Published in final edited form as:

Ann Thorac Surg. 2011 December ; 92(6): 2184–2192. doi:10.1016/j.athoracsur.2011.06.008.

Richard E. Clark Paper: Variation in Outcomes for Benchmark Operations: An analysis of the STS Congenital Heart Surgery Database

Jeffrey Phillip Jacobs, MD¹, Sean M. O'Brien, PhD², Sara K. Pasquali, MD², Marshall Lewis Jacobs, MD³, Francois G. Lacour-Gayet, MD⁴, Christo I. Tchervenkov, MD⁵, Erle H. Austin III, MD⁶, Christian Pizarro, MD⁷, Kamal K. Pourmoghadam, MD⁸, Frank G. Scholl, MD⁹, Karl F. Welke, MD¹⁰, and Constantine Mavroudis, MD³

¹The Congenital Heart Institute of Florida, Saint Petersburg, FL

²Duke University School of Medicine and Duke Clinical Research Institute, Duke University Medical Center, Durham, NC

³Cleveland Clinic Lerner School of Medicine, Cleveland, OH

⁴Children's Hospital at Montefiore, New York, NY

⁵Montreal Children's Hospital, Montreal, Canada

⁶University of Louisville, Louisville, KY

⁷Alfred I. duPont Hospital for Children, Wilmington, DE

⁸University of Oklahoma, OK

⁹Joe DiMaggio Children's Hospital, Hollywood, FL

¹⁰Mary Bridge Children's Hospital, Tacoma, WA

Abstract

Background—We evaluated outcomes for common operations in the STS Congenital Heart Surgery Database (STS-CHSDB) to provide contemporary benchmarks and examine variation between centers.

Methods—Patients undergoing surgery from 2005–2009 were included. Centers with >10% missing data were excluded. Discharge mortality and postoperative length of stay (PLOS) among patients discharged alive were calculated for eight benchmark operations of varying complexity. Power for analyzing between-center variation in outcome was determined for each operation. Variation was evaluated using funnel plots and Bayesian hierarchical modeling.

Results—18,375 index operations at 74 centers were included in the analysis of eight benchmark operations. Overall discharge mortality (range) was: ventricular septal defect repair (VSD) 0.6% (0%–5.1%), tetralogy of Fallot repair (TOF) 1.1% (0%–16.7%), complete atrioventricular canal repair (AVC) 2.2% (0%–20%), arterial switch (ASO) 2.9% (0%–50%), ASO+VSD 7.0% (0%–100%), Fontan 1.3% (0%–9.1%), truncus repair 10.9% (0%–100%), Norwood 19.3% (2.9%–100%). Funnel plots revealed the number of centers characterized as outliers were: VSD=0, TOF=0, AVC=1, ASO=3, ASO+VSD=1, Fontan=0, Truncus=4, Norwood=11. Power calculations showed statistically meaningful comparisons of mortality rates between centers could only be made for Norwood, for which the Bayesian-estimated range (95% Probability Interval) was 7.0%

(3.7%-10.3%) to 41.6% (30.6%-57.2%). Between-center variation in PLOS was analyzed for all operations and was larger for more complex operations.

Conclusions—This analysis documents contemporary benchmarks for common pediatric cardiac surgical operations and the range of outcomes among centers. Variation was most prominent for the more complex operations. These data may aid in quality assessment and quality improvement initiatives.

Keywords

database; outcomes

Background

The Congenital Heart Surgery Database of the Society of Thoracic Surgeons (STS–CHSDB) is the largest database in North America that tracks the outcomes of pediatric and congenital cardiac surgery [1,2,3]. As of January 1, 2011, participants in the STS–CHSDB include 96 of the estimated 122 congenital cardiac surgical programs in the United States [4]. One of the major goals of the STS–CHDB is to facilitate the improvement of quality in pediatric cardiac surgical programs in North America.

