Adverse Outcomes Prediction for Congenital Heart Surgery: A Machine Learning Approach

Dimitris Bertsimas, PhD[1] ⓘ **, Daisy Zhuo, PhD[2,3]** ⓘ **, Jack Dunn, PhD[2,3]** ⓘ **,**
Jordan Levine, MEng[2,3] ⓘ **, Eugenio Zuccarelli, MBAn, MSc[1]** ⓘ **,**
Nikos Smyrnakis[4] ⓘ **, Zdzislaw Tobota, MD[5],**
Bohdan Maruszewski, MD, PhD[5] ⓘ **, Jose Fragata, MD, PhD[6],**
and George E. Sarris, MD, PhD[7] ⓘ

Abstract

Objective: Risk assessment tools typically used in congenital heart surgery (CHS) assume that various possible risk factors interact in a linear and additive fashion, an assumption that may not reflect reality. Using artificial intelligence techniques, we sought to develop nonlinear models for predicting outcomes in CHS. **Methods:** We built machine learning (ML) models to predict mortality, postoperative mechanical ventilatory support time (MVST), and hospital length of stay (LOS) for patients who underwent CHS, based on data of more than 235,000 patients and 295,000 operations provided by the European Congenital Heart Surgeons Association Congenital Database. We used optimal classification trees (OCTs) methodology for its interpretability and accuracy, and compared to logistic regression and state-of-the-art ML methods (Random Forests, Gradient Boosting), reporting their area under the curve (AUC or c-statistic) for both training and testing data sets. **Results:** Optimal classification trees achieve outstanding performance across all three models (mortality AUC = 0.86, prolonged MVST AUC = 0.85, prolonged LOS AUC = 0.82), while being intuitively interpretable. The most significant predictors of mortality are procedure, age, and weight, followed by days since previous admission and any general preoperative patient risk factors. **Conclusions:** The nonlinear ML-based models of OCTs are intuitively interpretable and provide superior predictive power. The associated risk calculator allows easy, accurate, and understandable estimation of individual patient risks, in the theoretical framework of the average performance of all centers represented in the database. This methodology has the potential to facilitate decision-making and resource optimization in CHS, enabling total quality management and precise benchmarking initiatives.

Introduction

Despite great progress achieved in the surgical management congenital heart disease (CHD), there is still considerable associated risk of death and complications, nonuniformly distributed across specific conditions and treatments. Complications may lead to prolonged length of postoperative mechanical ventilatory support time (MVST) and overall hospital length of stay (LOS). In the context of efforts to promote quality improvements in congenital heart surgery (CHS), large databases of such operations have been developed, along with methodologies seeking to establish benchmarks for surgical results, and to devise risk models for predicting important outcome parameters.[1-18] However, methodologies used typically assume that various possible risk factors interact in a linear and additive fashion, an erroneous assumption. Using artificial

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

Corresponding Author:
George E. Sarris, Athens Heart Surgery Institute, Kifissias Avenue #2, Amaroussion, Athens 15125, Greece.
Email: gsarris@mac.com

Abbreviations	
AI	artificial intelligence
AUC	area under the curve (or c-statistic)
CHD	congenital heart disease
CHS	congenital heart surgery
CPB	cardiopulmonary bypass
EACTS	European Association for Cardiothoracic Surgery
ECDB	European Congenital Heart Surgeons Association Congenital Database
ECHSA	European Congenital Heart Surgeons Association
GPRF	general preoperative risk factor
LOS	length of hospital stay
ML	machine learning
MVST	mechanical ventilatory support time
OCT	optimal classification trees
STS-CHSD	Society of Thoracic Surgeons Congenital Heart Surgery Database

Table 1. The Preoperatively Known Variables (Potential Risk Factors) Analyzed.

Variable
Age
Weight
Gender
History of any prior cardiac procedure(s)
Days since previous admission, if any
Number of preoperative diagnoses
Antenatal diagnosis known
Any noncardiac abnormalities present
Any general preoperative risk factor present
Case category (CPB vs non-CPB)
Number of concomitant procedures performed
Year of procedure
Procedure

Abbreviation: CPB, cardiopulmonary bypass.

intelligence (AI) techniques, we developed nonlinear models devoid of any assumptions regarding possible variable interactions, aiming to predict outcomes after CHS.

