Development of a Congenital Heart Surgery Composite Quality Metric: Part 2 – Analytic Methods

Online Supplement

1 Detailed Methods

Statistical Modeling

Hospital-specific operative mortality rates, major complication rates among survivors, and distributions of LOS among survivors were estimated in a multivariate hierarchical model with hospital-specific random effects. The term multivariate refers to the fact that the three endpoints were analyzed together in a single model, not estimated one at a time in separate models. Random-effects refers to the assumption that the hospital-specific parameters of interest arise from a probability distribution defined by parameters that are also estimated in the modelling process. The strategy of modeling multiple endpoints jointly was intended to improve estimation efficiency by borrowing information across multiple endpoints per patient both within and across hospitals.

Estimation of Risk Scores

In order to adjust for case mix, we first estimated a set of risk scores for predicting each of the three endpoints on the basis of preoperative prognostic factors. For operative mortality, risk scores were obtained by applying the published STS congenital mortality model methodology [ref]. Factors adjusted in the model include: age, weight among infants and neonates, prior cardiothoracic operation, any noncardiac congenital anatomic abnormality, any chromosomal abnormality or syndrome, prematurity, preoperative/preprocedural mechanical circulatory support, shock persistent at the time of the operation, renal dysfunction or renal failure requiring dialysis (or both), mechanical ventilation to treat cardiorespiratory failure, preoperative neurological deficit, any other preoperative factor, and strata defined by the cross-classification of primary procedure and age group (neonate, infant, child, adult). In accordance with the published methodology, coefficients for age group $\times$ primary procedure strata were estimated using empirical Bayes shrinkage estimators with an auxiliary adjustment for STAT Mortality Categories [ref] in the modeling of shrinkage targets. A patient's mortality risk score was then calculated as $x_1\hat{\beta}_1 + x_2\hat{\beta}_2 + \ldots + x_q\hat{\beta}_q$ where x_i denotes the patient's numerical value for the i-th covariate and $\hat{\beta}_i$ denotes the corresponding estimated regression coefficient. An analogous method was used to create risk scores for major complications among operative mortality survivors and LOS among survivors. The form of the risk score model was a logistic regression for major complications and a linear regression for log(LOS). Covariates were identical to the mortality model with the exception that STS Morbidity Categories [ref] were used in place of STAT Mortality Categories in the modeling of shrinkage targets for shrinkage estimation. Before using risk scores to adjust for case mix, each risk score model's fit to the current study data was tested. Calibration was assessed by comparing observed versus expected outcomes overall and within subgroups based on deciles of predicted risk and by analyzing the distribution of LOS residuals.

Multivariate Hierarchical Model

For the i-th of n_j patients at the j-th hospital ($j = 1, 2, \ldots, N$), let Y_{1ji} be a binary indicator of operative mortality status (0=alive, 1=dead), Y_{2ji} be an indicator of major complications (0 = none, 1 = at least one), and let $Y_{3ji} = \log(\text{LOS}_{ji})$, where LOS_{ji} denotes the patient's length of stay truncated at 90 days. Risk scores for the k-th endpoint are denoted by x_{kji} where $k = 1$ refers to operative mortality, $k = 2$ refers to major complications, and $k = 3$ refers to LOS. Define:

$$\begin{array}{cl} \text{(operative mortality)} & \pi_{1ji} = \Pr(Y_{1ji} = 1 |\, x_{1ji}, \text{hospital} = j) \\ \text{(major complications)} & \pi_{2ji} = \Pr(Y_{2ji} = 1 |\, x_{1ji}, \text{hospital} = j, Y_{1ji} = 0) \\ \text{(LOS mean)} & \pi_{3ji} = E(Y_{3ji} = 1 |\, x_{1ji}, \text{hospital} = j, Y_{1ji} = 0) \\ \text{(LOS variance)} & \sigma^2 = V(Y_{3ji} = 1 |\, x_{1ji}, \text{hospital} = j, Y_{1ji} = 0). \end{array}$$

In words, π_{1ji} is the probability of mortality, π_{2ji} is the probability of major complications conditional on being an operative mortality survivor, π_{3ji} is the average log(LOS) conditional on being an operative mortality survivor, and σ^2 is the variance of log(LOS) conditional on being an operative mortality survivor (assumed constant). Variations in the π_{kji} are assumed to be described by a multivariate generalized linear mixed model with a logistic link function for mortality, a logistic link function for complications, and an identity link function for log(LOS). At the first level, we assume:

$$\begin{array}{cc} \text{(operative mortality)} & Y_{1ji} \overset{\text{ind}}{\sim} \text{Bernoulli}(\pi_{1ji}) \\ \text{(major complications)} & Y_{2ji} \overset{\text{ind}}{\sim} \text{Bernoulli}(\pi_{2ji}) \\ \text{(LOS)} & Y_{3ji} \overset{\text{ind}}{\sim} \text{Normal}(\pi_{3ji}, \sigma^2) \end{array}$$

