Original Article

World Journal for Pediatric and Congenital Heart Surgery
2022, Vol. 13(1) 23-35

DOI: 10.1177/21501351211051227
journals.sagepub.com/home/pch
SAGE

Benchmarking in Congenital Heart Surgery Using Machine Learning-Derived Optimal Classification Trees

Dimitris Bertsimas, PhD[1], Daisy Zhuo, PhD[2,3], Jordan Levine, MEng[2,3], Jack Dunn, PhD[2,3], Zdzislaw Tobota, MD[4], Bohdan Maruszewski, MD, PhD[4], Jose Fragata, MD, PhD[5], and George E Sarris, MD, PhD[6]

Abstract

Background: We have previously shown that the machine learning methodology of optimal classification trees (OCTs) can accurately predict risk after congenital heart surgery (CHS). We have now applied this methodology to define benchmarking standards after CHS, permitting case-adjusted hospital-specific performance evaluation. **Methods:** The European Congenital Heart Surgeons Association Congenital Database data subset (31 792 patients) who had undergone any of the 10 "benchmark procedure group" primary procedures were analyzed. OCT models were built predicting hospital mortality (HM), and prolonged postoperative mechanical ventilatory support time (MVST) or length of hospital stay (LOS), thereby establishing case-adjusted benchmarking standards reflecting the overall performance of all participating hospitals, designated as the "virtual hospital." These models were then used to predict individual hospitals' expected outcomes (both aggregate and, importantly, for risk-matched patient cohorts) for their own specific cases and case-mix, based on OCT analysis of aggregate data from the "virtual hospital." **Results:** The raw average rates were HM = 4.4%, MVST = 15.3%, and LOS = 15.5%. Of 64 participating centers, in comparison with each hospital's specific case-adjusted benchmark, 17.0% statistically (under 90% confidence intervals) overperformed and 26.4% underperformed with respect to the predicted outcomes for their own specific cases and case-mix. For MVST and LOS, overperformers were 34.0% and 26.4%, and underperformers were 28.3% and 43.4%, respectively. OCT analyses reveal hospital-specific patient cohorts of either overperformance or underperformance. **Conclusions:** OCT benchmarking analysis can assess hospital-specific case-adjusted performance after CHS, both overall and patient cohort-specific, serving as a tool for hospital self-assessment and quality improvement.

Keywords

Congenital heart disease, congenital heart surgery, database (all types), outcomes, statistics, risk analysis/modeling, statistics-survival analysis

Submitted April 28, 2021; Accepted September 5, 2021

Introduction

Quality improvement efforts in congenital heart surgery (CHS) depend on the determination of appropriate benchmarking standards and on comparison of measured outcomes to the benchmark, while adjusting for variation in case-mix and in various patient and institutional factors that may affect such risk. Adjustment for variability in outcome attributed to the inherent risk of different procedures and different patient characteristics has evolved from initial attempts of risk stratification based on expert consensus (Risk Adjustment for Congenital Heart Surgery [RACHS-1][1] and Aristotle[2] methods) to those based on outcome measures provided by real data (STS-EACTS [STAT] Mortality Score, recently updated as STAT 2020, and Society of Thoracic Surgeons [STS] Morbidity Score.[3–15]) Risk prediction models have been developed and refined (STS Congenital Heart Surgery Database [CHSD] Mortality Risk Model[7–11] and the UK PRAiS2 models[15]), adjusting for differences in both case-mix (risk stratification) and patient-specific factors, achieving remarkable accuracy (area under the curve [AUC] in the range of 0.852-0.875).[7,8,11] These approaches have received criticism for both data limitations (eg, the availability of only a limited subset of many potentially important patient-related factors) and for methodological issues, such as emphasis on a single summary measure of hospital performance, the overall observed/expected (O/E) mortality ratio, which may either mask underperformance of

[1] Operations Research Center and Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA
[2] Alexandria Health, Cambridge, MA, USA
[3] Alexandria Health, Providence, RI, USA
[4] Children's Memorial Health Institute, Warsaw, Poland
[5] Hospital de Santa Marta and NOVA University, Lisbon, Portugal
[6] Athens Heart Surgery Institute, Athens, Greece

Corresponding Author:
George E. Sarris, Athens Heart Surgery Institute, Kifissias Avenue #2, Amaroussion, Athens 151 25, Greece.
Email: gsarris@mac.com

Abbreviations

AI	artificial intelligence
AUC	area under the curve (or c-statistic)
CHS	congenital heart surgery
CHSD	CHS database
CPB	cardiopulmonary bypass
ECHSA	European Congenital Heart Surgeons Association
ECDB	ECHSA congenital database
LOS	length of hospital stay
ML	machine learning
MVST	mechanical ventilatory support time
OCT	optimal classification trees
VSD	ventricular septal defect

centers for some low-volume complex procedures or may do injustice when applied to centers performing rare or even unique and high-risk procedures.[14] This latter issue has been partially addressed by comparative performance analyses focusing on "benchmark procedures" as well as on additional metrics including adjusted mortality rates, both overall and for each STAT Mortality Category.[4,10]

An important limitation of traditional analytical methods (linear regression) is the incorrect assumption that various possible risk factors interact in a linear and additive fashion, ie, that the odds ratio for each risk factor is the same for all patients and does not interact with other factors. This limitation has been partially addressed in the STS CHSD Mortality Risk Model in which intuitively preselected feature interactions have been considered.[11]

We have previously shown[16] that the artificial intelligence (AI)–machine learning (ML)-based nonlinear methodology of optimal classification trees (OCTs) can be used to accurately and interpretably predict risk after CHS even at the individual patient level, taking into account all relevant recorded risk factors, automatically identifying important ones without any a priori assumption of factor interaction, thereby avoiding human bias introduction. We now apply this methodology to demonstrate how benchmarking standards after CHS can be calculated objectively, with case-mix adjustment customized to individual hospitals, permitting practical self-evaluation of hospital performance, both overall and also regarding cohort-specific outcomes. We emphasize that we aimed to define the methodology to be used for center self-assessment using European data, and not to establish generally applicable benchmarking standards or for the purpose of public reporting. Furthermore, in recognition of the fact that there is a wide spectrum of CHS procedures that are performed quite infrequently even in large centers, we limited this analysis to the 10 so-called "benchmark procedure groups," a practice also adopted by the STS CHSD.[4,10]

Material and Methods

The data, provided by the European Congenital Heart Surgeons Association (ECHSA) congenital database (ECDB) after study review and approval by the ECDB Committee regarding compliance with all ECHSA ethical and patient data protection policies, validated in accordance with ECDB procedures, encompass fully anonymized information regarding patients who have undergone CHS in participating hospitals. The ECDB is fully compliant with the European General Data Protection Regulation, and all its participating hospitals have agreed to provide fully anonymized data to the ECDB to be used in data analyses for research and patient care quality improvement initiatives, in full compliance with every applicable local law and internal Institutional Review Board Procedure. This study focuses on a subset of the total data of >235 000 patients and 295 000 operations, pertaining to 64 202 patients who had undergone operations belonging to the 10 "benchmark procedure groups" (see Table 1 and Supplemental Table S1) from January 1, 2010, to December 31, 2018. Non-European hospitals and their data have been specifically excluded from this analysis. After limiting to European centers only, the outcomes of 31 792 "benchmark" operations were analyzed. The 2010-2015 data subset was used to train the models and those from 2016 to 2019 for testing.

Methods

The collective data of all 64 ECDB participating hospitals are considered to define the "virtual hospital," ie, as if the total patient population and all surgical outcomes reflected the practice of CHS at a single theoretical hospital containing all participating hospitals. The OCT ML-based method previously described[16] was applied. OCTs, compared to traditional methods such as logistic regression, can capture nonlinear variable interactions while providing individual-level interpretability. The output is a single decision tree where the user can trace the decision path that leads to the prediction, hence providing the level of interpretability that other black box methods such as random forests and gradient boosting are unable to achieve. The modern optimization techniques used in OCTs also allow them to perform competitively with the black-box methods.