The purpose of this analysis is to document current outcomes for common operations in the STS–CHSDB to provide contemporary benchmarks and examine variation in outcomes between centers. In this manuscript, the terms “centers” and “participants” are used as synonyms to denote pediatric and congenital cardiac surgical programs that participate in STS–CHSDB. The approach of using benchmark operations to assess the quality of care of pediatric cardiac surgical operations has been previously described [5]. The goal of the analysis was to describe discharge mortality and postoperative length of stay (PLOS) for eight common potential benchmark operations of varying complexity and to examine between-participant variation in these endpoints. A related goal was to assess the feasibility of comparing institutions with these endpoints.

Material and Methods

Study Population

The study population includes patients who underwent operations with one of the Primary Procedures listed in Table 1 and met the inclusionary and exclusionary criteria listed in Table 1. Patients undergoing ASO are in a separate cohort from those undergoing ASO +VSD because the outcomes of these two groups are quite different [6]. Furthermore, the presence or absence of a VSD is a nonmodifiable variable that is an intrinsic characteristic of the patient. In the Fontan cohort, patients undergoing “Fontan revision or conversion (Re-do Fontan)” were excluded. Patients ≥ 7 years undergoing Fontan were excluded because it was felt to be less likely the patient was undergoing primary Fontan operation.

Analytic Methods

1. Outcome variables—Outcome variables in this analysis are mortality prior to discharge from the hospital (“discharge mortality”) and PLOS among patients discharged alive. In this manuscript, the word “mortality” is used to represent “discharge mortality” [7,8]. Previous publications from the STS–CHSDB have used PLOS as one measure of operative morbidity [7,8,9]. In these prior analyses, prolonged PLOS was regarded as a very general proxy measure of morbidity [9].

2. Raw data summary—For each type of procedure, the overall and participant-specific discharge mortality rates and the overall and participant-specific average PLOS were calculated. Participants-specific results were summarized by the mean, median (50th percentile), interquartile range (25th and 75th percentiles), and range (minimum and maximum).

3. Funnel plots—Participant-specific unadjusted mortality rates were depicted graphically in relation to the participant's number of eligible cases (i.e. the participant's sample size). Lines depicting exact 95% binomial prediction limits were overlaid to make a “funnel plot” [10]. For each individual participant, the probability of observing a mortality rate that falls on or outside of the plotted prediction limits is <5%, if the participant's true mortality rate is equal to the overall aggregate mortality rate of all STS participants in the analysis.

4. Feasibility of analyzing between-center variation—The feasibility of analyzing between-participant variation in mortality was assessed by counting the number of participants that met the sample size required to achieve 50% power to detect a two-fold increase in the mortality rate [11] (vs. the overall aggregate mortality rate of all participants) using a one-sided type-I error rate of 0.05. For example, assuming an overall aggregate mortality rate of 7%, a sample size of 48 operations would be required to attain 50% power to detect a doubling of the mortality rate to 14%.

The feasibility of analyzing between-participant variation in PLOS was assessed by counting the number of participants that met the sample size required to achieve 50% power to detect a doubling of the mean PLOS with a one-sided 0.05-level test. For simplicity, power was calculated by assuming an exponential distribution for time to hospital discharge. (This assumption was only made for sample size calculations, not for the actual data analysis.)

5. Bayesian estimation of between-participant variation—Bayesian hierarchical modeling was used to estimate the distribution of true unadjusted and adjusted participant-specific mortality rates and average PLOS. For unadjusted mortality, the observed number of deaths was modeled as a binomial distribution with different probability parameters (log-odds) for each participant. The log-odds parameters were assumed to be normally distributed across participants. For unadjusted PLOS, the patient-level variable $y = \log(1 + \text{PLOS})$ was modeled as a normal distribution with a different mean parameter for each participant and a single variance parameter that was common to all participants. Similar to the mortality model, mean parameters were assumed to be normally distributed across participants. For analyzing risk-adjusted outcomes, a hierarchical logistic regression model was used for mortality, and a hierarchical linear regression model was used for the variable $y = \log(1 + \text{PLOS})$. Covariates in each model included age (linear and quadratic), weight-for-age-and-sex z-score, sex (male vs. female/other/missing), any preoperative risk factor (yes/no), and any noncardiac abnormality (yes/no). The STS–CHSDB contains standard definitions adopted in 2007 for pre-operative risk factors and noncardiac abnormalities [12]. In addition, each model included normally distributed participant-specific random intercepts. The Bayesian approach to data analysis requires the analyst to specify prior beliefs about unknown model parameters using a probability distribution. Because our prior knowledge was limited, we specified a vague proper prior distribution that consisted of independent normal distributions for regression coefficients and inverse gamma distributions for variances. Inferences were based on Markov Chain Monte Carlo (MCMC) simulations as implemented in WinBUGS version 1.4 software. Bayesian point estimates (posterior means) and 95% probability intervals (PIs) were calculated using 420,000 MCMC iterations following a burn-in period of 5,000 iterations. To facilitate interpretation, parameters from the mortality models were converted to probabilities and parameters of the PLOS models were converted from the scale of $\log(1 + \text{PLOS})$ to the scale of untransformed PLOS. The