where

$$\begin{array}{cr} \text{(operative mortality)} & \log\left(\frac{\pi_{1ji}}{1-\pi_{1ji}}\right) = \alpha_{1j} + x_{1ji}\beta_1 \\ \text{(major complication)} & \log\left(\frac{\pi_{2ji}}{1-\pi_{2ji}}\right) = \alpha_{2j} + x_{2ji}\beta_2 \\ \text{(LOS)} & \pi_{3ji} = \alpha_{3j} + x_{3ji}\beta_3 \end{array}$$

where $\alpha_{1j}, \alpha_{2j}, \alpha_{3j}$ denote a set of unknown hospital-specific random intercepts and $\beta_1, \beta_2, \beta_3$ denote a set of unknown regression coefficients. Within patients, the outcomes $Y_{1ji}, Y_{2ji}, Y_{3ji}$ are assumed to be conditionally independent given the parameters $\pi_{1ji}, \pi_{2ji}, \pi_{3ji}$. Outcomes of patients at different hospitals are assumed to be statistically independent, and outcomes of patients at the same hospital are assumed to be conditionally independent given $(\alpha_{1j}, \alpha_{2j}, \alpha_{3j})$. The assumption that $Y_{1ji}, Y_{2ji}, Y_{3ji}$ are conditionally independent given $\pi_{1ji}, \pi_{2ji}, \pi_{3ji}$ is likely to be violated in practice but is made in order to facilitate computation. Although the model assumes *conditional* independence between $Y_{1ji}, Y_{2ji}, Y_{3ji}$, the model does not assume *marginal* independence between these three variables, as the underlying parameters $\pi_{1ji}, \pi_{2ji}, \pi_{3ji}$ depend on random effects parameters which account for within-hospital correlation.

At the second level, variation in the α_{kj} parameters was modeled by assuming

$$(\alpha_{1j}, \alpha_{2j}, \alpha_{3j}) \overset{\text{iid}}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a trivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ and covariance $\boldsymbol{\Sigma} = (\sigma_{11}, \sigma_{12}, \sigma_{13}, \sigma_{22}, \sigma_{23}, \sigma_{33})$.

Estimation

Model parameters were estimated in a Bayesian framework by specifying a prior probability distribution for the unknown model parameters and using Markov Chain Monte Carlo (MCMC) simulations for inference. Briefly, Bayesian inference uses the language of probability to express beliefs about clinically

interesting hypotheses and quantities. The output of a Bayesian analysis is a probability distribution describing the most likely numerical estimates of unknown model parameters. MCMC simulations are used to generate representative samples of parameter values, which are then analyzed to create appropriate estimates and summary measures. Advantages of fully Bayesian estimation include the ability to perform inference about complex functions of unknown quantities and the ability to calculate the probability of any clinically interesting hypothesis (e.g., the probability that a given hospital's composite score is greater than the STS average). Unlike frequentist confidence intervals, Bayesian interval estimates (known as credible intervals [CrI's]) have an intuitively direct interpretation as an interval containing the true value with a specified probability (eg, 95%). Because our prior knowledge was limited, we specified a vague proper prior distribution that consisted of independent normal distributions for regression coefficients (β_1, β_2, β_3), an inverse gamma distribution for σ^2, and an inverse Wishart distribution for $\mathbf{\Sigma} = (\sigma_{11}, \sigma_{12}, \sigma_{13}, \sigma_{22}, \sigma_{23}, \sigma_{33})$. Posterior means and credible intervals were calculated using MCMC simulations as implemented in WinBUGS version 1.4.3 software. Posterior summaries were calculated by generating 175,000 sets of simulated parameter values after a burn-in period of 5,000 MCMC iterations to ensure convergence. Adequacy of the number of MCMC iterations was assessed by the methods of Raftery and Lewis (1992) and Geweke (1991) as implemented in the CODA add-on package for R statistical software. To facilitate subsequent data processing, we reduced the number of samples by retaining only 1 of every 25 MCMC iterations for a final sample size of 7000 MCMC iterations. Let θ_j denote the j-th hospital's composite score. The parameter θ_j was estimated as $\hat{\theta}_j = \sum_{l=1}^{7000} \theta_j^{(l)} \Big/ 7000$, where $\theta_j^{(l)}$ denotes the simulated values of θ_j at the l-th iteration of the MCMC procedure. A 95% Bayesian credible interval was obtained by calculating the 175-th lowest and 175-th highest values of θ_j across the 7000 simulated values.