The risk factors entered in the model, all being preoperative features, are shown in Table 2. The risk of postoperative adverse outcomes previously defined[16] (hospital mortality [HM], prolonged postoperative mechanical ventilatory support time [MVST], and prolonged postoperative length of hospital stay [LOS]) was calculated. The model for each outcome presents itself in the form of a decision tree, with predictive power (out-of-sample AUC) 0.871, 0.814, and 0.813, respectively. Other performance characteristics of the models are shown in Supplemental Table S2 and Figures S1 and S2. The relevant preoperative risk factors for each model are shown as split variables in the respective virtual hospital decision tree, of which, for simplicity, only a portion for the mortality model is shown in Figure 1.

In addition, the OCT algorithm automatically determines "patient pathways" along the virtual hospital decision tree, each of which (a) describes a patient cohort with a combination of particular characteristics having a similar risk profile and (b) presents outcome statistics for the particular patient cohort, averaged for all hospitals.

Table 1. The 10 Benchmark Procedures Studied.

Procedure	Training (2010-2015)		Testing (2016-2019)	
	Number of procedures	Mortality (%)	Number of procedures	Mortality (%)
Off-bypass coarctation repair	3519	2.2	1248	1.0
Fontan procedure	1264	4.7	420	2.9
Glenn or hemi-Fontan procedure	832	7.5	423	4.0
Arterial switch operation with ventricular septal defect (VSD) repair	895	8.2	312	7.4
Arterial switch operation	2175	4.0	757	5.4
Complete atrioventricular canal repair	2157	4.7	821	3.8
Tetralogy of Fallot repair	728	1.5	249	1.2
VSD repair	7257	0.9	2503	0.6
Norwood procedure	1288	31.4	564	28.4
Truncus repair	3265	4.0	1115	2.4
Overall	23 380	4.6	8412	4.1

Results

Assessment of the Virtual Hospital

The full virtual hospital decision tree has many splits, defining 18 terminal patient cohorts. For simplicity, we show in detail (Figure 1) only part (first 2 levels) of the entire virtual hospital tree for the mortality outcome. The virtual hospital average HM (for 23 380 patients in the training set) is 4.6%. The tree part shown magnified consists of 4 pathways leading to 4 distinct cohorts:

Cohort 1 (mortality risk: 7.5%): This cohort underwent atrioventricular septal defect repair, arterial switch operation, arterial switch operation/ventricular septal defect (VSD) repair, off-bypass coarctation repair, tetralogy of Fallot repair, and VSD repair, and it has been <1569 days since the previous operation.

Cohort 2 (mortality risk: 2.0%): This cohort underwent the same set of procedures as cohort 1, but has either never had a previous operation, or it has been at least 1569 days since the previous one.

Cohort 3 (mortality risk: 31.4%): This cohort underwent the Norwood procedure.

Cohort 4 (mortality risk: 9.9%): This patient cohort underwent a procedure belonging to 1 of the following 3 procedure groups: Fontan, Glenn Hemi-Fontan, and Truncus Repair.

Table 2. Preoperative Risk Features (Potential Risk Factors) Analyzed and the Features' Relative Importance for Mortality Prediction.

Predictive variable	Importance (%)
Procedure	71.0
Days since the previous operation, if any	11.9
Weight	9.2
Age (months)	3.7
Number of preoperative diagnoses	3.1
Any general preoperative risk factor present	1.0
Case category (cardiopulmonary bypass [CPB] vs non-CPB)	0.0
Gender	0.0
Number of concomitant procedures performed	0.0
Any noncardiac abnormality present	0.0
Any prior operation	0.0
Year of procedure	0.0

In this nonlinear methodology, all patients of a cohort share similar risk by virtue of various features, but may well have undergone different procedures. The same procedure may appear in different risk cohorts, depending on other features, but a given patient belongs only to 1 same-level cohort.

The collection of terminal nodes at the end of each pathway forms a set of patient cohorts, each defined by specific features and including patients of similar risk. The 18 terminal leaves of the entire tree define all of the similar risk cohorts of the virtual hospital.

The entire collection of cohorts makes up the whole patient population, and any patient analyzed by the virtual hospital OCT will be categorized into 1 and only 1 terminal cohort. Furthermore, a comparison of the percentage of patients in each cohort to the overall population provides an appreciation for patient case-mix as illustrated in Figure 2, which sorts all cohorts from low to high risk. Some cohorts have a high risk (cohort 3) and some have a low risk (cohort 2). It is evident that some nodes have a high percentage of patients (cohort 2); some do not (cohorts 1 and 3). Bar height represents the percentage of patients in each cohort. Thus, the pathways enable a logical presentation of the virtual hospital case-mix.

Assessment of the Individual Hospital

We assess each individual hospital's ("index hospital") performance compared to the virtual hospital by calculating the potential outcome of each individual hospital's patient population in