risk-adjusted mortality rate was defined as the mortality rate that would be predicted for a patient with risk factor values that are equal to the STS population average. The risk-adjusted mean PLOS was defined similarly.

The ratio of the maximum and minimum value was estimated for each endpoint to illustrate the scale of between-center differences. Also, the Gini index (GINI) was calculated for each operation as a measure of spread. GINI ranges from 0 to 1. A larger number means more variation between hospitals. GINI is one half of the average absolute difference of the mortality rates of two hospitals, averaging over all possible pairs of hospitals in the analysis, divided by the average mortality rate. We did not provide p-values because p-values are not used in Bayesian analyses. Instead, 95% Bayesian PIs are provided.

All analyses were performed using SAS version 9.2, R version 2.8, and WinBUGS version 1.4.

Institutional Review Board Approval

This study was approved by the Duke University Health System Institutional Review Board. Because the data used in analysis represent a limited data set (no direct patient identifiers) that was originally collected for non-research purposes, and the investigators do not know the identity of individual patients, the analysis of these data was declared by the Duke University Health System Institutional Review Board to be research not involving human subjects [13].

Results

From 2005–2009, inclusive, 85 centers (USA and Canada) submitted data to STS–CHSDB, and discharge mortality of index cardiac operations was 4.0% (3,418/86,297). For patients age<18 years, from 2005–2009, inclusive, 85 centers submitted data to STS–CHSDB, and discharge mortality of index cardiac operations was 4.1% (3,309/81,062). 18,375 index operations at 74 centers were included in the analysis of eight benchmark operations.

Raw Data and Funnel Plots

Table 2 summarizes overall aggregate and participant-specific results for mortality and PLOS for each operation. Mortality data are also displayed as funnel plots for these eight benchmark operations (Figure 1). These funnel plots demonstrate that for the majority of these benchmark operations, very few programs can be classified as outliers for discharge mortality, i.e., most programs fall within the 95% prediction limits and are not considered outliers. In fact, for some operations such as VSD, TOF, and Fontan, no programs are outliers. For other operations such as AVC, ASO, ASO+VSD, truncus, and Norwood, some participants are outliers. The number of “outliers” (based on two one-sided .025-level tests) were: VSD=0, TOF=0, AVC=1, ASO=3, ASO+VSD=1, Fontan=0, Truncus=4, Norwood=11. By design, approximately 5% of participants would be expected to have mortality rates that fall outside of the 95% prediction interval even if true probability of mortality did not vary across centers. For each operation except Norwood, the number of centers falling outside of the 95% prediction interval was consistent with the number that would be expected under the null hypothesis of no between-center variation. However, the small number of outliers should not be interpreted as evidence of no between-center variation in mortality. Power for detecting between-center variation for low complexity operations was minimal, as described below.

Feasibility of analyzing between-center variation

The number of cases required to detect a two-fold increase in the mortality rate with at least 50% power ranged from 17 for Norwood to 599 for VSD repair (Table 3). In the Norwood group, 40 participants met this required sample size. (Power to detect a smaller 1.5-fold increase in Norwood mortality was at least 50% for 12 participants and at least 80% for 4 participants.) For procedures other than Norwood, at most 1 participant met the sample size required to detect a doubling of mortality with at least 50% power. Based on these results, between-participant variation in mortality was analyzed with Bayesian methodology only for Norwood. For Bayesian analyses of Norwood, all participants were included regardless of sample size.

