estimate	Est.Error	l.95..CI	u.95..CI
Intercept	-4.87	0.31	-5.44	-4.25
procedure_typeSTATMortalityCategory2	1.46	0.12	1.22	1.70
procedure_typeSTATMortalityCategory3	1.76	0.14	1.47	2.03
procedure_typeSTATMortalityCategory4	2.93	0.12	2.69	3.15
procedure_typeSTATMortalityCategory5	3.56	0.16	3.25	3.85
logtotal_procedures	-0.15	0.05	-0.26	-0.05
sd(Intercept)	0.45	0.23	0.10	0.95
sd(logtotal_procedures)	0.06	0.05	0.00	0.17
cor(Intercept,logtotal_procedures)	-0.50	0.52	-0.97	0.82

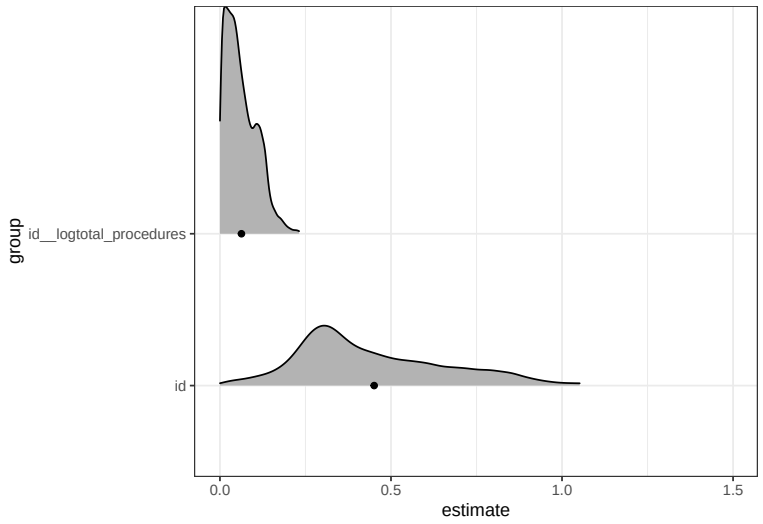

Figure 5: Density plot of Variance Estimation

We futher applied our model to rank hospitals based on the ratio of observed to expected mortality rate (O/E score) and provide stars based on 95% credible intervals of O/E scores. Table 2 and Table 3 list the top and bottom 10 hospitals.

Table 2: Top Ten Best Hospital

hospital_name	OE	OE_l	OE_u	star
Geisinger Medical Center	0.000	0.000	0.000	3
University of Kentucky Healthcare	0.000	0.000	0.000	3
Connecticut Children's Medical Center	0.233	0.144	0.450	3
Nemours Children's Hospital	0.442	0.260	0.940	3
Penn State Children's Hospital	0.499	0.334	0.846	3
UF Health Shands Children's Hospital	0.525	0.353	0.880	3
University of Maryland Children's Hospital	0.614	0.416	1.016	2
Helen DeVos Children's Hospital	0.636	0.433	1.014	2
University of Wisconsin Hospitals and Clinics	0.665	0.434	1.103	2
Mercy Medical Center	0.683	0.440	1.136	2

Table 3: Top Ten Killing Hospital

hospital_name	OE	OE_l	OE_u	star
St. Christopher's Hospital for Children	1.440	0.923	2.422	2
Children's Hospital of the King's Daughters	1.305	0.794	2.165	2
UNC Children's Hispital	1.282	0.903	1.930	2
Inova Children's Hospital	1.259	0.872	1.916	2
Mount Sinai Hospital	1.196	0.832	1.810	2
Maine Medical Center	1.196	0.732	2.064	2
Jackson Memorial Hospital	1.169	0.786	1.842	2
Dell Children's Medical Center	1.164	0.777	1.878	2
University of Minnesota Masonic Children's Hospital	1.161	0.807	1.744	2
Children's Hospital New Orleans	1.142	0.832	1.609	2

Evaluation on UNC Program

As a response to the New York Times' investigation in May 2019, the UNC Health Care announced to suspended complex heart surgeries (STAT 4 and 5 cases), which have higher mortality rates compared to other hospitals shown in the dataset. However, it argued in a published report that larger programs have lower morbidity and mortality in general while reporting mortality rate as in percentages doesn't do fair to small market share hospitals. UNC has on average 98 pediatric cardiac surgeries (STAT 1-5) annually, which is relatively small compared to the ideal range [100,150] suggested by the American Board of Thoracic Surgery (ABTS)[4]. We aim to utilize our model to evaluate the UNC program based on the adjusted estimator.