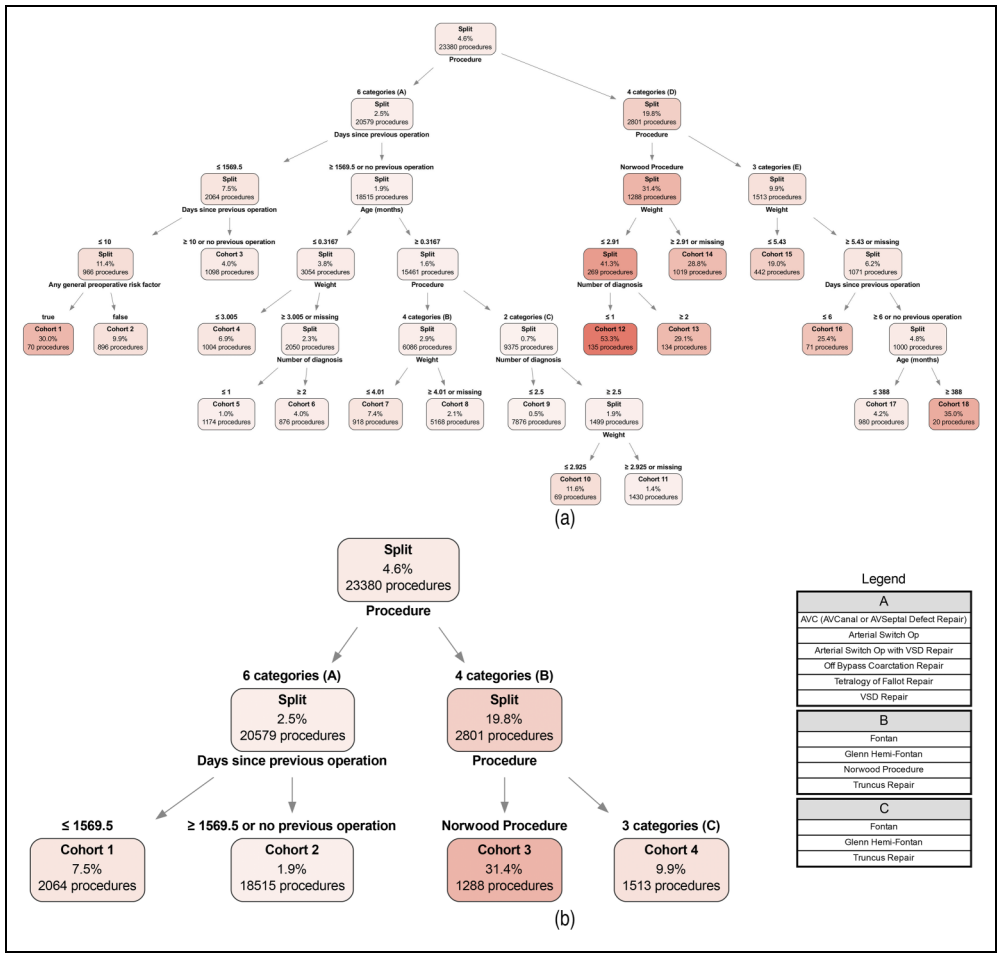

Figure 1. The mortality model optimal classification tree (OCT) for the virtual hospital (A), with a zoomed-in version of the top 2 levels shown (B). Each of the terminal boxes refers to a cohort defined by the criteria for its pathway. For each cohort, the average mortality and the number of procedures (in the training data subset) for that cohort are shown. At the top level, there are 23 380 cases of benchmark procedures analyzed. The first split divides the total sample into 2, based on the variable procedure: on the left branch, there are 20 579 procedures falling under 6 categories (details in the legend) with mortality 2.5% and on the right branch 2801 procedures in 4 categories, with mortality 19.8%. The left branch is further split based on the variable days since the previous operation if any, leading to cohorts 1 and 2, and the right branch is further split again according to the variable procedure, leading to cohorts 3 and 4. Note that the splits along the pathways are chosen by the algorithm to be the optimal ones and are based on different variables every time, as the methodology is not linear. The table insert shows, as an example, which procedures are used to split into cohorts 3 and 4. The terminal leaves of the entire tree (not shown in detail for simplicity) define all the similar risk cohorts revealed by the algorithm.

the virtual hospital's performance environment. This is a 2-step process: First, we calculate the aggregate outcome rate for mortality, prolonged MVST, and prolonged LOS of the index hospital's case-mix in the virtual hospital's performance environment (decision tree), yielding what we define as the expected rate, both cumulatively, for all patients, and, second, also for each patient cohort. In other words, the expected rates provide measures of performance assuming the index hospital's actual case-mix is being treated at the virtual hospital's objectively known performance level. Thus, the calculated expected rates serve as each index hospital's case-adjusted and hospital-specific benchmarks.

Stated alternatively, since, clearly, the performance of an individual hospital can only be "quantified" in relation to their own unique case-mix, our methodology achieves such comparison of the performance of a hospital to their own predicted outcomes for their own specific cases and case-mix, based on our OCT analysis of aggregate data from the "virtual hospital."

The calculation of these metrics is illustrated with an example of hospital's case-mix in Figure 2. To adjust for the difference in case-mix, we calculate the expected rate by applying the virtual hospital performance on this hospital's case-mix. We multiply the mortality rates observed at the virtual hospital for each cohort by the number of patients that this index hospital sees for each corresponding cohort and divided by the total number of patients at this hospital, yielding the expected rate.

Next, to further analyze the index hospital's performance, we identify: first, "areas of distinction," ie, cohorts for which the observed index hospital's measured performance is statistically better than that of the same patient cohort's calculated risk

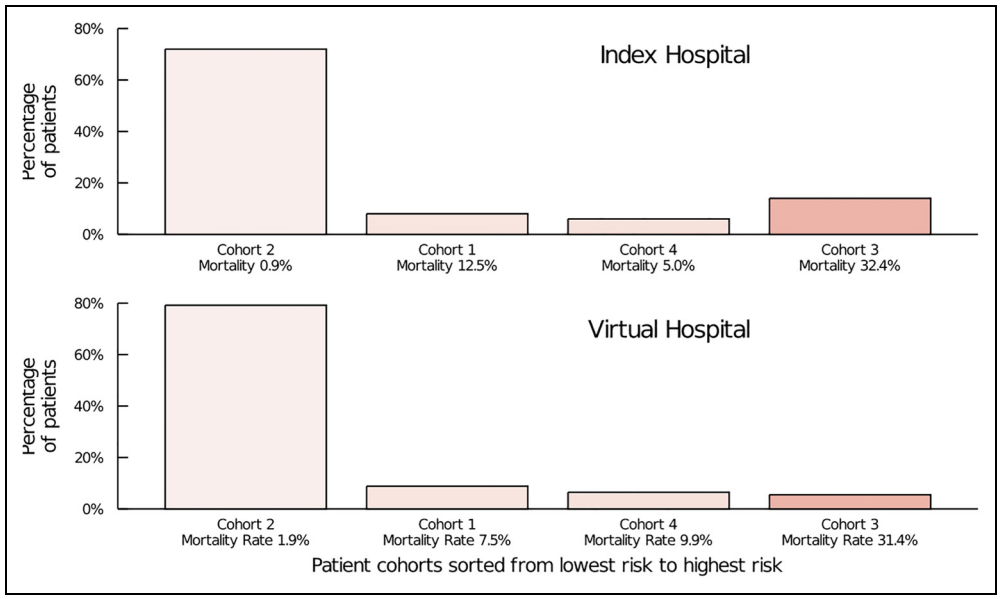

Figure 2. Example case-mix for the virtual hospital, and an example hospital for comparison. The cohorts appear on the *x*-axis (ordered from lower to highest risk). The *y*-axis represents the percentage of the total patient number in each cohort. Shading represents a risk, higher risk is indicated by darker shading. Comparison of the case-mix of the index hospital to that of the virtual hospital shows that the index hospital has relatively fewer low-risk patients (cohort 2) and more high-risk patients (cohort 3).

in the virtual hospital ("expected rate") and second, "areas of opportunity," ie, cohorts of greater observed risk in the index hospital compared to the "expected rate," ie, the calculated risk of a patient cohort with identical characteristics in the virtual hospital. This is accomplished by comparing the entire virtual hospital tree to the entire individual index hospital tree, permitting a direct comparison of identical terminal nodes of the pathways of the index and virtual hospitals, ie, direct comparison of outcomes of patient cohorts with similar risk characteristics.

Comparative Assessment of all Hospitals

The results of comparison of overall outcome performance of each of the 64 index hospitals against their respective expected rate are shown in Figures 3 to 5 for mortality, prolonged MVST, and prolonged LOS, respectively. In Figures 6 and 7, the observed-to-expected (O/E) mortality ratio of all hospitals is shown, sorted from the lowest to the highest. For mortality, 17.0% of hospitals perform better and 26.4% worse than expected, ie, with respect to the predicted outcomes for their own specific cases and case-mix, based on the OCT analysis of aggregate data from the "virtual hospital." For 11.3% of hospitals, the O/E mortality ratio is above 2. We observe a similar performance distribution for prolonged MVST and prolonged LOS.

The result of an individual hospital's detailed performance analysis is illustrated by the example of an index unnamed real hospital, labeled as "Hospital A," with 1254 patients included in this study. The summary assessment of this hospital is shown in Table 3, with the actual observed versus the expected values for HM, prolonged MVST, and prolonged LOS rates being 7.2% versus 7.2% ($p > .05$), 22.7% versus 19.7% ($p = .011$), and 25.7% versus 21.6% ($p = .002$). The expected rates are the predicted outcomes for this hospital's own specific cases and case-mix, based on the OCT analysis of aggregate data from the "virtual hospital."