The required sample size to detect a doubling of the mean PLOS is five operations (Table 3). Based on these results, between-participant variation in PLOS was analyzed for all operations. All participants were included regardless of sample size.

Bayesian estimation of between-participant variation

Table 4 documents unadjusted and risk adjusted Bayesian estimation of between-participant variation for mortality and PLOS. The estimated 25th and 75th percentiles for Norwood mortality are 15.5% and 27.0%. We estimate that 25% of participants have a true mortality rate <15.5% and 75% of participants have a true mortality rate <27.0%. The estimated minimum and maximum true mortality rates are 7.3% and 47.0%. We estimate that the highest mortality rate is approximately 7-fold higher than the lowest. The 95% PI for the max/min ratio is 3.7–13.9, implying that we are highly confident that there is at least a 3.7-fold difference and no more than a 13.9-fold difference between the highest and lowest participant-specific true mortality rate. The between-center variation in mortality was only marginally attenuated when adjusting for case mix (estimated max/min ratio=6.5; 95% PI: 3.3–13.0). Variation in PLOS was also substantial, with a trend suggesting greater variation for higher-complexity operations. The estimated GINI index for adjusted PLOS ranged from 0.069 (95% PI: 0.056-0.082) for TOF to 0.142 (95% PI: 0.117-0.171) for Norwood.

Comment

The STS–CHSDB is the largest Congenital Heart Surgery Database in North America. This analysis documents (1) contemporary benchmarks for common pediatric cardiac surgical operations of varying levels of complexity, and (2) the degree of variation in outcome between centers. Variation in outcome was most prominent for the more complex operations. These data can aid in quality assessment and quality improvement initiatives. Variation in outcomes across centers demonstrates opportunities for multi-institutional collaboration to improve quality.

Knowledge of the distribution (e.g. percentiles) of adverse event rates across hospitals can be used to prioritize improvement efforts and establish benchmarks. However, estimation of hospital percentiles is not straightforward because the number of patients per hospital is often quite small. Percentiles calculated directly from the observed event rates are misleading because hospitals with a very small number of subjects are likely to have extreme event rates (e.g. 0% or 100%), and these rates may not be representative of their true long-run performance. Although the raw data are skewed towards having an unrealistically large amount of spread, a statistical model can be used to recover the true underlying distribution of hospital-specific probabilities. The Bayesian hierarchical modeling approach used in this article is particularly well suited for this purpose as it is designed explicitly to model true variation between units while accounting for purely random variation. In addition to estimating percentiles and other measures of between-

hospital variation, the Bayesian approach also allows calculating an appropriate measure of uncertainty (95% probability intervals) for these estimates.

Because Norwood was the only operation (of the 8 benchmark operations in this analysis) with more than one participant performing the minimum number of operations to detect a doubling of mortality, between-participant variation in mortality was analyzed with Bayesian methodology only for Norwood. With the other seven operations, the sample size (number of events) was too small to produce a valid estimate of the magnitude of between-center variation. In a Bayesian analysis, beliefs about unknown quantities are expressed using a probability distribution. One specifies a prior distribution (i.e. what one believes before seeing any data) and then one calculates the posterior distribution (i.e. what one believes after seeing the data) by using Bayes theorem. In the other seven operations, with such a small sample size, the results of a Bayesian analysis would be largely driven by the prior distribution, rather than the data. Although sensitivity to the prior distribution is an issue specific to Bayesian analysis, no alternative method could produce a meaningful estimate of between center variation for these seven operations.

It is apparent that even with 5 years of data, many individual operations are not performed frequently enough at any given institution to detect a doubling of mortality. A previous analysis of the potential to use mortality after “marker operations” to assess pediatric cardiac surgical performance concluded that, “There were relatively small data sets for individual hospitals and surgeons, which made statistical evaluation difficult. For setting standards, data from more departments for a longer period will be required. Statistical methods alone cannot be used as a sole arbiter of what is considered acceptable performance.” [5] Nevertheless, the strategy of analyzing mortality using funnel plots can help to identify programs that are outliers with respect to mortality for specific operations. Since 2000, this strategy has been utilized by the United Kingdom Central Cardiac Audit Database and forms the basis of their public reporting initiative [2].