Definition of Risk-Adjusted Outcome Metrics

Based on this model, the j-th hospital's risk-adjusted ratio (RAR) was defined for operative mortality (RAR_{MORT}), major complications (RAR_{COMP}), and LOS (RAR_{LOS}) as follows:

$$(\text{RAR}_{\text{MORT}}) \qquad \theta_{1j} = \frac{\sum_{i=1}^{n_j} \text{expit}(\alpha_{1j} + x'_{ji}\beta_1)}{\sum_{i=1}^{n_j} \text{expit}(\mu_1 + x'_{ji}\beta_1)}$$

$$(\text{RAR}_{\text{COMP}}) \qquad \theta_{2j} = \frac{\sum_{i=1}^{n_j} \text{expit}(\alpha_{2j} + x'_{ji}\beta_2)}{\sum_{i=1}^{n_j} \text{expit}(\mu_2 + x'_{ji}\beta_2)}$$

$$(\text{RAR}_{\text{LOS}}) \qquad \theta_{3j} = \exp(\alpha_{3j} - \mu_3).$$

The j-th hospitals risk-adjusted mortality rate (RAMR), risk-adjusted complication rate (RACR), and risk-adjusted median LOS were defined as $\text{RAMR}_j = \theta_{1j} \times \overline{Y}_1$, $\text{RACR}_j = \theta_{2j} \times \overline{Y}_2$, and $\text{RAMLOS}_j = \theta_{3j} \times \overline{Y}_3$, respectively, where $\overline{Y}_1$ denotes the overall aggregate observed rate of operative mortality in the study sample, $\overline{Y}_2$ denotes the overall aggregate observed rate of major complication among operative mortality survivors, and $\overline{Y}_3$ denotes the overall median LOS among operative mortality survivors. The j-th hospital's composite score was defined by the formula

$$\theta_j = \frac{[\theta_{1j} + (\theta_{2j} + \theta_3)/2]}{2}.$$

Methods for Figures 1 and 3 (Technical Details)

To create the top panel of Figure 1, the range of possible risk-adjusted mortality rate (RAMR) values (0 to 100%) was partitioned into 201 equally-sized categories with cutpoints 0, 0.5, 1.0, etc. Let κ_c denote the unknown number of hospital's (out of 100) that have true RAMR values falling in the interval $(\frac{c-1}{2}, \frac{c}{2})$,

for $c = 1, 2, \ldots, 200$. The posterior mean of each κ_c was calculated as $E[\kappa_c|\text{data}] = \frac{1}{7000}\sum_{l=1}^{7000}\kappa_c^{(l)}$ where 7000 is the number of MCMC iterations and $\kappa_c^{(l)}$ is the value of κ_c on the l-th MCMC iteration. The top panel of Figure 1 was obtained by plotting bars of width $(\frac{c-1}{2}, \frac{c}{2})$ and height $E[\kappa_c|\text{data}]$ over the range for which $E[\kappa_c|\text{data}] > 0$. Numerical summaries of the RAMR distribution were estimated as follows. Let $\gamma_p^{(l)}$ denote the empirical p-th percentile of the set of numbers $\text{RAMR}_1^{(l)}$, $\text{RAMR}_2^{(l)}$, ..., $\text{RAMR}_{100}^{(l)}$, where $\text{RAMR}_j^{(l)}$ denotes the simulated value of the RAMR of the j-th hospital on the l-th MCMC iteration. An estimate of the empirical p-th percentile of RAMR's was obtained as $\hat{\gamma}_p = (1/7000)\sum_{l=1}^{7000}\gamma_p^{(l)}$, where p = 10, 50 (median), or 90. Methods for the other two panels of Figure 1 were essentially identical to the top panel and are omitted for brevity.

To create the top left panel of Figure 3, we focused on the 9 hospitals that were classified as having worse-than-expected composite outcomes according to the methods described above and in the main methods section. Let κ_c denote the unknown number of these hospitals (out of 9) that have true RAMR's falling in the interval $(\frac{c-1}{2}, \frac{c}{2})$. The top left panel of Figure 3 was created by plotting bars of width $(\frac{c-1}{2}, \frac{c}{2})$ and height $E[\kappa_c|\text{data}]$ over the range for which $E[\kappa_c|\text{data}] > 0$ (see prior paragraph for details). Methods for the top middle and right panels of Figure 3 were exactly analogous. Methods for the bottom panels were identical except that they focused on the 16 hospitals that were classified as having better-than-expected composite outcomes.