Methods

Data

The data, provided by the European Congenital Heart Surgeons Association (ECHSA) Congenital Database (ECDB; https://echsacongenitaldb.org) after study review and approval by the ECDB Committee regarding compliance with all ECHSA ethical and patient data protection policies, comprise fully anonymized information regarding patients with CHD who have undergone surgery in participating hospitals from Europe and around the world. The full spectrum of pediatric acquired and congenital, and adult congenital cardiothoracic operations is represented in the database. Exclusion and inclusion criteria used to select the data cohort of 221,335 procedures analyzed, covering admissions from January 1, 2000, to September 30, 2019, are shown in supplemental Table S1-A, and relevant summary descriptive statistics in Tables S1-B and S1C.

Machine Learning Methodology

In the quest to achieve improved patient outcomes after CHS, it is important to be able to make accurate predictions of the risk of *patient-specific* adverse outcomes such as mortality, MVST intensive care unit (ICU), and LOS. Furthermore, it is important to be able to understand the predictive model's decision process, that is, the model needs to be logically interpretable and intuitively understandable. Traditionally used linear models (eg, logistic regression) do not capture nonlinear effects of risk-factor interplay and are accordingly limited in this regard. On the other hand, "black box" machine learning (ML)-based models can be very accurate but are complex and not intuitively understandable. However, in medicine, it is crucial that prediction models be transparent and understandable by clinicians. Accordingly, we decided to use ML methods providing

interpretability, ranging from population level to individual level, therefore limiting the use of "black box" approaches. Methods providing *individual-level interpretability* allow understanding of which variables are important for a particular patient, as opposed to other variables (eg, body weight), which may be significant for the population but may indeed be unimportant for an individual. We used a recent breakthrough in ML methodology, the optimal classification trees (OCTs), which can capture nonlinear variable interactions while providing individual-level interpretability.[19,20] For comparison, we also used the following ML methodologies: logistic regression, random forests, and gradient boosting. Logistic regression, widely used in medical statistics and a classic example of linear models, is simple to understand. Random forests and gradient boosting are state of the art in performance but are considered "black box" methods, as they do not provide individual-level interpretability.

The entire data set was analyzed for estimating mortality risk and surgical survivors for MVST and LOS. For each of the three models built (mortality, MVST, and LOS), only the *preoperatively known* patient variables listed in Table 1 are used, with more detailed information provided in Table S2. The data were split into *a training set* (procedures prior to January 1, 2016, N = 175,239), used to train the ML model, and a *test set* (procedures on or after January 1, 2016, N = 46,096), that is, data hidden from the model, in order to test the model's performance on previously "unseen" data. Compared with training data, the test data set describes "future" operations.

The training set was further split (using stratified sampling, ensuring a similar proportion of survivors and nonsurvivors in each subset) to produce a separate *validation set*, the role of which was to tune *algorithm hyperparameters*: We select a set of hyperparameters (eg, for OCT models, tree depth, or the minimum number of patients in each leaf), a classification tree was built using training data, and then those hyperparameter values which perform best in the validation set were chosen.

The ML predictive task consists of a binary classification problem aiming to predict one of two possible outcomes for mortality, MVST, and LOS:

Table 2. Comparative Performance of Different Methods of Building Risk Estimation Models for Mortality, Prolonged MVST, and Prolonged LOS.

Method	AUC for mortality		AUC for prolonged MVST		AUC for prolonged LOS	
	Training (%)	Testing (%)	Training (%)	Testing (%)	Training (%)	Testing (%)
Logistic regression	75.3	77.2	80.9	81.3	71.5	72.8
Optimal classification trees	85.5	86.2	85.6	84.8	80.4	81.0
Random forest	85.9	87.3	85.2	85.4	80.1	82.1
Gradient boosting	85.6	87.4	85.5	85.6	80.4	82.0

Abbreviations: AUC, area under the curve; LOS, length of hospital stay; MVST, mechanical ventilatory support time.

For *Mortality*, the binary outcome is whether the patient died in the hospital (or within 30 days) after the procedure or not.

For prolonged *Postoperative MVST* defined as *Time of extubation in ICU (or, time of last extubation in cases of reintubation(s))—Time of end of surgery*, the binary classification problem was to predict whether the duration of MVST was prolonged, selected as being above or below the 85th percentile MSVT (94.5 hours).