Our assessment of the feasibility of analyzing between-center variation in mortality and PLOS revealed that statistically meaningful comparisons of mortality between centers could only be made for Norwood, while variability in PLOS could be analyzed for every operation. PLOS has been used as a surrogate for morbidity [9] and can help assess variation in outcome between centers. These data create an opportunity for inter-institutional collaboration in optimizing structure and process with a goal of improving overall quality of care and outcome.

Complexity stratification using the five STS-EACTS Categories [6] allows the grouping of operations into similar strata of risk and therefore permits analysis of higher volumes of cases than using benchmark operations. An analysis of variation in outcomes of mortality and PLOS stratified by the five STS-EACTS Categories represents an important opportunity for future investigation. Such an analysis may create additional opportunities for inter-institutional sharing of structure and process in order to improve overall quality of care and outcome. Similarly, Bayesian estimation of between-participant variation based on the five STS-EACTS Categories represents an important area of future investigation.

Conclusions

This analysis documents (1) contemporary benchmarks for common pediatric cardiac surgical operations of varying levels of complexity, and (2) the range of outcomes among centers. Variation in outcome was most prominent for more complex operations. Even with the use of 5 years of data, because of the relatively small datasets for many operations at most centers, it is not possible to perform statistically meaningful comparisons between

centers of mortality after benchmark operations. Funnel plots of mortality after benchmark operations can help to identify outliers. Grouping of operations into strata of similar complexity may further facilitate inter-institutional comparisons. These data can aid in quality assessment and quality improvement initiatives. Variation in outcomes across centers demonstrates opportunities for multi-institutional collaboration to improve quality.

References

1. Jacobs, JP.; Jacobs, ML.; Mavroudis, C.; Lacour-Gayet, FG.; Tchervenkov, CI. Executive Summary: The Society of Thoracic Surgeons Congenital Heart Surgery Database - Twelfth Harvest – (January 1, 2006 – December 31, 2009). The Society of Thoracic Surgeons (STS) and Duke Clinical Research Institute (DCRI); Duke University Medical Center, Durham, North Carolina, United States: Spring 2010 Harvest
2. Jacobs ML, Jacobs JP, Franklin RCG, et al. Databases for assessing the outcomes of the treatment of patients with congenital and paediatric cardiac disease – the perspective of cardiac surgery. *Cardiol Young*. 2008; 18(suppl 2):101–115. [PubMed: 19063780]
3. Jacobs JP, Maruszewski B, Kurosawa H, et al. Congenital heart surgery databases around the world: do we need a global database? *Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu*. 2010; 13(1):3–19. [PubMed: 20307856]
4. Jacobs ML, Mavroudis C, Jacobs JP, et al. Report of the 2005 STS Congenital Heart Surgery Practice and Manpower Survey: A report from the STS Work Force on Congenital Heart Surgery. *Ann Thorac Surg*. 2006; 82:1152–1159. [PubMed: 16928571]
5. Stark JF, Gallivan S, Davis K, et al. Assessment of mortality rates for congenital heart defects and surgeons' performance. *Ann Thorac Surg*. Jul; 2001 72(1):169–74. [PubMed: 11465173]
6. O'Brien SM, Clarke DR, Jacobs JP, et al. An empirically based tool for analyzing mortality associated with congenital heart surgery. *J Thorac Cardiovasc Surg*. 2009; 138:1139–1153. [PubMed: 19837218]
7. Jacobs JP, Mavroudis C, Jacobs ML, et al. What is operative mortality? Defining death in a surgical registry database: a report of the STS Congenital Database Task Force and the Joint EACTS-STS Congenital Database Committee. *Ann Thorac Surg*. 2006; 81:1937–41. [PubMed: 16631716]
8. Jacobs JP, Jacobs ML, Mavroudis C, et al. What is operative morbidity? Defining complications in a surgical registry database: a report from the STS Congenital Database Task Force and the Joint EACTS-STS Congenital Database Committee. *Ann Thorac Surg*. 2007; 84:1416–1421.
9. O'Brien SM, Jacobs JP, Clarke DR, et al. Accuracy of the Aristotle Basic Complexity Score for Classifying the Mortality and Morbidity Potential of Congenital Heart Surgery Procedures. *Ann Thorac Surg*. 2007; 84:2027–2037. [PubMed: 18036930]
10. Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Stat Med*. Apr 30; 2005 24(8):1185–202. [PubMed: 15568194]
11. Dimick JB, Welch HG, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: the problem with small sample size. *JAMA*. Aug 18; 2004 292(7):847–51. [PubMed: 15315999]
12. [Accessed January 1, 2011]
<http://www.sts.org/sections/stsnationaldatabase/datamanagers/congenitalheartsurgerydb/datacollection/index.html>
13. Dokholyan RS, Muhlbaier LH, Falletta J, et al. Regulatory and ethical considerations for linking clinical and administrative databases. *Am Heart J*. 2009; 157:971–82. [PubMed: 19464406]