Estimation of Reliability (Tables 2 and 3)

Calculations for Overall 4 Year Cohort

Reliability is conventionally defined as the proportion of variation in a measure that is due to true between-unit differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). Equivalently, it is the squared correlation between a measurement and the true value. Accordingly, reliability was defined as the square of the Pearson correlation coefficient (ρ^2) between the set of hospital-specific estimates $\hat{\theta}_1, \ldots, \hat{\theta}_N$ and the corresponding unknown true values $\theta_1, \ldots, \theta_N$, that is:

$$\rho^2 = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h)(\theta_j - \frac{1}{N}\sum_{h=1}^{N}\theta_h)}{\sum_{j=1}^{N}(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h)^2\sum_{j=1}^{N}(\theta_j - \frac{1}{N}\sum_{h=1}^{N}\theta_h)^2}$$

The quantity ρ^2 was estimated by its posterior mean, namely,

$$\hat{\rho}^2 = \frac{1}{7000}\sum_{l=1}^{7000}\rho_{(l)}^2$$

where

$$\rho_{(l)}^2 = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h)(\theta_j^{(l)} - \frac{1}{N}\sum_{h=1}^{N}\theta_h^{(l)})}{\sum_{j=1}^{N}(\hat{\theta}_j - \frac{1}{N}\sum_{h=1}^{N}\hat{\theta}_h)^2\sum_{j=1}^{N}(\theta_j^{(l)} - \frac{1}{N}\sum_{h=1}^{N}\theta_h^{(l)})^2}$$

with $\theta_h^{(l)}$ denoting the value of θ_j on the l-th MCMC sample $\hat{\theta}_j = \sum_{l=1}^{7000}\theta_j^{(l)}/7000$ denoting the posterior mean of θ_j. A 95% credible interval for ρ^2 was obtained by calculating the 125th smallest and 125th largest values of $\rho_{(l)}^2$ across the 7000 MCMC samples.

Reliability as a Function of the Measurement Window

We used Brown's prophecy formula to estimate the reliability that would be achieved hypothetically if composite scores were to be re-estimated with a narrower measurement window assuming each hospital's performance and case volumes remain constant over time. Suppose reliability using all 4

years of data is denoted by ρ. According to Brown's prophecy formula, reliability would be equal to $\rho k/(\rho k+(1-\rho))$ if reliability were to be estimated using a fraction k of each hospital's data. The numbers in the top row of Table 3 were calculated by substituting $\rho = 0.73$ and plugging in $k = 0.75$ for 3 years of data, $k = 0.50$ of 2 years of data, and $k = 0.25$ for 1 year of data. Note: Most publications discussing the Brown prophecy formula assume that all units have the same sample size, that the measure of interest is a simple average (no shrinkage or risk adjustment), and that the measure of interest is based on a continuous variable with constant error variance. Using basic probability arguments, the same formula can be derived assuming that (1) a hospital's # of eligible cases per unit of time is a random variable, (2) a hospital's # of eligible cases per unit time is potentially correlated with a hospital's true performance, (3) the measure of interest is a weighted average of multiple individual measures (e.g. mortality rate, complication, rate, average LOS), and (4) the error variance is not necessarily constant. The main additional assumption is that a hospital's number of eligible cases per unit time and its performance remain constant over time. However, this formula is only a rough approximation of actual reliability because it assumes that the individual measures being combined in the composite are simple averages and does not account for the use of shrinkage estimation or risk adjustment.

Reliability as a Function of Sample Size

In order to shed light on the minimum sufficient sample size for the estimation of composite scores, we estimated the reliability that would be achieved hypothetically if composite scores were to be estimated using data from a stratified random sample of n eligible operations per hospital. For each hospital, we assumed that n patients are randomly sampled from a large (conceptually infinite) population that is unique to that particular hospital. The distribution of risk scores in a hospital's population was assumed to be identical to the hospital's risk score distribution in the current analysis. In addition, hospital-specific outcomes were assumed to follow distributions described by this paper's multivariate hierarchical model. To make these calculations tractable, we assumed that composite scores are unadjusted for case mix and that estimation is based on a simple non-hierarchical modeling analysis. The formula for estimating the j-th hospital's composite score is:

$$\hat{\theta}_j^* = \frac{\bar{Y}_{1j}}{2c_1} + \frac{\bar{Y}_{2j}}{4c_2} + \frac{\exp(\bar{Y}_{3j})}{4\exp(c_3)}$$

where $\bar{Y}_{1j}$, $\bar{Y}_{2j}$, $\bar{Y}_{3j}$ are the j-th hospital's observed mortality rate, observed major complication rate, and observed average log(LOS), respectively, and c_1, c_2, c_3 are constants representing the average hospital-specific mortality rate, average hospital-specific major complications rate, and average hospital-specific mean log(LOS), respectively, across the $N = 100$ hospitals. We also assume that $\bar{Y}_{1j}$, $\bar{Y}_{2j}$, $\bar{Y}_{3j}$ are conditionally independent conditional on the set of hierarchical model parameters. The following additional

notation will be used in this section:

$$\bar{\pi}_{1j} = \frac{1}{n_j}\sum_{i=1}^{n_j}\pi_{1ji}, \quad \bar{\pi}_{2j} = \frac{1}{n_j}\sum_{i=1}^{n_j}\pi_{2ji}, \quad \bar{\pi}_{3j} = \frac{1}{n_j}\sum_{i=1}^{n_j}\pi_{3ji}$$

$$\bar{\bar{\pi}}_1 = \frac{1}{N}\sum_{j=1}^{N}\bar{\pi}_{1j}, \quad \bar{\bar{\pi}}_2 = \frac{1}{N}\sum_{j=1}^{N}\bar{\pi}_{2j}, \quad \bar{\bar{\pi}}_3 = \frac{1}{N}\sum_{j=1}^{N}\bar{\pi}_{3j},$$

$$\theta_j^* = \frac{\bar{\pi}_{1j}}{2\bar{\bar{\pi}}_1} + \frac{\bar{\pi}_{2j}}{4\bar{\bar{\pi}}_2} + \frac{\exp(\bar{\pi}_{3j})}{4\exp(\bar{\bar{\pi}}_3)}, \qquad \bar{\theta}^* = \frac{1}{N}\sum_{j=1}^{N}\theta_j^*$$

$$V_j(n) = \frac{\bar{\pi}_{1j}(1-\bar{\pi}_{1j})/n}{4{\bar{\bar{\pi}}_1}^2} + \frac{\bar{\pi}_{2j}(1-\bar{\pi}_{2j})/n}{16{\bar{\bar{\pi}}_2}^2} + \frac{[\exp(\sigma^2/n)-1]\exp(2\bar{\pi}_{3j}+\sigma^2/n)}{16[\exp(\bar{\bar{\pi}}_3)]^2}$$

$$B = \frac{1}{N}\sum_{j=1}^{N}(\theta_j^* - \bar{\theta}^*)^2, \quad \bar{V}_n = \frac{1}{N}\sum_{j=1}^{N}V_j(n)$$

In order to estimate reliability, we started with the reliability definition

$$\text{reliability} = \rho_n = \{\text{cor}[\hat{\theta}_J^*, E(\hat{\theta}_J^*)]\}^2$$

and then used mathematical properties of the binomial and normal distribution to derive the expression:

$$\rho_n = B/(B + \bar{V}_n)$$

where B and $\bar{V}_n$ are defined above. In the definition of reliability given above, J is a random variable representing the index of a randomly selected hospital. The expectation $E[\hat{\theta}_J^*]$ is taken over the random selection of a single hospital out of the $N = 100$ hospitals, the random sample of n operations from this hospital's population, and the set of all possible outcomes of these n operations. The expression $\rho_n = B/(B + \bar{V}_n)$ then follows from basic probability arguments under the assumption that

$$n\bar{Y}_{1j}|\bar{\pi}_{1j} \sim \text{Binomial}(n, \bar{\pi}_{1j})$$
$$n\bar{Y}_{2j}|\bar{\pi}_{2j} \sim \text{Binomial}(n, \bar{\pi}_{2j})$$
$$\bar{Y}_{3j}|\bar{\pi}_{3j} \sim \text{Normal}(\bar{\pi}_{3j}, \sigma^2/n).$$

The quantity ρ_n cannot be observed directly because it depends on the unknown π_{mji}'s and σ^2. Instead, we estimated ρ_n using MCMC methods. Let $\rho_n^{(l)}$ be the value obtained when ρ_n is calculated from the l-th set of randomly sampled parameter values from the MCMC procedure. Our estimate of ρ_n was the posterior mean

$$\hat{\rho}_n = \frac{1}{M}\sum_{l=1}^{M}\rho_n^{(l)}.$$

Methods for Quantifying Impact of a Change in a Single Endpoint

As noted above and in the manuscript body, the final composite score equation can be expressed as

$$\frac{1}{2} \times \text{RAR}_{\text{MORT}} + \frac{1}{4} \times \text{RAR}_{\text{COMP}} + \frac{1}{4} \times \text{RAR}_{\text{LOS}},$$

and each RARis interpreted as the ratio of observed to expected outcomes. In other words, it has the form

$$\frac{1}{2} \times \frac{\text{actual \% mortality}}{\text{expected \% mortality}} + \frac{1}{4} \times \frac{\text{actual major \% complications}}{\text{expected \% major complications}} + \frac{1}{4} \times \frac{\text{actual median LOS}}{\text{expected median LOS}}.$$

If a hospital's case mix was similar to the overall study population, then its expected outcomes would be 3.1% for mortality, 11.3% for major complications, and median 7 days for LOS, and so this hospital's composite score would have the form

$$\frac{1}{2} \times \frac{\text{actual \% mortality}}{3.1} + \frac{1}{4} \times \frac{\text{actual \% major complications}}{11.3} + \frac{1}{4} \times \frac{\text{actual median LOS}}{7}.$$