For prolonged *LOS*, defined as date of discharge–date of surgery, the binary problem was to predict whether the duration of hospitalization is above or below the selected threshold of 85th percentile (19 days).

Results

Average mortality is 5.1% in the training, 3.0% in the test, and 4.7% in the overall data sets analyzed. Univariate associations between preoperative variables and mortality in the training set are shown in Table S3. By definition, the average rate for prolonged LOS and prolonged MVST is 15%. For the three models built, the area under the receiver operating characteristic curve (AUC) for the outcomes of mortality, LOS, and MVST, using the different methodologies, is reported in Table 2. Area under the curve was chosen as a metric particularly useful to evaluate binary classification tasks, especially for highly unbalanced data sets, like the one analyzed. Calibration plots for the OCT models in training and testing data are shown in supplemental Figure S1.

The importance of each preoperative variable in predicting outcome was also computed and listed in Table S4. The most important risk factors are *procedure* and patient's age and *weight*. Procedure alone accounts for >50% of the features' contribution. A short time (<12 days) since the last (of any) prior operation and the presence of any general preoperative risk factors (GPRFs) also contribute significantly to the likelihood of mortality. The remaining factors have marginal significance, with each contributing less than 2% toward the odds of mortality.

The mortality risk model presents itself clearly as a decision tree, shown for the outcome of mortality in Figure 1A. It is possible to follow this tree along its branches to the lowest level terminal leaves, which identify 23 distinct patient cohorts and

their respective predicted mortalities. A summary description of an example cohort is presented in Table S5. At each branching step along the decision tree, the combination of factors involved and the logic behind each step is fully transparent. Thus, given the preoperative features of an individual patient, one can read the tree along its branching points and arrive at the predicted risk, as shown in Figure 1B. By direct analogy, the decision trees for predicting MVST and LOS (shown in Figures S3 and S4) are also fully interpretable. Therefore, it becomes clear why our OCT methodology is fully interpretable at the individual patient level.

Congenital Heart Surgery Adverse Outcome Prediction Tool

The OCT-derived risk models were used to produce a clinician-friendly software tool that aims to permit easy estimation of an individual patient's risk for a given procedure in the theoretical framework of the average performance of all centers represented in the database. One can input *preoperatively* available patient information on the smartphone app (Figure S2), leading to the immediate calculation of risk estimates (mortality, prolonged MVST, or prolonged LOS) for the specific patient and the proposed procedure.