Figure 1.

Table 1
Benchmark Operation included in this analysis

Procedure type	Abbreviation	STS-CHSDB Primary Procedure codes
1. VSD Repair	VSD	110="VSD repair, Patch"
2. TOF Repair	TOF	350="TOF repair, No ventriculotomy"
		360="TOF repair, Ventriculotomy, Nontransanular patch"
		370="TOF repair, Ventriculotomy, Transanular patch"
3. Complete AV Canal Repair	AVC	170=AVC (AVSD) repair, Complete (CAVSD)
4. Arterial Switch	ASO	1110="Arterial switch operation (ASO)"
5. Arterial Switch + VSD repair	ASO+VSD	1120="Arterial switch operation (ASO) and VSD repair"
6. Fontan	Fontan	970="Fontan, TCPC, Lateral tunnel, Fenestrated"
		980="Fontan, TCPC, Lateral tunnel, Nonfenestrated"
		1000="Fontan, TCPC, External conduit, Fenestrated"
		1010="Fontan, TCPC, External conduit, Nonfenestrated"
7. Truncus Repair	Truncus	230="Truncus arteriosus repair"
8. Norwood	Norwood	870="Norwood procedure"

The study population includes patients who underwent operations with one of the Primary Procedures listed in Table 1 and met the following inclusionary and exclusionary criteria:

1. Age<18 years
2. Surgical dates: January 1, 2005 – December 31, 2009, inclusive
3. Operations type: "CPB" or "No CPB Cardiovascular" (CPB=cardiopulmonary bypass)
4. The operation was the Index Operation of the admission (i.e. the first operation of a given admission where operation type was "CPB" or "No CPB Cardiovascular")
5. STS participant had at least 90% complete data for discharge mortality, PLOS, pre-operative risk factors, and non-cardiac abnormalities
6. Fontan: exclude patients age>7 years
7. Truncus: exclude any operation that includes a secondary procedure of "240=Valvuloplasty, Truncal valve" or "250=Valve replacement, Truncal valve"
8. From included participants, patients with missing discharge mortality status or PLOS were excluded from analysis
9. For certain analyses, we restricted inclusion to participants with a certain number of cases, as discussed in the manuscript.