If this hospital's complication rate and LOS were exactly as expected but its mortality rate was 1 percentage point higher than expected (e.g. 4.1% actual mortality versus 3.1% expected mortality), then the hospital's composite score would be

$$\frac{1}{2} \times \left(\frac{3.1+1}{3.1}\right) + \frac{1}{4} \times (1) + \frac{1}{4} \times (1) = 1.16.$$

In order to determine the difference in major complications and LOS that would have the same impact as a 1 percentage point increase in mortality, we solved for δ_{COMP} and δ_{LOS} in the equations

$$\frac{1}{2} \times (1) + \frac{1}{4} \times \left(\frac{13.3+\delta_{\text{COMP}}}{13.3}\right) + \frac{1}{4} \times (1) = 1.16$$

and

$$\frac{1}{2} \times (1) + \frac{1}{4} \times (1) + \frac{1}{4} \times \left(\frac{7+\delta_{\text{LOS}}}{7}\right) = 1.16$$

to obtain $\delta_{\text{COMP}} = 7.2\%$ and $\delta_{\text{LOS}} = 4.5$ days. In other words, a 1 percentage point excess in mortality has the same impact as a 7.2 percentage point excess in major complications or a 4.5 day excess in LOS.

Impact of Change in a Single Endpoint in Selected Operations

As described above and in the main manuscript, we found that an absolute change of 1 percentage point in a hospital's adjusted mortality rate (e.g. 4.1% versus 3.1%) would have the same impact on the composite score as an absolute change of 7.2 percentage points in a hospital's adjusted complication rate or a change of 4.5 days in a hospi-tal's adjusted LOS. To further understand the influence of the individual component metrics on the overall composite, and to ensure that complications and LOS did not have an undue influence, we estimated the impact on the composite score of changes in a hospital's complication rate and LOS vs. mortality across several representative operations spanning the spectrum of case complexity: Tetralogy of Fallot (TOF) repair, arterial switch operation (ASO), and the Norwood operation.

Methods. We repeated the calculation of δ_{COMP} assuming a case mix typical for Norwood operations (expected median LOS = 32 days), ASO operations (expected median LOS = 13 days), and TOF repair operations (expected median LOS = 8 days), to obtain:

$$\frac{1}{2} \times (1) + \frac{1}{4} \times (1) + \frac{1}{4} \times \left(\frac{32+\delta_{\text{LOS}}^{\text{Norwood}}}{32}\right) = 1.16 \quad \Longrightarrow \quad \delta_{\text{LOS}}^{\text{Norwood}} \approx 20 \text{ days}$$

$$\frac{1}{2} \times (1) + \frac{1}{4} \times (1) + \frac{1}{4} \times \left(\frac{13+\delta_{\text{LOS}}^{\text{ASO}}}{13}\right) = 1.16 \quad \Longrightarrow \quad \delta_{\text{LOS}}^{\text{ASO}} \approx 8 \text{ days}$$

$$\frac{1}{2} \times (1) + \frac{1}{4} \times (1) + \frac{1}{4} \times \left(\frac{8+\delta_{\text{LOS}}^{\text{TOF}}}{8}\right) = 1.16 \quad \Longrightarrow \quad \delta_{\text{LOS}}^{\text{TOF}} \approx 5 \text{ days}.$$

Results. These changes in overall LOS on the hospital level equate to a change of $\sim$20 days for the Norwood operation (for reference, median Norwood LOS among survivors is 32 days, interquartile range 21–51 days), $\sim$8 days for the ASO (median ASO LOS among survivors is 13 days, interquartile range 9–19 days), and $\sim$5 days for TOF repair (median TOF repair LOS among survivors is 8 days, interquartile range 6-11 days). In other words, these changes in LOS are relatively large and essentially equate to moving from the median to near or beyond the interquartile range for all operations examined. Overall, these values suggest that complications and LOS do not have too great an influence on the composite measure with the final weighting scheme chosen.

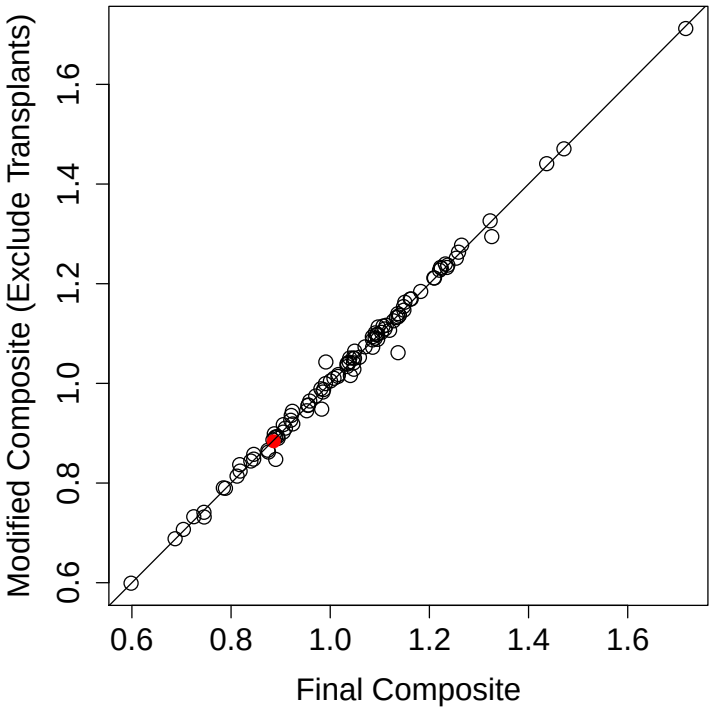
2 Length of Stay (LOS) Sensitivity Analysis