Comment

The ECHSA ECDB was established in 1995, originally as the European Congenital Heart Defects Database and renamed as the European Association for Cardiothoracic Surgery (EACTS) Congenital Database in September 1999, acquiring its current final name in 2015. The ECDB collects data from participating centers regarding pediatric and adult congenital cardiac operations, aiming to assess results for scientific study and to provide tools for individual programs' own quality improvement efforts. The ECDB has collaborated closely with the Society of Thoracic Surgeons (STS) Congenital Heart Surgery Database (CHSD), developing common nomenclature of cardiac defects and identical data fields, enabling sharing data in joint research efforts.[5] European Congenital Heart Surgeons Association Congenital Database research is documented in multiple publications listed on its website: https://www.echsacongenitaldb.org/publications/.
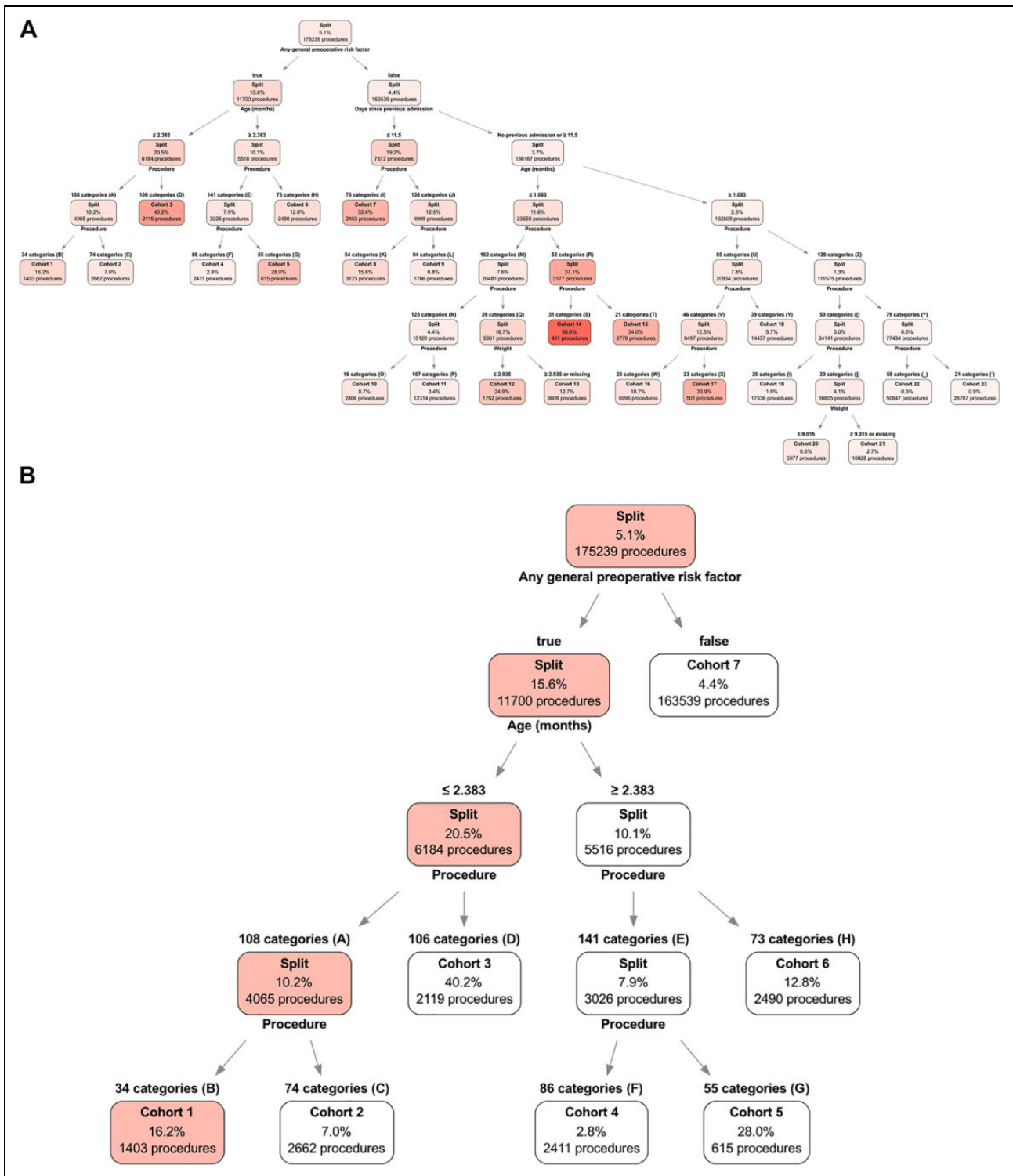
Figure 1. A, Mortality decision tree based on the optimal classification tree (OCT) algorithm. Greater intensity of the color in each leaf depicts increased mortality of the cohort. At the root level, the overall mortality rate for 175,239 procedures analyzed in the training data set is 5.1% (note that since the rate is limited to training data prior to 2016, the rate is higher than the overall database average of 4.7%, and of 3.0% in the test data set of 3.0%). The first branch split occurs based on the presence of any general preoperative risk factors (GPRF). If we follow the right branch (no GPRF), we see that the next branching is based on days since previous admission, age, procedure, and weight. If we follow the left branch again, depending on the procedure and age, the mortality can range from 2.8% to 40.2%. It is therefore possible to follow the tree along its branches to the lowest level terminal leaves, which identify 23 patient cohorts and their respective predicted mortalities. At each branching step along the decision tree, the combination of the factors involved and the logic behind each step is fully transparent. Thus, given the preoperative features of an individual patient, one can read the tree along its branching points and arrive at the predicted risk, as shown in B. (B) The prediction of mortality risk (16.2%) for a patient with diagnosis transposition of the great arteries with VSD, weight 3 kg, preoperative ventilation and resolved shock at the time of surgery, age 2 weeks, undergoing an arterial switch operation with VSD closure, is highlighted on the decision tree.