Focusing on mortality, the observed HM for Hospital A is 7.2%, compared with average raw mortality in the virtual hospital of 4.4%. On surface, Hospital A is doing worse than average. However, as illustrated in Figure 8, comparison of the risk-stratified cohorts of Hospital A and the virtual hospital, it is evident Hospital A's case-mix is at higher risk. Accordingly, when we calculate Hospital A's expected rate, which is predicted based on its own specific case-mix, conceptually, as if this hospital's patients were to be treated at the virtual hospital, its HM should be 7.2%, indeed higher than the average raw mortality of the virtual hospital. Since Hospital A's observed mortality rate is 7.2%, similar (by chance, here, equal) to the adjusted, expected rate, Hospital A's performance is not statistically significantly different from the virtual hospital. Indeed, in Figure 6 and 7, where the O/E ratio of all hospitals is shown, O/E for Hospital A is 1.0, the error bar indicating the absence of statistically significant difference. Analyzing mortality performance at a deeper level, examination of the full Hospital A decision tree may reveal "areas of distinction" and "areas of opportunity," the color-coding indicating overperformance (green) or underperformance (red), compared to expected, the case adjusted benchmark. Despite the overall performance being as expected, this full tree (Figure 9), showing the pathways to the 8 distinct cohorts revealed by the algorithm, demonstrates no cohorts of distinction, and some (cohorts 3, 5, and 8) providing opportunities

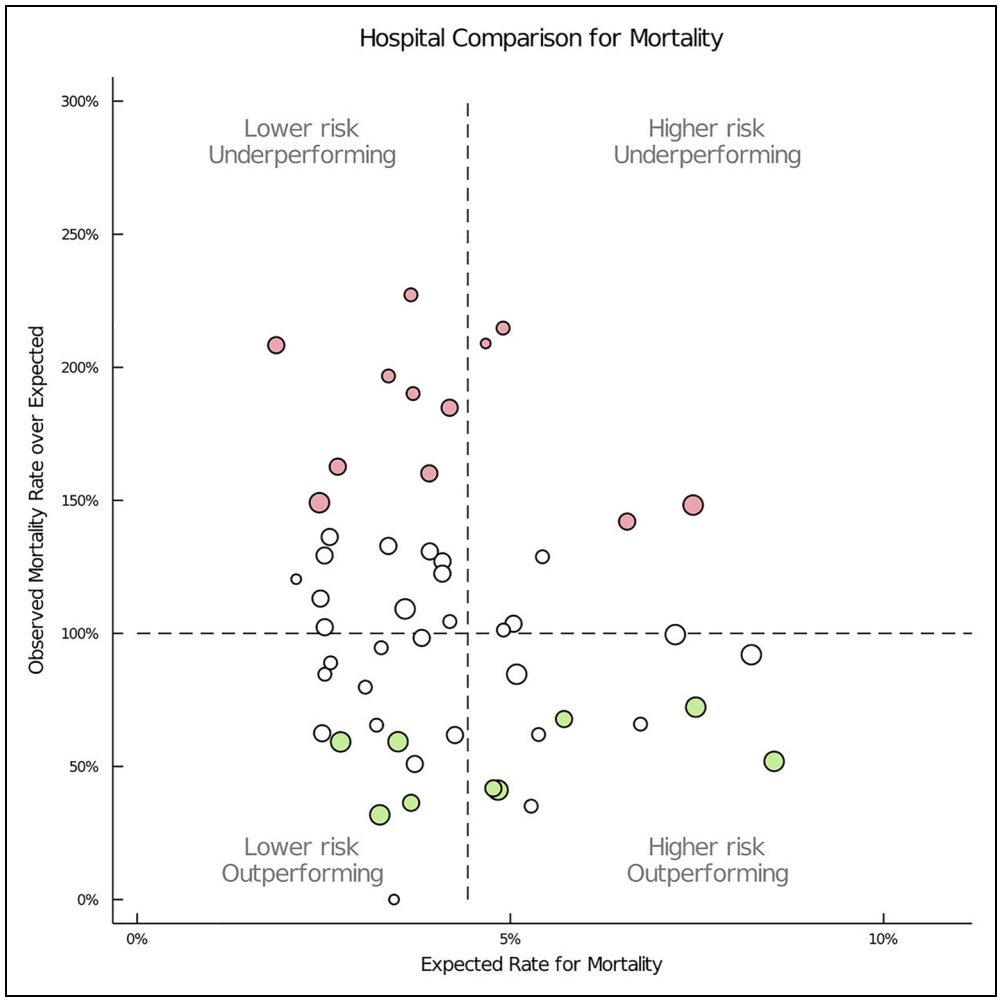

Figure 3. Comparison of mortality for all hospitals. The O/E mortality ratio (*y*-axis) is plotted against the expected mortality for each hospital. Dot size reflects hospital size (ie, number of patients). The vertical solid line represents the raw mortality of the virtual hospital. The horizontal line corresponds to HM equals to its expected mortality, which is calculated by our OCT analysis as the predicted mortality for the hospital's own specific cases and case-mix, based on the OCT analysis of aggregate data from the "virtual hospital." These 2 lines divide the *x*–*y*-plane into 4 quadrants: "lower-risk cases" and underperforming hospitals, "lower-risk cases" and overperforming hospitals, "higher risk cases" and underperforming hospitals, and "higher risk cases" and overperforming hospitals. Higher or lower risk is in comparison to the aggregate risk in the virtual hospital. Red indicates statistically significant underperformance, green indicates overperformance, and white indicates no statistically significant performance difference.
Abbreviations: O/E, observed-to-expected; HM, hospital mortality; OCT, optimal classification tree.

for improvement (worse than expected performance). Figure 10 shows the characteristics of these cohorts, and Figure 11 summarizes mortality performance for Hospital A, further breaking down cohorts by benchmark procedure.

Comment

We have previously shown[16] analyzing data for more than 235 000 patients and more than 295 000 operations in the ECDB that the nonlinear AI–ML methodology of OCTs permits accurate predictions of adverse outcomes after CHS in a fully intuitively and interpretable manner, presenting themselves as decision trees. The predictive power of this methodology, which some of us have previously used successfully in other medical applications,[17] is devoid of frequently assumptions of risk factor linearity which have been traditionally employed in linear regression-based methods and involves no assumptions about the potential importance of preoperative features, assumptions which may introduce bias. The benchmarking analysis presented herein utilizes the power of OCTs, focusing on the ECDB data subset pertaining to 10 common "benchmark" operation groups, to reduce the variability related to a wide spectrum of many more but much rarer procedures.[4,10] Furthermore, we limited our analysis to European data only, to limit potential variability related to the geographic heterogeneity of CHS practices.

Our risk model for each adverse outcome studied, presented as a decision tree, takes into account all preoperatively known variables recorded in the database, including patient-specific general