A sensitivity analysis was performed to understand the impact of the inclusion of LOS in the composite measure on hospitals whose typical practice involved keeping patients undergoing the Norwood operation in the hospital until Stage II. Based on the distribution of the data in the study population, hospitals with a high proportion of such patients were defined as those where $>$20% of their patients stayed in the hospital from the Norwood operation through Stage II palliation. In the five hospitals that met this criteria, we investigated whether their performance as assessed by the composite measure (classification as same, better, or worse-than expected) changed if LOS was included or excluded. None of these hospitals changed their performance category (see table below) suggesting that this practice does not negatively impact these hospitals, and supporting the retention of LOS in the composite.

	Hospital #1	**Hospital #2**	**Hospital #3**	**Hospital #4**	**Hospital #5**
Number of Norwoods	9	13	44	3	54
% of Norwoods kept in-house though Stage II	22.2%	23.1%	27.3%	33.3%	40.7%
Composite Score RAR — LOS included	1.11 (0.92, 1.33)	1.05 (0.88, 1.24)	0.99 (0.89, 1.11)	0.89 (0.68, 0.16)	0.72 (0.65, 0.81)
Composite Performance Category — LOS included	Same as Expected	Same as Expected	Same as Expected	Same as Expected	Better than Expected
Composite Score RAR — LOS excluded	1.17 (0.93, 1.47)	1.03 (0.81, 1.28)	0.97 (0.84, 1.13)	0.84 (0.56, 1.20)	0.60 (0.50, 0.71)
Composite Performance Category — LOS excluded	Same as Expected	Same as Expected	Same as Expected	Same as Expected	Better than Expected

3 Heart and Lung Transplant Sensitivity Analysis

A sensitivity analysis was performed to examine the influence of inclusion of heart and lung transplant procedures in the study population. We repeated the estimation of composite scores after excluding transplant procedures (heart transplant, lung transplant, and combined heart/lung transplant). As shown in the figure below, composite estimates calculated with and without the inclusion of transplant procedures were highly similar (correlation = 0.997). One hospital was re-classified from “better-than-expected performance” to“same-as-expected performance” after transplant procedures were excluded (red dot in figure). The lower limit of this hospital’s 95% credible interval fell a tiny amount above the STS average when transplants were included and was a tiny amount below the STS average after transplants were excluded. Based on these findings, transplant procedures were retained in the final study population to be consistent with other STS reporting conventions.



4 Composite Classifications by Center Volume and Case Mix

We examined the distribution of hospital performance categories across different categories of center volume and case-mix. With regard to volume, more hospitals in the higher vs. lower volume categories were classified in the better-than-expected category, which is anticipated given the known volume-outcome relationship demonstrated by many previous analyses in the field.

Volume Category	Number of Hospitals	Worse-Than-Expected	Same-As-Expected	Better-Than-Expected
$<$75	25	2 (8%)	22 (88%)	1 (4%)
75 – 149	24	3 (12%)	20 (83%)	1 (4%)
150 – 249	19	0 (0%)	17 (89%)	2 (11%)
250 – 349	17	3 (18%)	8 (47%)	6 (35%)
350+	15	1 (7%)	8 (53%)	6 (40%)
Total	100	9 (9%)	75 (75%)	16 (16%)

With regard to case-mix, there is no current gold standard for assessment. In the absence of this we used the metric of percent of STAT 5 cases. There was a generally similar distribution of center performance across different percentiles or categories of case-mix. This is anticipated based on the methodology used for quality measures both in the present analyses and across many other fields, which allows one to discern how a hospital is performing in relation to what would be expected for their particular case-mix, as described in the discussion of the main manuscript. Thus, hospitals performing better- or worse-than-expected can be found across all levels of case-mix as shown in these tables. It is also for this reason that such quality metrics are not meant to be used to "rank" hospitals with differing case-mix one against another, as it cannot be assumed for example that a hospital with a relatively low complexity case-mix would achieve the same performance if faced with a relatively high complexity case-mix.