It is established that assessment of surgical results must take into account, on one side, the inherent complexity of the diagnosis, the individual patient's characteristics, the disease's severity, and the nature of the procedure, and on the other side, the performance of the surgeon, team, and institution. In this article, we have focused on the impact of various features of the underlying pathologies and preoperative patient characteristics recorded in the ECDB. Important outcomes to study include not only mortality but also the duration of stay in the ICU, on ventilator support (MVST), or in the hospital (LOS). Other quality metrics are also important, as well as costs, which, in turn, are influenced by complications, MVST, and LOS. *Such additional outcome metrics along with relevant institutional factors are the focus of further studies by our group.*

Major efforts to adjust for "case-mix," that is, for variability in the inherent risk of various procedures, and to account for effects of variation in patient characteristics, have evolved from initial attempts based on expert consensus (RACHS-1[1,2] and Aristotle[3,4] methods), to those empirically derived, based on outcome measures provided by real data[5-18] (STS-EACTS Mortality Score and Categories, so-called STAT Score and Categories, and STS Morbidity Score and Categories, and use of accurate risk prediction models, such as the STS-CHSD Mortality Risk Model[14] and the UK PRAiS2 Model).[18] Still, it is recognized that patient factors typically used in these models only explain a small proportion of variation in mortality.[16]

However, the aforementioned analytical methods, although achieving very high predictive accuracy, are flawed in assuming that various possible risk factors for adverse outcome interact in a linear and additive fashion, an assumption which frequently does not reflect reality. For example, although prematurity and low birth weight are intuitively understood risk factors for CHS, their influence is practically absent in older patients and for most procedures. In recognition of the existence of potential nonlinear interactions of risk factors, the STS-CHSD Mortality Risk Model has included several variables and adjustments, such as condition/age interactions and condition/age/procedure interactions. For example, the effect of Down syndrome was estimated based on age and procedure subgroups, including atrioventricular canal repair and single ventricle palliation.[14] However, capturing the importance of such interactions required *prior manual creation* of possible candidates of interactions and use of a Bayesian model to identify groups as similar as possible.

Furthermore, published methods have not provided physicians with a practical tool to predict mortality or morbidity of a given patient with CHD considered for surgery, given his preoperative risk profile.

To circumvent such limitations, we have used AI techniques, which some of us have previously used in other medical applications,[19-20] seeking to develop (1) nonlinear models, free from any assumptions regarding the importance and interplay of preoperative variables or their combinations, aiming to predict outcome after CHS and (2), a user-friendly tool enabling clinicians to estimate outcome risks by inputting a specific patient's characteristics. Thus, we sought to develop models to predict mortality after CHS, and for survivors, to estimate the risk of prolonged MVST and LOS.

Our results demonstrated that OCTs consistently outperformed the simpler linear regression and had similar predictive accuracy (within 1.5%) as the more complex ML models, however, with the significant advantage over the latter of ensuring full interpretability. It is, in fact, a major advantage of OCT methodology that the prediction model generated, presenting itself as a decision tree, is fully transparent. Furthermore, in addition to enabling a clear understanding of how any prediction is reached for an individual patient, the methodology automatically reveals patient cohorts of similar risk. These model features are evident by studying the OCT presented in Figure 1. It is easy to see the influence of each feature ("risk factor") by following the tree to each successively deeper level. Given the preoperative characteristics of a patient, one can follow the decision "path" along the branches of the tree, understanding at each step the sequential decisions taken, and also comprehend the characteristics of resulting patient cohorts with similar risk, as revealed by the tree's terminal leaves. The ability to understand the decisions of the OCT algorithm is in advantageous contrast with findings of various studies in similar settings showing the use of "black-box" deep learning solutions,[21-27] which, despite higher performance indicators, have architectural inner workings which are obscure not just for clinicians, but for AI experts as well. Such black-box methods cannot provide interpretable explanations of "why" a given patient is assigned a certain outcome risk. The interpretability of our OCT methodology is extremely important since physicians, not typically versed in the complex AI and ML algorithms, need to understand how mathematical models achieve their predictions if they are to trust their results.

The risk factors identified by this analysis are consistent with clinical experience: procedure, age, and weight are the most powerful predictors of aggregate mortality, with significant contributions from the occurrence of a very recent other cardiac operation (<12 days, suggesting an unplanned early procedure possibly addressing some complication), and the presence of any of the various GPRFs listed in Table S2 (such as mechanical circulatory or ventilatory support). Of note, in our model, we have not used the factor of *preoperative diagnosis leading to operation* because it neither increased model accuracy nor did it result in clinically more useful decision trees.