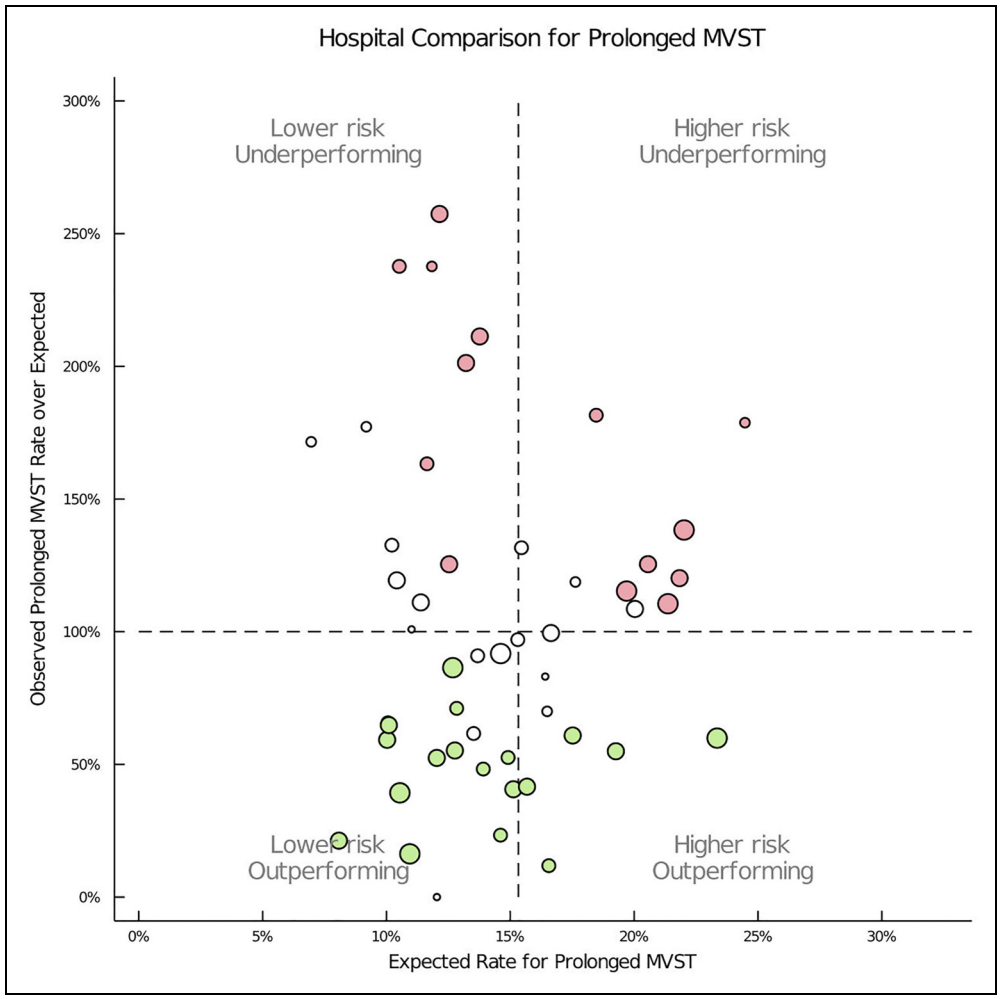

Figure 4. Comparison of prolonged mechanical ventilatory support time (MVST) for all hospitals.

preoperative factors (eg, diagnosis, age, weight, prematurity, history of prior cardiac procedures, and presence of other noncardiac anomalies such as genetic and chromosomal defects), and other preoperative factors indicating clinical status (eg, preoperative mechanical ventilation or circulatory support, etc). Importantly, this methodology does not assume that these factors have any a priori relationship between themselves, nor that they interact in a linear and additive fashion, as linear regression does. Once the model's decision tree is established, it may be considered conceptually as demonstrating the performance of a "virtual" hospital where all ECDB patients have been treated. The uniqueness of our methodology lies in that when an individual hospital's performance is analyzed, our model can calculate what the predicted outcomes would be if the given hospital's specific case-mix were treated in the virtual hospital, thereby determining a hospital-specific case-adjusted benchmark, against which the real observed performance of the hospital can be assessed. Thus, the case-mix of the index hospital becomes identical, in so far as the recorded data reflect, to the case-mix of the virtual hospital, practically eliminating the contributions of case-mix variation to estimated risk. Accordingly, each hospital's performance is assessed by comparison to their own predicted outcomes for their own specific cases and case-mix, based on the OCT analysis of aggregate data from the "virtual hospital."

It should be noted that, alternatively, we could assess the reverse, ie, the potential performance of each individual hospital if it treated "the virtual hospital's" patient population. However, if the index hospital sees none or very few patients of a particular virtual hospital cohort, eg, if a hospital rarely performs a particular operation, such as the Norwood operation, which may be the case in situations of differing national policies, it would be unfair to extrapolate this hospital's very limited experience relevant to such a procedure to predict its performance as if it treated the virtual hospital's population. Therefore, we focus on the expected rate comparison described above as the hospital-specific benchmark metric.

Importantly, this methodology is not limited to evaluation of the aggregate performance of each hospital and has the additional power to automatically assess the performance of the same patient cohorts in the index as in the virtual hospital, thereby revealing possible areas of strength (overperformance) and areas of opportunity (underperformance). The analysis of performance is available for each patient cohort, each of which comprises patients with a similar risk. The criterion for

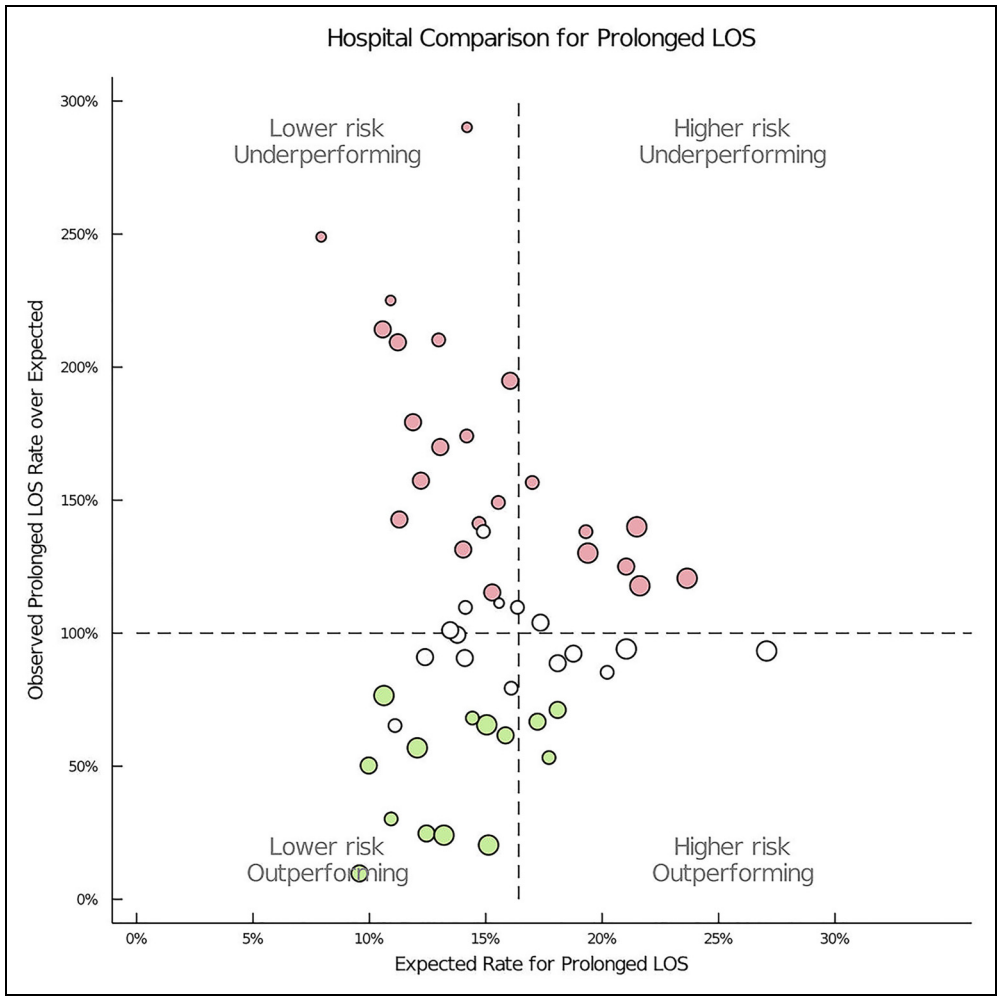

Figure 5. Comparison of prolonged length of hospital stay (LOS) for all hospitals.

assigning a patient to a cohort is the combination of shared characteristics and statistically similar risks described by the relevant pathway. Therefore, heterogeneous procedures may well be included in the same cohort. Obviously, as more data accumulates in the database, greater granularity and statistical differentiation of subcohorts can be revealed with the algorithm proceeding to deeper levels.

In addition, our analysis automatically further breaks down the performance of each cohort into its components derived from each of the 10 benchmark procedures (Figure 11). Therefore, a detailed risk-adjusted view of the participating hospital is provided, analyzed both by the risk group and by the procedure. Accordingly, our benchmarking analysis, which is not limited to overall performance assessment, but dissects into important components of performance, can serve as a powerful self-assessment and quality improvement tool for each hospital.

Limitations

While limiting the analysis to the 10 benchmark procedure groups reduces data heterogeneity and facilitates statistical analysis, considering that benchmark procedures represent only 45.7% of the total recorded in the ECDB, exclusion of many albeit infrequently performed procedures may mask potentially important overperformance or underperformance of centers regarding procedures not studied. For many tertiary care hospitals, the percentage of rare, complex, and high-risk cases (some of which may be performed almost exclusively in specialized centers) not included in the 10 benchmark procedures may be substantial, and use of additional metrics, such as adjusted outcome estimates for all procedures, both overall and by risk category, is important. Accordingly, in our ongoing research, hospital performance for both benchmark and non-benchmark procedures will be addressed.