Percent of STAT 5 Cases	Number of Hospitals	Worse-Than-Expected	Same-As-Expected	Better-Than-Expected
$<$2.4% (lowest 25% of hospitals)	25	3 (12%)	19 (76%)	3 (12%)
2.4% – 4.8% (middle 50% of hospitals)	50	4 (8%)	36 (72%)	10 (20%)
$\geq$4.9% (highest 25% of hospitals)	25	2 (8%)	20 (80%)	3 (12%)
Total	100	9 (9%)	75 (75%)	16 (16%)

5 Comparison of Alternative Composite Weighting Strategies

We evaluated several potential methods for calculating the composite score. For each of the methods, we assessed composite measure properties including the hospital-level correlations between each of the individual component metrics and the overall composite score (to examine clinical-face validity and the overall influence of the individual components on the composite score, reliability (signal-to-noise ratio), and the proportion of hospitals classified in different composite performance categories.

Selection of Candidate Composite Formulas

Our initial set of candidate composite formulas was based on combinations of:

- 3 methods of standardizing the individual metrics to account for unequal standard deviations, and
- 2 methods of weighting the individual metrics after they were standardized.

Standardization Methods.

As noted in the manuscript, the construction of a composite measure must account for unlike measurement scales (e.g. mortality measured by percentages compared to LOS measured in days). RAR's address this issue by producing a metric that is unitless and has a similar numerical interpretation for each endpoint as a ratio of observed to expected outcomes. However, even if all metrics or on the same scale numerically, their actual standard deviations may differ, and the resulting composite is likely to be influenced most heavily by items with the largest standard deviations. This may be desirable if items with large standard deviations are regarded as most important by users of the composite measure but may be undesirable otherwise. To address this issue, we considered 3 different transformations of the individual metrics which result in different ratios of standard deviations across them.

1. **Use unstandardized RAR's.** The resulting composite is a weighted average of RAR's.
2. **Normalize RAR's by dividing by their respective standard deviations.** The RAR for each endpoint is divided by its standard deviation across hospitals. The final composite is a weighted average of the normalized RAR's.
3. **Use log-transformed RAR's.** The log(RAR) values are averaged together and the resulting average is then exponentiated in order to be on the same scale as the original RAR's. This method is equivalent to calculating a geometric average of the RAR's.

Choice of Weights.

As described in the manuscript, the final composite score was calculated as an equally weighted average of case-mix adjusted mortality and case-mix adjusted morbidity. Mathematically, the form of the composite measure calculation was: composite score = (mortality + morbidity) / 2 where mortality = RAR_{MORT} and morbidity = $(\text{RAR}_{\text{COMP}} + \text{RAR}_{\text{LOS}})/2$. A mathematically equivalent representation is to say that RAR_{MORT}, RAR_{COMP}, and RAR_{LOS} are weighted in a 2:1:1 ratio. The set of weights that we considered for alternative composite measure formulas was as follows:

2:1:1 weighting of mortality, major complications, and LOS

1:1:1 weighting of mortality, major complications, and LOS

4:2:1 weighting of mortality, major complications, and LOS

The first two sets of weights were tested in combination with all 3 standardization methods, whereas the 4:2:1 weighting was only applied to unstandardized RAR's. Thus, a total of 7 composite formulas were tested.

Results

As shown in the table, across all methods reliability and the proportion of hospitals classified as statistical performance outliers were generally similar (although these increased somewhat with increasing weight placed on LOS). Therefore, the investigator team primarily considered clinical face-validity (see manuscript for further details) and ease of interpretation in selection of the final weighting scheme, and the alternate methods were rejected as they resulted in greater influence of complications and LOS vs. mortality on the overall composite score which was less desirable.

	Final Comp. Method	Alternate Method #1	Alternate Method #2	Alternate Method #3	Alternate Method #4	Alternate Method #5	Alternate Method #6
Composite Method							
Standardization method	RAR	RAR/sd	logRAR	RAR	RAR/sd	logRAR	RAR
Weighting*	2:1:1	2:1:1	2:1:1	1:1:1	1:1:1	1:1:1	4:2:1
Correlation of individual components with overall composite score							
Mortality	0.87	0.81	0.87	0.74	0.74	0.73	0.87
Major Complications	0.70	0.63	0.70	0.82	0.66	0.82	0.72
LOS	0.47	0.67	0.45	0.49	0.71	0.46	0.39
Number of hospitals classified in composite performance categories							
Worse-than-expected	9	14	8	13	16	13	8
Same-as-expected	75	64	74	66	60	66	78
Better-than-expected	16	22	18	21	24	21	14
Reliability	0.73	0.80	0.75	0.81	0.84	0.82	0.72

*Relative weighting of mortality, major complications, and LOS, respectively.