Table 2

Raw Data Summary

	VSD	TOF	AVC	ASO	ASO+VSD	Fontan	Truncus	Norwood
ALL PARTICIPANTS								
Number of participants	74	74	73	70	65	71	66	69
Number of operations	5150	3264	2325	1499	655	2691	449	2342
Average participant-specific sample size	69.7	44.2	32.0	21.4	10.1	37.9	6.8	34.4
Range of participant-specific sample sizes	2-288	2-199	1-171	1-131	1-56	1-267	1-37	1-205
Aggregate mortality rate	0.64%	1.1%	2.2%	2.9%	7.0%	1.3%	10.9%	19.3%
Median participant-specific mortality rate	0%	0%	0%	0%	1.8%	0%	0%	22.6%
Range of participant-specific mortality rates	0%-5.1%	0%-16.7%	0%-20%	0%-50%	0%-100%	0%-9.1%	0%-100%	0%-100%
Interquartile range of participant-specific mortality rates	0%-1.1%	0%-2.0%	0%-4.5%	0%-5.3%	0%-16.7%	0%-1.5%	0%-16.7%	14.3%-31.6%
Aggregate average PLOS per patient (days)	7.6	9.8	12.2	14.5	17.4	13.4	25.5	35.2
Median participant-specific average PLOS (days)	7.6	9.9	12.3	14.6	17.5	13.3	21.7	36.5
Range of participant-specific average PLOS (days)	4.1-25.4	5.5-34.4	4.8-30.5	6.9-42.0	7.0-110.5	7.0-35.5	7.0-52.8	13.7-109.0
Interquartile range of participant-specific average PLOS	6.2-8.6	8.2-11.8	8.9-14.6	12.0-17.4	13.3-23.3	11.6-17.1	16.3-29.5	30.1-45.3
SITES WITH N>=10								
Number of participants	72	69	57	43	27	56	17	48
Number of operations	5139	3240	2255	1351	502	2624	262	2250
Average participant-specific sample size	71.5	47.0	39.8	31.5	18.6	46.9	15.5	47.5
Range of participant-specific sample sizes	10-288	10-199	10-171	11-131	10-56	11-267	10-37	10-205
Aggregate mortality rate	0.64%	1.1%	2.2%	2.3%	5.2%	1.3%	9.2%	18.9%
Median participant-specific mortality rate	0%	0%	1.5%	0%	6.3%	0%	8.3%	22.7%
Range of participant-specific mortality rates	0%-5.1%	0%-10.0%	0%-15.4%	0%-12.5%	0%-36.4%	0%-9.1%	0%-40.0%	0%-50.0%
Interquartile range of participant-specific mortality rates	0%-1.1%	0%-2.0%	0%-5.3%	0%-4.8%	0%-10.0%	0%-2.8%	0%-10.3%	14.6%-30.4%
Aggregate average PLOS per patient (days)	7.6	9.8	12.2	14.5	17.4	13.4	25.5	35.2
Median participant-specific average PLOS (days)	7.6	9.9	12.3	14.7	17.0	13.3	26.5	37.4
Range of participant-specific average PLOS (days)	4.1-25.4	5.8-20.7	6.2-22.1	6.9-24.4	10.1-25.4	8.2-26.7	15.3-48.9	13.7-87.1
Interquartile range of participant-specific average PLOS	6.2-8.7	8.3-11.8	9.3-14.4	12.4-17.2	12.7-19.2	12.0-16.9	21.7-34.9	31.6-44.9

Table 3**Feasibility of analyzing between-center variation**

	VSD	TOF	AVC	ASO	ASO+VSD	Fontan	Truncu s	Norwood
Number of participants included in analysis	74	74	73	70	65	71	66	69
Aggregate mortality rate	0.64%	1.1%	2.2%	2.9%	7.0%	1.3%	10.9%	19.3%
Sample size required for 50% power to detect 2x increase in mortality	599	338	172	131	48	304	31	17
Number of participants meeting mortality requirement	0	0	0	1	1	0	1	40
Number of participants meeting PLOS requirement (≥ 5 cases)	73	72	65	60	44	64	33	59