Although all preoperative factors recorded in the database have been entered in the analysis, there are unrecorded or unknown patient factors that may be significant risk contributors: genetic factors independent of other patient and operative features may play a significant role (such as the apolipoprotein E-ϵ2 allele in neurodevelopmental outcome after neonatal CHS[18]), yet these are generally not clinically tracked. We also know that there are specific features relevant only to certain procedures (eg, coronary anatomy for transposition of

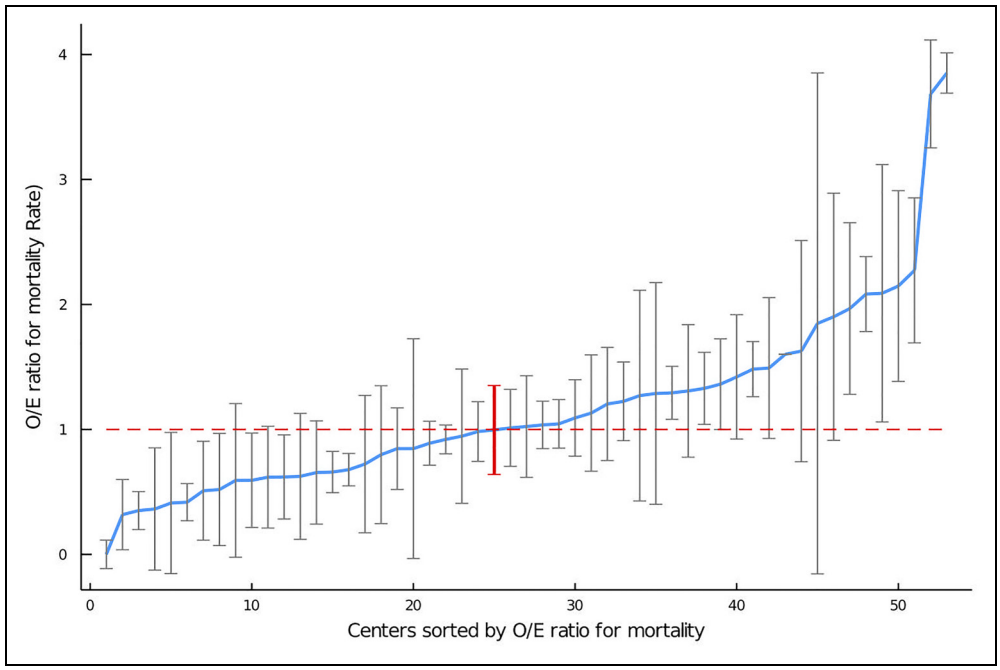

Figure 6. A plot of hospital-specific observed-to-expected (O/E) ratios for mortality with 95% confidence intervals (grey bars), for each hospital, sorted by increasing the O/E ratio. The red dotted line is at ratio 1.0 for observed equally to expected mortality. The O/E ratio of the example hospital, highlighted in red, is at unity, no different from expected.

the great arteries), which have not been tracked in the ECDB, as it was, at its inception, structured to capture the exact same dataset for all procedure types. Thus, procedure-specific factors have not been recorded in the ECDB, which is now in the process of adding such features (data fields) to its data collection software. We hope that such important additional granular information will permit the achievement of even more accurate predictions in the future.

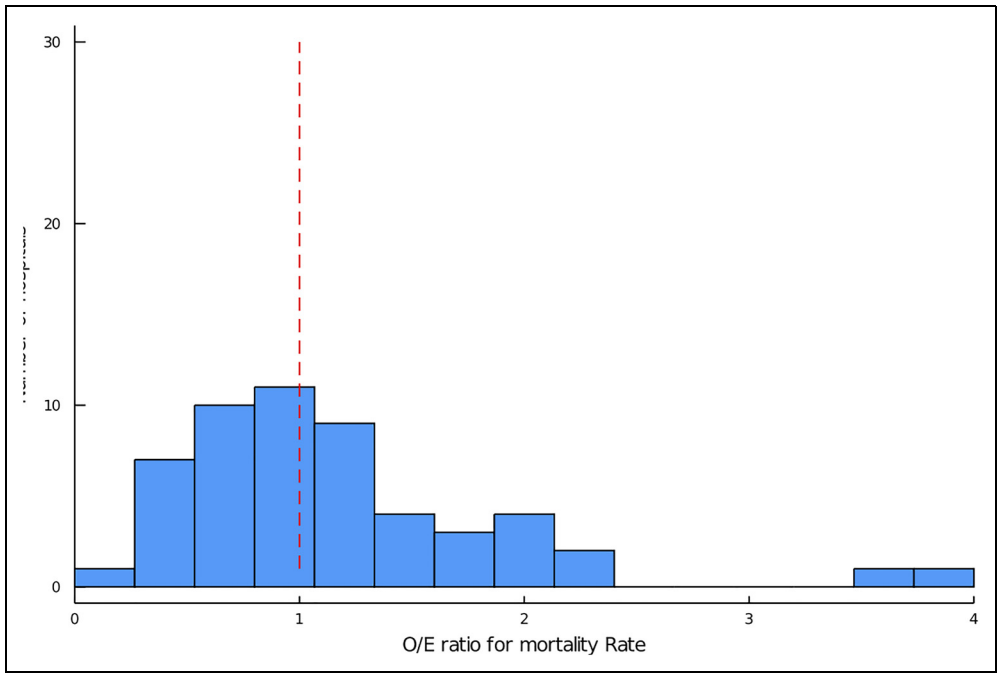

Figure 7. The distribution of hospitals by the observed-to-expected (O/E) ratios with a ratio of 1.0 for no difference, in red. Of all hospitals, 11.3% have O/E ratios above 2.

Table 3. Example "Hospital A" Performance Summary.

Outcome	Actual rate (90% confidence interval)	Average rate of virtual hospital (%)	Expected rate (%)	Performance difference referenced to the expected rate
Mortality	7.2% [6.0%, 8.5%]	4.4	7.2	Not statistically significant ($p = 1.0$)
Prolonged MVST	22.7% [20.7%, 24.8%]	15.3	19.7	Worse ($p = .011$)
Prolonged LOS	25.7% [23.4%, 27.6%]	15.5	21.6	Worse ($p = .002$)

Abbreviations: MVST, mechanical ventilatory support time; LOS, length of hospital stay.