Interestingly, the contribution to the models' predicting power of other preoperative variables was very small (eg, *case category,* cardiopulmonary bypass [CPB] *vs non-CPB*) or even absent (eg, *year of surgery*), despite the univariate association of these variables with the outcome (mortality), as shown in Table S3. In this nonlinear analytic system, the importance of such variables is overshadowed by the more powerful predictors listed above, especially when considering the effect of a variable such as *year of surgery* on *procedures with a small*

number of yearly observations, which precludes statistical significance.

Several additional points deserve emphasis: Because of well-known progress achieved over the last several years in CHS, particularly with regard to complex high-risk lesions,[13] we expect the model to overestimate mortality. When we recalibrate the model with a more recent validation set (2016-2017) and show results with validation and testing (2018-2019) calibration plots (Figure 1S), the model still overestimates risks for higher risk procedures, but, its predictions stabilize. In addition, our models still provide correct risk estimation in terms of order as reflected by high AUCs. In other words, AUC is a measure of ordering, not of the absolute magnitude of risk. Thus, although the overall risk in the "new data" is lower than predicted for higher risk procedures, AUC is still high because the order is correct. Clearly, to keep the predictive capacity of the model always current, our plan is to recalculate the model periodically, as new data accumulate.

We emphasize that our model predictions result mathematically *from the given data set* which is inclusive of data submitted from surgical practices of all ECDB participants, comprising both European and *(for 1/3 of data)* non-European centers. We did not limit our analysis to European centers, which might have resulted in greater data uniformity and model predictive accuracy, aiming to provide self-assessment tools applicable to *all* ECDB participating centers. It is possible that *specific predictions* of the model regarding particular patients and procedures may well be different if a different data set were analyzed, for example, our *European* data subset, or the STS-CHSD data reflecting North American practices. Although important lessons could be drawn from such possible future comparative studies, the focus of this study is to establish the new ML methodology of OCTs, not to derive specific clinical lessons from analysis of CHS data for any particular clinical application. In other words, we acknowledge that the emphasis of our current work is to demonstrate the utility of this ML methodology in the evaluation of CHS data in the framework of the total (geographically and temporally heterogeneous) experience recorded in the ECDB. Accordingly, precisely because of this heterogeneity, our presented model is not to be considered "a final product" to be used for assessment by all CHS centers, nor by all ECHSA or European centers. Such applications will obviously require recalculation of the models using the appropriate region-specific and contemporaneous homogeneous data sets.

We also acknowledged that the *same data fields are recorded for all patients* in the ECDB, it being a registry aiming to record all CHS activity in participating centers. Therefore, the ECDB has not captured all features, which may be important for all pathologies, procedures, and their innumerable combinations. Accordingly, efforts are underway to develop additional data fields to capture *important diagnosis and procedure-specific preoperative variables,* such as anatomic information of coronary arteries in Transposition of the Great Arteries (TGA) or Tetralogy of Fallot (TOF), and so on.

Future availability of such granular data may well further increase the predictive power of our methodology.

We note that all predictions are made on the procedure instead of on the patient level. Therefore, the cumulative estimated risk for a patient who may undergo multiple procedures would be higher than the individual procedure prediction.

We also note that surgical outcomes do not only depend on the disease, procedure, and patient-specific factors but also on other variables relating to the availability, quality, and organization of necessary health care resources, determined, for example, by geography, center, program size, and so on. The effects of such factors external to the patient were not addressed in this article, but are being analyzed in our continuing research, using the presented methodology, and will be presented in subsequent reports. Consequently, although our current risk calculator provides a user-friendly tool to predict risks of various procedures taking into account individual patient characteristics, we emphasize that such outcome predictions are based on cumulative data of all participating centers and do not reflect any individual center experience and performance level. Accordingly, our current version of the calculator is to be considered as a pilot tool for theoretical general risk prediction and is not intended to advise patients in the context of care provided by individual centers.

Any real-world, real-time model and application used in clinical practice to predict CHS risks requires the use of fairly contemporaneous data to develop the model. Although the models reported in this article demonstrate the potential of ML in the analysis of CHS outcomes, they are not yet ready for actual clinical application in a bedside calculator designed to predict CHS risks in specific centers. However, we hope that our ongoing evolving research, focusing on deep analysis of individual center performance, will allow center-specific predictions. Such more refined models to be developed, based on more homogeneous data subsets and taking into account institutional and possibly surgeon factors, may form the basis of future versions of the calculator which could have more clinical relevance to specific patient counseling.