Table 4

Results of Bayesian Hierarchical Models

Estimate (95% Bayesian Probability Interval)							Estimated Gini Index (×100)
Minimum	25th	50th	75th	Maximum	Max/Min		
Estimated percentiles of the distribution of hospital-specific mortality rates							
Norwood Unadjusted Mortality (%)	7.3 (3.8-10.8)	15.5 (13.2-18.0)	20.8 (18.2-23.6)	27.0 (23.5-31.2)	47.0 (35.5-61.8)	7.0 (3.7-13.9)	22.0 (16.3-27.6)
Norwood Risk-Adjusted Mortality (%)	7.0 (3.7-10.3)	14.0 (11.6-16.6)	18.6 (15.9-21.3)	23.9 (20.3-27.9)	41.6 (30.6-57.2)	65. (3.3-13.0)	21.1 (15.0-27.3)
Estimated percentiles of the distribution of hospital-specific average PLOS:							
Unadjusted							
VSD	5.8 (5.1-6.3)	7.2 (6.9-7.4)	7.9 (7.7-8.1)	8.7 (8.4-9.0)	11.2 (10.2-12.5)	1.9 (1.7-2.3)	8.0 (6.8-9.3)
TOF	7.5 (6.9-8.0)	9.4 (9.1-9.8)	10.4 (10.0-10.7)	11.3 (10.9-11.7)	14.0 (12.7-15.9)	1.9 (1.6-2.2)	7.5 (6.2-9.0)
AVC	8.5 (7.4-9.5)	11.2 (10.6-11.7)	12.4 (11.9-12.9)	13.7 (13.0-14.4)	17.9 (15.7-21.0)	2.1 (1.7-2.7)	8.6 (6.7-10.8)
ASO	9.2 (7.8-10.5)	13.6 (12.9-14.3)	15.4 (14.7-16.1)	17.4 (16.4-18.4)	24.8 (20.8-31.5)	2.7 (2.1-3.7)	10.9 (8.6-13.4)
ASO+VSD	11.9 (9.3-14.2)	15.9 (14.6-17.2)	18.0 (16.8-19.4)	20.5 (18.6-22.7)	30.2 (22.4-44.4)	2.6 (1.6-4.2)	11.1 (6.4-16.3)
Fontan	9.7 (8.8-10.4)	13.0 (12.4-13.6)	14.5 (13.9-15.1)	16.1 (15.4-16.9)	22.2 (19.2-26.9)	2.3 (1.9-2.8)	9.5 (7.8-11.5)
Truncus	18.4 (13.3-23.6)	23.0 (20.3-25.8)	25.3 (23.0-27.7)	27.9 (24.9-31.3)	35.8 (27.2-47.6)	2.0 (1.2-3.4)	8.2 (2.1-14.2)
Norwood	16.8 (13.9-19.9)	31.7 (29.6-33.8)	37.8 (35.5-40.4)	44.5 (41.4-47.9)	70.4 (57.3-92.5)	4.2 (3.1-5.9)	15.0 (12.5-17.9)
Estimated percentiles of the distribution of hospital-specific average PLOS:							
Risk Adjusted							
VSD	5.8 (5.2-6.3)	7.1 (6.9-7.3)	7.6 (7.4-7.8)	8.3 (8.0-8.6)	10.4 (9.7-11.4)	1.8 (1.6-2.1)	7.1 (6.1-8.1)
TOF	7.5 (6.9-8.0)	9.4 (9.1-9.7)	10.2 (9.9-10.5)	11.0 (10.7-11.4)	13.5 (12.2-15.4)	1.8 (1.6-2.1)	6.9 (5.6-8.2)
AVC	8.6 (7.4-9.5)	11.0 (10.5-11.6)	12.2 (11.7-12.7)	13.5 (12.9-14.2)	17.6 (15.6-20.6)	2.1 (1.7-2.6)	8.6 (6.7-10.8)

Estimate (95% Bayesian Probability Interval)						
	Minimum	25th	50th	75th	Maximum	Max/Min
ASO	9.3 (7.9-10.6)	13.6 (12.9-14.2)	15.1 (14.5-15.8)	17.0 (16.2-18.0)	24.3 (20.1-31.2)	2.6 (2.0-3.6)
ASO+VSD	12.5 (10.1-14.7)	15.8 (14.6-17.1)	17.5 (16.4-18.7)	19.4 (17.9-21.3)	25.7 (20.5-35.4)	2.1 (1.4-3.2)
Fontan	10.0 (8.8-10.8)	13.1 (12.5-13.7)	14.5 (14.0-15.1)	16.2 (15.5-16.9)	22.4 (19.3-27.5)	2.2 (1.9-2.9)
Truncus	18.4 (13.4-23.5)	22.6 (20.1-25.2)	24.7 (22.7-26.9)	27.0 (24.3-30.2)	33.9 (26.0-45.0)	1.9 (1.2-3.2)
Norwood	17.5 (14.0-21.3)	31.4 (29.5-33.5)	37.1 (34.9-39.4)	43.7 (40.7-47.1)	67.5 (55.9-86.3)	3.9 (2.9-5.5)
						Estimated Gini Index ($\times 100$)
						10.3 (8.1-12.8)
						8.7 (4.6-13.4)
						9.3 (7.5-11.3)
						7.4 (1.7-13.5)
						14.2 (11.7-17.1)