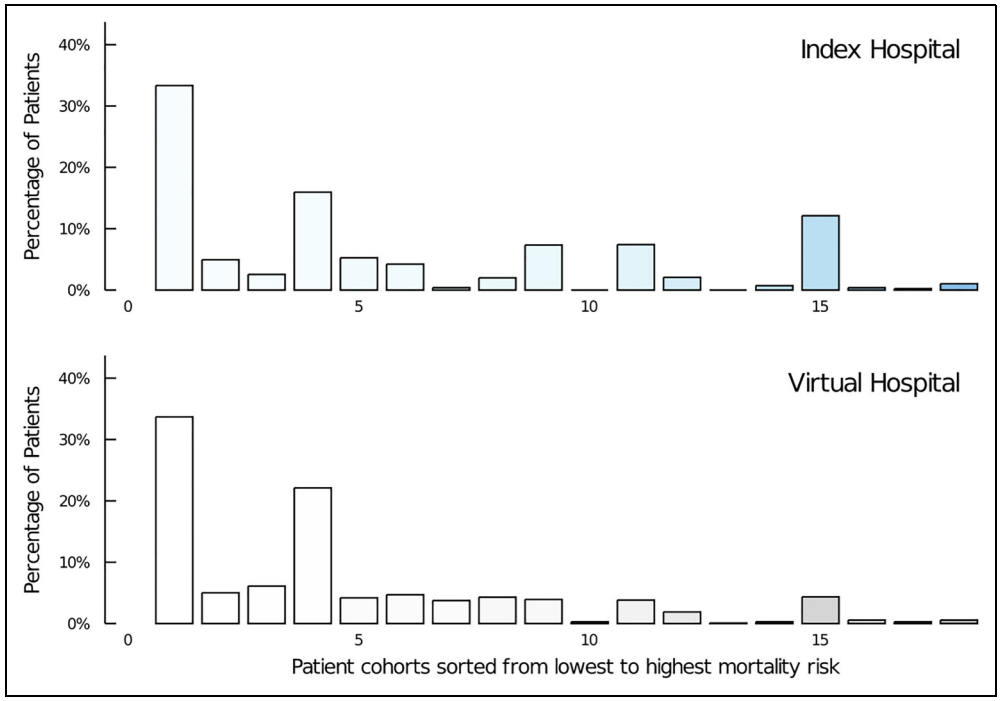

Figure 8. Overall case-mix for the index hospital compared to the virtual hospital regarding mortality. The *y*-axis refers to the percentage of patients in each cohort. Shading intensity reflects risk, darker bars indicating higher mortality risk. The case-mix at this hospital has an overall mortality risk (ie, expected mortality rate) of 7.1% compared to the virtual hospital's 4.4%, hence it is seeing overall "higher risk" patients.

An example of the importance of improving data granularity may be seen in the examination of the OCT model calibration plots (Supplemental Figures S1 and S2). The model calibration plots produced for each individual procedure group and for each outcome separately (Supplemental Figure S2, all necessarily based on aggregated data for all procedures and from all hospitals), generally indicate good model performance, as do the calibration plots for all the procedure groups considered together (Supplemental Figure S1). However, for some procedure groups (eg, Fontan and Hemi-Fontan), the calibration plots indicate overprediction, and the number of points, which correspond to the number of statistically distinct OCT terminal leaves (ie, to the number of similar risk cohorts which include patients who have undergone the procedure examined) is small. We believe that this is a reflection of a smaller number of data records for these procedures and, importantly, of limited data granularity, ie, of lack of information recorded in the database on procedure-specific risk factors. Availability of large numbers of patients for each individual procedure examined along with relevant procedure-specific information (over and above the features collected for all procedures in the database) should result in the OCT algorithm revealing more terminal leaves, ie, would permit our OCT methodology to detect more distinct risk patient cohorts. In other words, we believe that a smaller number of distinct risk cohorts and overprediction in specific procedures is a data limitation, not 1 of the OCT methodology. Enrichment of the database with procedure-specific risk factors may enable us to build separate OCTs for each procedure, a task beyond the scope of this paper. We are well aware of the importance of adding to the ECHSA database procedure-specific risk factors, as this should lead to further increases of the predictive accuracy of our AI and ML-based OCT approach (as it also would for traditional statistical approaches). Accordingly, this project is

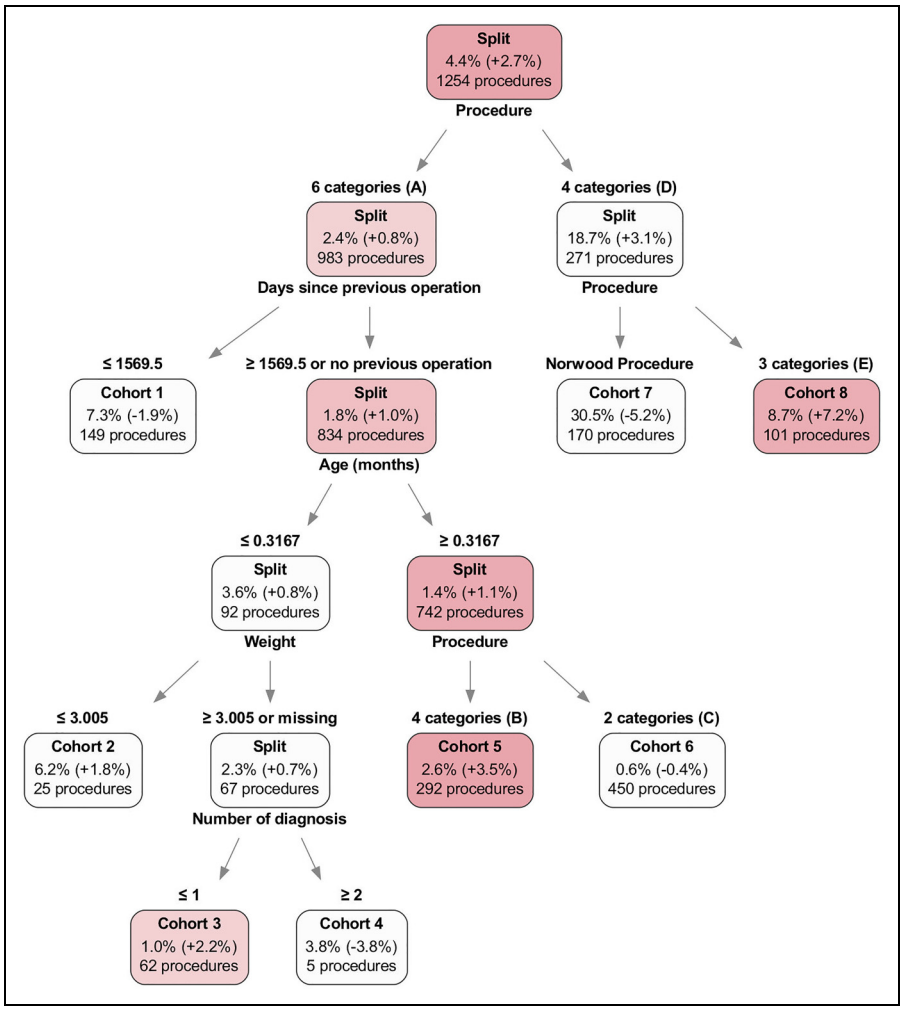

Figure 9. Optimal classification tree (OCT) for mortality for the index hospital. The tree identifies 8 unique patient cohorts of similar risk for this hospital. Cohorts of distinction (better than expected performance), if any, would have been in green. Cohorts of opportunity (lower than expected performance) are in red.

Cohort 3 Virtual Hospital Mortality 1.0% This Hospital Mortality 3.2% (p = 0.085)	1) Procedure is AVC (AVCanal or AVSeptal Defect Repair), Arterial Switch Op, Arterial Switch Op with VSD Repair, Off Bypass Coarctation Repair, Tetralogy of Fallot Repair, VSD Repair 2) Days since previous admission ≥ 1569.5 or no previous operation 3) Age (months) ≤ 0.3167 4) Weight ≥ 3.005 or missing 5) Number of diagnosis ≤ 1.5
Cohort 5 Virtual Hospital Mortality 2.6% This Hospital Mortality 6.2% (p = ≤ 0.001)	1) Days since previous admission ≥ 1569.5 or no previous operation 2) Age (months) ≥ 0.3167 3) Procedure is AVC (AVCanal or AVSeptal Defect Repair), Arterial Switch Op, Arterial Switch Op with VSD Repair, Tetralogy of Fallot Repair
Cohort 8 Virtual Hospital Mortality 8.7% This Hospital Mortality 15.8% (p = 0.009)	1) Procedure is Fontan, Glenn Hemi-Fontan, Truncus Repair

Figure 10. Characteristics of cohorts with higher than expected mortality.