Limitations

This methodology depends on a large number of good quality data, as are those in the ECDB. Although only a minority of data (approximately 14%) in the ECDB has been subjected to our on-site data verification process (in which independent database auditors visit volunteering centers and perform, on-site, detailed verification of 100% of submitted data fields and for 100% of the submitted patients), the outcomes of these verification analyses are regularly updated and published on the ECDB website (https://echsacongenitaldb.org/data_verification_results/), the results to date consistently demonstrating no statistically significant differences between the verified and unverified subsets.[28] Another limitation is that our analyses have involved a large variety of diagnoses, procedures, and their combinations, with major variation in their frequencies, ranging from relatively common to extremely rare. This

variability precludes having large numbers of data for rare conditions, for which many centers and surgeons may have little or no experience. Accordingly, the terminal leaves of the decision trees indicate frequently heterogeneous patient cohorts of similar risk. Finally, due to the large number of heterogeneous procedures with few observations, we chose not to use procedure-specific but rather cumulative 85th percentile MSVT and LOS. Although this could introduce bias against more complex procedures, auxiliary analyses using procedure-specific MSVT and LOS demonstrated similar model performance for various procedure types (Table S6). Such limitations originating in data heterogeneity are also addressed by our ongoing research analyzing the more common and most clinically important operations, chosen to be, for comparative purposes, the same ten "benchmark operation groups" as introduced by the STS-CHSD analyses.[5] Finally and most importantly, we acknowledge that the wide temporal and geographic heterogeneity of the data on which the models were trained, which likely reflect widely varying patterns of practice and outcomes, precludes the specific application of these models to any one specific regional practice (eg, Europe or North America). However, the same methodology can be easily applied to recalculate the models, for example, using European only data for application in Europe.

Conclusions

In summary, in seeking improved tools for risk prediction in CHS via analysis of the large data set in the ECHSA ECDB, the presented AI and ML-based models of OCTs, which are devoid of assumptions such as risk factor linearity, provide predictive power superior to traditional logistic regression and other competitive ML models, and, due to their intrinsic power, may need fewer variables than traditional methods to achieve accurate statistically significant predictions. Our methods operate entirely objectively, yet their decision-making process is easily understandable at each step. In other words, our approach has the added advantage of full intuitive interpretability of the method and its results, and its predictive models can be easily updated as new data accumulate. Our current risk calculator, based on these models, allows easy estimation of risks for individual patients given different preoperative scenarios, *in the theoretical framework of the average performance of all centers represented in the Database.*

Our work represents an initial evaluation of powerful ML tools and their application in the estimation of adverse outcome risks after CHS in the context of the challenges and limitations presented by the diverse spectrum of diagnoses, patients, procedures, and important limitations noted, including heterogeneity of patients, procedures, practices, and outcomes in a great variety of ECDB centers across multiple geographic regions, as well as necessary use of "historical" and limited data in developing predictive algorithms. *Therefore, we consider the strength of this study to rest more on its focus and demonstration of the potential of the methodology presented, rather than* on the actual results and clinical conclusions generated with the current "training" or development data set.

In short, our article demonstrates the potential of ML in the analysis of pediatric and congenital cardiac surgical outcomes. The approach described represents the initial step in an iterative process that will certainly evolve over time. Our methodology, which we plan to develop further to take into account hospital-derived features, has the potential to contribute significantly to quality control initiatives in CHS, opening the door to more precise and transparent benchmarking of outcomes. The current model reported in this article is a preliminary model; this current model is a scientific analysis that is not ready for clinical application. This current model is a general prototype developed based on heterogeneous sources of data, including both geographic heterogeneity and temporal heterogeneity. In order for such a model to be suitable for clinical application, it would need to be recalculated using data from a contemporary interval of time and a geographically proximate source (ie, using only European data rather than global data if the model is to be used clinically in Europe).