As Table 3 shows, UNC is ranked as bottom 3 based on O/E score. We further quantify the ranking based on posterior samples. The probability of UNC ranked in the bottom 5 based on O/E scores is 0.86 (Duke is 0.23), suggesting UNC have higher O/E scores compared to other hospitals. We further compare the posterior mean of random effects of UNC to those of other hospitals, which is shown in Figure 6. The light blue indicate estimated random effects of UNC

[4]https://www.abts.org

children's Hospital($b_{0,UNC}$ and $b_{1,UNC}$ in the model), and we also marked Duke University Hospital by duke blue as a reference. As illustrated in this plot, the point estimate $b_{0,UNC}$ and $b_{1,UNC}$ are relatively extreme, suggesting UNC is more likely to fail a produre than a typical hospital and less likely to learn from experience than a typical hospital. Thus we are not surprising to see it nominated as the 10 "killing" hospitals.

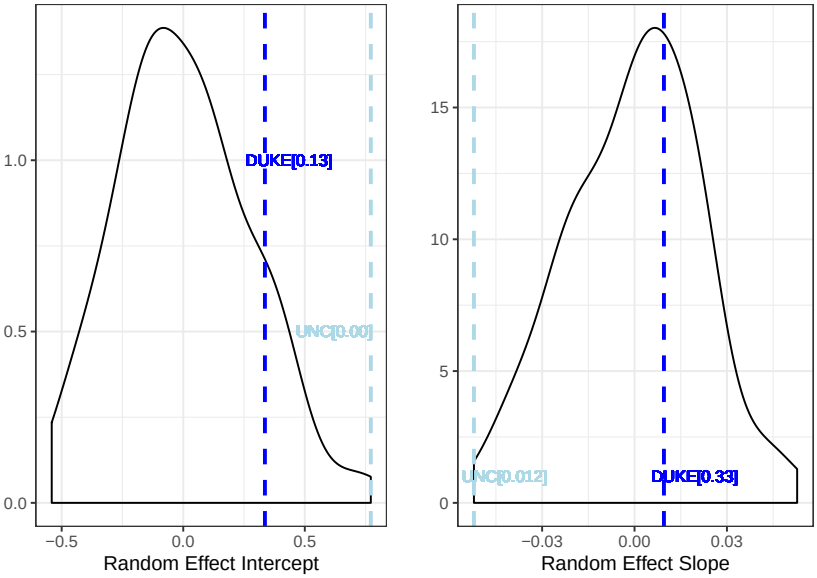

Figure 6: distribution of estimated random effects

Conclusion

Our model show excellent performance especially on addressing the volume-based shrinkage and considering case mix problem. Our model fits the data well as shown by model diagnosis and comparison among alternative models. Our model is valuable for benefiting patients' decision making and supervising the quality of each program. As suggested by our model, UNC performs worse compared to other programs which is also validated by the investigation of New York's Times. Apart from UNC hospitals, other hospitals in the bottom 10 list are also needed attentions. Since a success of complex surgery require efforts from both dedicated surgical team and recovery units, a more practical improvement is to merge competing programs in a nearby neighborhood for optimal resource allocation. For a small program unable to handle complex surgeries, enhance the patient referrals to nearby better programs.

Appendix

Sensitive Analysis on Prior Settings

To futher improve our model, we set several different priors based on the model6, and we compared the `waic` and `bayes_ratio` of them. We built 4 models for 3 different priors:
1. In model61: we assume b_{0i} and $b_{1,i}$ are independent.
2. In model62: we set tau_{11} and tau_{22} follow $cauchy(0, 1)$. The default prior is $cauchy(0, 5)$. We set the smaller varaince to make the prior more imformative.
3. In model 63: For the risk levels 1 2 and 3, we decreased the varaince of t-distribution from 10 to 1 and increased the degree of freedom from 3 to 8 to make the prior more informative.
By comparing the `waic` table below, we fond that both model6 has smaller `waic` value than model61 and model62, and has very similar `waic` value with model63.
From the Bayes_ratio table, we found that our model6 is better than model61, model63, and only model62 is a little superior than model6, but it is at the "not worth more than a bare mention" level.
Compared two criterions, we found that the model result is not sensitive to the prior, so we decided use the default prior and use model6 as our final model.

Table 4: Bayes Factor comparision

model	BF
model61 $ model6	-0.8496145
model62 & model6	0.2374132
model63 & model6	-6.7481922
model64 & model6	-6.6177451

Table 5: waic comparision

models	waic
model63	1569.069
model64	1569.069
model6	1569.509
model62	1569.823
model61	1571.926