Cohort	Overall	AVC (AVCanal or AVSeptal Defect Repair)	Arterial Switch Op	Arterial Switch Op with VSD Repair	Fontan	Glenn Hemi-Fontan	Norwood Procedure	Off Bypass Coarctation Repair	Tetralogy of Fallot Repair	Truncus Repair	VSD Repair
All Procedures	Actual 7.2% Virtual 7.2% 1254 procedures	Actual 11.6% Virtual 4.4% 95 procedures	Actual 4.0% Virtual 5.0% 125 procedures	Actual 20.0% Virtual 6.2% 50 procedures	Actual 0.0% Virtual 4.2% 2 procedures	Actual 13.2% Virtual 7.3% 76 procedures	Actual 25.3% Virtual 30.6% 170 procedures	Actual 0.0% Virtual 1.6% 130 procedures	Actual 1.0% Virtual 3.9% 201 procedures	Actual 26.1% Virtual 19.0% 23 procedures	Actual 0.8% Virtual 1.3% 382 procedures
Cohort 1	Actual 5.4% Virtual 7.3% 149 procedures	Actual 13.0% Virtual 9.6% 23 procedures	Actual 0.0% Virtual 12.2% 7 procedures	Actual 33.3% Virtual 12.0% 6 procedures				Actual 0.0% Virtual 12.4% 5 procedures	Actual 1.3% Virtual 4.5% 75 procedures		Actual 6.1% Virtual 4.0% 33 procedures
Cohort 2	Actual 8.0% Virtual 6.2% 25 procedures		Actual 0.0% Virtual 5.0% 11 procedures	Actual 66.7% Virtual 8.4% 3 procedures				Actual 0.0% Virtual 5.9% 11 procedures			
Cohort 3	Actual 3.2% Virtual 1.0% 62 procedures		Actual 4.4% Virtual 0.9% 45 procedures	Actual 0.0% Virtual 3.2% 7 procedures				Actual 0.0% Virtual 0.5% 10 procedures			
Cohort 4	Actual 0.0% Virtual 3.8% 5 procedures		Actual 0.0% Virtual 3.7% 1 procedures	Actual 0.0% Virtual 6.1% 1 procedures				Actual 0.0% Virtual 2.8% 3 procedures			
Cohort 5	Actual 6.2% Virtual 2.6% 292 procedures	Actual 11.1% Virtual 3.2% 72 procedures	Actual 4.9% Virtual 5.8% 61 procedures	Actual 18.2% Virtual 8.4% 33 procedures					Actual 0.8% Virtual 1.5% 126 procedures		
Cohort 6	Actual 0.2% Virtual 0.6% 450 procedures							Actual 0.0% Virtual 0.7% 101 procedures			Actual 0.3% Virtual 0.6% 349 procedures
Cohort 7	Actual 25.3% Virtual 30.5% 170 procedures						Actual 25.3% Virtual 30.5% 170 procedures				
Cohort 8	Actual 15.8% Virtual 8.7% 101 procedures				Actual 0.0% Virtual 4.7% 2 procedures	Actual 13.2% Virtual 6.3% 76 procedures				Actual 26.1% Virtual 17.4% 23 procedures	

Figure 11. Cohort summary for mortality, showing performance for each cohort broken down by benchmark procedure.

underway at the ECDB. Furthermore, we emphasize that our models reflect the current status of the field. Ongoing developments will be reflected in future data harvests. It is an advantage of our methodology that, after the algorithms are set, routine periodic recalculations based on updated data are readily feasible.

We also recognize that benchmarks are calculated with reference to the case-mix and results recorded only from European participating centers. Therefore, although the models calculated are internally valid, they may not accurately predict performance in other practice environments. In the future, in the context of international cooperative efforts, it would be of value to test our methodology with other large datasets of CHS.

Finally, we emphasize our work's focus is to demonstrate how our ML OCT methodology can be applied to calculate benchmarking standards for the benefit of individual ECDB participating center's self-assessment and quality improvement efforts (using European centers' data as an example), and not to present generally applicable standards, nor to engage in hospital comparisons for the public.

Acknowledgments

We gratefully acknowledge the contributions of ECDB participating surgeons and centers, whose CHS data have made this study possible.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

George E Sarris https://orcid.org/0000-0002-9341-2850

Supplemental Material

Supplemental material for this article is available online.

References

1. Jenkins KJ, Gauvreau K, Newburger JW, Spray TL, Moller JH, Lezzoni LI. Consensus-based method for risk adjustment for surgery for congenital heart disease. *J Thorac Cardiovasc Surg*. 2002;123:110–118.
2. Lacour-Gayet F, Clarke D, Jacobs J, et al. The Aristotle score: a complexity-adjusted method to evaluate surgical results. *Eur J Cardiothorac Surg*. 2004;25:911–924.
3. 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.
4. Jacobs JP, O'Brien SM, Pasquali SK, et al. Variation in outcomes for benchmark operations: an analysis of The society of thoracic surgeons congenital heart surgery database. *Ann Thorac Surg*. 2011;92:2184–2192.
5. Jacobs JP, O'Brien SM, Pasquali SK, et al. Variation in outcomes for risk-stratified pediatric cardiac surgical operations: an analysis of the STS congenital heart surgery database. *Ann Thorac Surg*. 2012;94:564–572.
6. Jacobs JP, O'Brien SM, Pasquali SK, et al. The importance of patient-specific preoperative factors: an analysis of the society of

thoracic surgeons congenital heart surgery database. *Ann Thorac Surg*. 2014;98:1653–1659.
7. O'Brien SM, Jacobs JP, Pasquali SK, et al. The society of thoracic surgeons congenital heart surgery database mortality risk model: part 1—statistical methodology. *Ann Thorac Surg*. 2015;100:1054–1062.
8. Jacobs JP, O'Brien SM, Pasquali SK, et al. The society of thoracic surgeons congenital heart surgery database mortality risk model: part 2—clinical application. *Ann Thorac Surg*. 2015;100:1063–1070.
9. Pasquali SK, Jacobs ML, O'Brien SM, et al. Impact of patient characteristics on hospital-level outcomes assessment in congenital heart surgery. *Ann Thorac Surg*. 2015;100:1071–1077.
10. Jacobs JP, Mayer JEJr, Pasquali SK, et al. The society of thoracic surgeons congenital heart surgery database: 2019 update on outcomes and quality. *Ann Thorac Surg*. 2019;107:691–704.
11. Jacobs JP, O'Brien SM, Hill KD, et al. Refining the society of thoracic surgeons congenital heart surgery database mortality risk model with enhanced risk adjustment for chromosomal abnormalities, syndromes, and noncardiac congenital anatomic abnormalities. *Ann Thorac Surg*. 2019;108(2):558–566.
12. Pasquali SK, Gaies M, Banerjee M, et al. The quest for precision medicine: unmeasured patient factors and mortality after congenital heart surgery. *Ann Thorac Surg*. 2019;108:1889–1894.
13. Jacobs ML, Jacobs JP, Thibault D, et al. Updating an empirically based tool for analyzing congenital heart surgery mortality. *World J Pediatr Congenit Heart Surg*. 2021;12(2):246–281.
14. Spray TL, Gaynor WL. A word of caution in public reporting. *Semin Thorac Cardiovasc Surg Pediatr Card Surg Ann*. 2017;20:49–55.
15. Rogers L, Brown KL, Franklin RC, et al. Improving risk adjustment for mortality after pediatric cardiac surgery: the UK PRAiS2 model. *Ann Thorac Surg*. 2017;104:211–219.
16. Bertsimas D, Zuccarelli E, Smyrnakis N, et al. Adverse outcomes prediction for congenital heart surgery. A machine learning approach. *World J Pediatr Congenit Heart Surg*. 2021;12(4):453–460. Published online, April 28, 2021. DOI:10.1177/21501351211007106
17. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMA. Surgical risk is not linear. Derivation and validation of a novel, user-friendly, and machine-learning-based predictive optimal trees in emergency surgery risk (POTTER) calculator. *Ann Surg*. 2018;268:574–583.
18. Gaynor JW, Kim DS, Arrington CB, et al. Validation of association of the apolipoprotein E ϵ2 allele with neurodevelopmental dysfunction after cardiac surgery in neonates and infants. *J Thorac Cardiovasc Surg*. 2014;148(6):2560–2568.