ORCID iDs

Dimitris Bertsimas https://orcid.org/0000-0002-1985-1003
Daisy Zhuo https://orcid.org/0000-0002-8579-932X
Jack Dunn https://orcid.org/0000-0002-6936-4502
Jordan Levine https://orcid.org/0000-0002-8766-9983
Eugenio Zuccarelli https://orcid.org/0000-0003-4847-083X
Bohdan Maruszewski https://orcid.org/0000-00034144-6677
Nikos Smyrnakis https://orcid.org/0000-0001-8522-7703
George E. Sarris, MD, PhD 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(1): 110-118.
2. McSharry B, Straney L, Alexander J, et al.; for the Australian and New Zealand Intensive care society paediatric study group and centre for outcomes and resource evaluation. RACHS-ANZ: a modified risk adjustment in congenital heart surgery model for

outcome surveillance in Australia and New Zealand. *J Am Heart Assoc*. 2019;8(9): e011390.

3. Lacour-Gayet F, Clarke D, Jacobs J, et al.; and The Aristotle Committee. The Aristotle Score: a complexity-adjusted method to evaluate surgical results. *Eur J Cardiothorac Surg*. 2004;25(6): 911-924.

4. Bojan M, Gerelli S, Gioanni S, Pouard P, Vouhé P. The Aristotle Comprehensive Complexity score predicts mortality and morbidity after congenital heart surgery. *Ann Thorac Surg*. 2011;91(4): 1214-1221.

5. 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(5): 1139-1153.

6. 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(2): 564-572.

7. Jacobs JP, Jacobs ML, Maruszewski B, et al. Initial application in the EACTS and STS congenital heart surgery databases of an empirically derived methodology of complexity adjustment to evaluate surgical case mix and results. *Eur J Cardiothorac Surg*. 2012;42(5): 775-780.

8. 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(6): 2184-2192.

9. 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(5): 1653-1659.

10. 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(3): 1054-1062.

11. 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(3): 1063-1070.

12. 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(3): 1071-1077.

13. Jacobs JP, He X, Mayer JE Jr, et al. Mortality trends in pediatric and congenital heart surgery: an analysis of the Society of Thoracic Surgeons Congenital Heart Surgery Database. *Ann Thorac Surg*. 2016;102(4): 1345-1352.

14. 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.

15. Jacobs JP, Mayer JE Jr, Pasquali SK, et al. The Society of Thoracic Surgeons Congenital Heart Surgery Database: 2019 update on outcomes and quality. *Ann Thorac Surg*. 2019;107(3): 691-704.

16. 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(6): 1889-1894.

17. Spray TL, Gaynor WL. A word of caution in public reporting. *Semin Thorac Cardiovasc Surg Pediatr Card Surg Ann*. 2017;20: 49-55.

18. 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(1): 211-219.

19. Bertsimas D., Dunn J. Optimal classification trees. *Mach Learn*. 2017;106: 1039-1082.

20. 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(4): 574-583.

21. Jalali A, Lonsdale H, Do N, et al. Deep learning for improved risk prediction in surgical outcomes. *Sci Rep*. 2020;10(1): 9289-9302.

22. Jalali A, Simpao AF, Galvez JA, Licht DJ, Nataraj C. Prediction of periventricular leukomalacia in neonates after cardiac surgery using machine learning algorithms. *J Med Syst*. 2018;42(10): 177-188.

23. Johan Nilsson J, Ohlsson M, Thulin L, Höglund P, Nashef SAM, Brandt J. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. *J Thorac Cardiovasc Surg*. 2006;132(1): 12-19.

24. Cooper JN, Wei L, Fernandez SA, Minecci PC, Deans KJ. Preoperative prediction of surgical morbidity in children: comparison of five statistical models. *Comput Biol Med*. 2015;57: 54-65.

25. Cao Y, Fang X, Ottosson J, Naslund E, Stenberg E. A comparative study of machine learning algorithms in predicting severe complications after bariatric surgery. *J Clin Med*. 2019;8(5): 668-684.

26. Ahmed Z, Mohamed K, Zeeshan S, Dong XQ. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database*. 2020;1-35.

27. Hyer JM, White S, Cloyd J, et al. Can we improve prediction of adverse surgical outcomes? Development of a surgical complexity score using a novel machine learning technique. *J Am Coll Surg*. 2020;230(1): 43-52.

28. Maruszewski B, Lacour-Gayet F, Monro JL, Keogh BE, Tobota Z, Kansy A. An attempt at data verification at the EACTS congenital database. *Eur J Cardiothorac Surg*. 2005;28(3): 400-406.