# Benchmarking in Congenital Heart Surgery using Machine Learning-derived Optimal Classification Trees

<table>
<thead>
<tr>
<th>Journal:</th>
<th>World Journal for Pediatric and Congenital Heart Surgery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>Draft</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Original Article</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>n/a</td>
</tr>
</tbody>
</table>
| Complete List of Authors: | Bertsimas, Dimitris; Massachusetts Institute of Technology, Sloan School of Management and Operations Research Center
Zhuo, Daisy; Alexandria Health, Levine, Jordan; Alexandria Health
Dunn, Jack; Alexandria Health
Tobota, Zdzislaw; Warsaw University Hospital
Maruszewski, Bohdan; Children's Memorial Health Institute, Department for Paediatric Cardithoracic Surgery
Fragata, Jose; Hospital de Santa Marta,
Sarris, George; Athens Heart Surgery Institute, Congenital Heart Surgery; Iaso Children's Hospital, Pediatric Heart Surgery |
| Keywords: | Congenital heart disease (CHD), Congenital heart surgery, Database (all types), Outcomes (includes mortality, morbidity), Statistics, risk analysis/modeling, Statistics, survival analysis |

## Abstract:

Background. We have previously shown that the machine learning (ML) 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. From the European Congenital Heart Surgeons Association (ECHSA) Congenital Database (ECDB), the data subset describing 31,792 patients who had undergone any of ten "benchmark procedure group" operations as primary procedure was analyzed. OCT models were built predicting hospital mortality (HM) and prolonged post-operative mechanical ventilatory support time (MVST) and hospital length of 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 expected outcomes for each hospital, as if their patients could be treated in this "virtual hospital". The resulting OCTs accurately estimate expected surgical outcomes for each hospital, both aggregate and, importantly, for risk-matched specific patient cohorts.

Results. The raw average rates were HM= 4.4%, MVST= 15.3%, and LOS=15.5%. Of 64 participating centers, comparison with each hospital’s specific case-adjusted benchmark, 17.0% statistically (under 90% Confidence Intervals) overperformed and 26.4% underperformed the virtual hospital regarding mortality. For MVST and LOS,

http://mc.manuscriptcentral.com/wjpchs
overperformers were 34.0% and 26.4%, and underperformers 28.3% and 43.4%, respectively. OCT analyses reveal hospital-specific patient cohorts of either over- or under-performance.

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.
Benchmarking in Congenital Heart Surgery using Machine Learning-derived Optimal Classification Trees

Dimitris Bertsimas, PhD¹, Daisy Zhuo, PhD², Jordan Levine, MEng², Jack Dunn, PhD², Zdzislaw Tobota, MD³, Bohdan Maruszewski MD, PhD³, Jose Fragata MD, PhD⁴, George E Sarris, MD, PhD⁵

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

Keywords: Congenital heart disease (CHD), Congenital heart surgery, Database (all types), Outcomes, Statistics, risk analysis/modeling, Statistics-survival analysis

<table>
<thead>
<tr>
<th>Author</th>
<th>ORCID ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimitris Bertsimas</td>
<td>0000-0002-1985-1003</td>
</tr>
<tr>
<td>Daisy Zhuo</td>
<td>0000-0002-8579-932X</td>
</tr>
<tr>
<td>Jack Dunn</td>
<td>0000-0002-6936-4502</td>
</tr>
<tr>
<td>Jordan Levine</td>
<td>0000-0002-8766-9983</td>
</tr>
<tr>
<td>Zdzislaw Tobota</td>
<td></td>
</tr>
<tr>
<td>Bohdan Maruszewski</td>
<td>0000-00034144-6677</td>
</tr>
<tr>
<td>Jose Fragata</td>
<td>0000-0002-4493-8178</td>
</tr>
<tr>
<td>George E Sarris</td>
<td>0000-0002-9341-2850</td>
</tr>
</tbody>
</table>
Institutional addresses

Massachusetts Institute of Technology Operations Research Center: 1 Amherst St, Cambridge, MA 02142, USA

Massachusetts Institute of Technology Sloan School of Management: 100 Main St, Cambridge, MA 02142, USA

Alexandria Health: 14th Floor, One Broadway, Cambridge, MA 02142 and 225 Dyer St, Providence, RI 02903, USA

Children’s Memorial Health Institute: Aleja Dzieci Polskich 20, 04-730, Warsaw, Poland;

Hospital de Santa Marta: R. De Santa Marta 50,1169-024, Lisbon, Portugal

NOVA University: Campus de Campolide, 1099-085, Lisbon, Portugal

Athens Heart Surgery Institute: Kifissias Avenue #2, Amaroussion, Athens 151 25, Greece

Article word count: 4,972
ABSTRACT

**Background.** We have previously shown that the machine learning (ML) 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.** From the European Congenital Heart Surgeons Association (ECHSA) Congenital Database (ECDB), the data subset describing 31,792 patients who had undergone any of ten “benchmark procedure group” operations as primary procedure was analyzed. OCT models were built predicting hospital mortality (HM) and prolonged post-operative mechanical ventilatory support time (MVST) and hospital length of 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 expected outcomes for each hospital, as if their patients could be treated in this “virtual hospital”. The resulting OCTs accurately estimate expected surgical outcomes for each hospital, both aggregate and, importantly, for risk-matched specific patient cohorts.

**Results.** The raw average rates were HM= 4.4%, MVST= 15.3%, and LOS=15.5%. Of 64 participating centers, comparison with each hospital’s specific case-adjusted benchmark, 17.0% statistically (under 90% Confidence Intervals) overperformed and 26.4% underperformed the virtual hospital regarding mortality. For MVST and LOS, overperformers were 34.0% and 26.4%, and underperformers 28.3% and 43.4%, respectively. OCT analyses reveal hospital-specific patient cohorts of either over- or under-performance.

**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.
**Abbreviations**

AI: Artificial Intelligence

AUC: Area Under the Curve (or c-statistic)

CHD: Congenital Heart Disease

CHS: Congenital Heart Surgery

CHSD: Congenital Heart Surgery Database

CPB: Cardiopulmonary Bypass

ECHSA: European Congenital Heart Surgeons Association

ECDB: ECHSA Congenital Database

GPRF: General preoperative risk factor

LOS: Length of Hospital Stay

ML: Machine Learning

MVST: Postoperative Mechanical Ventilatory Support Time

OCT: Optimal Classification Trees
Introduction

Quality improvement efforts in CHS depend on 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 which may affect such risk. Adjustment for variability in outcome attributed to the inherent risk of different procedures and to different patient characteristics has evolved from initial attempts of risk stratification based on empiric expert consensus (RACHS-1 and Aristotle methods), to those based on outcome measures provided by real data (STAT Mortality Score, recently updated, STS Morbidity Score). Risk prediction models have been developed and refined (STS CHSD mortality and the UK PRAiS2 Models), comparing results of CHS centers, while adjusting for differences in both case mix (risk stratification) and in patient-specific factors, achieving remarkable accuracy (AUC’s in the range of 0.852 to 0.875). Nevertheless, these approaches, which aimed to provide comparative performance data to the public, have received criticism for both data limitations (e.g., 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 centers for some low-volume complex procedures, or may do injustice when applied to centers performing rare or even unique and high-risk procedures. 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.

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, i.e., 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 risk model in which intuitively pre-selected feature interactions have been considered.
We have previously shown\textsuperscript{16} that the artificial intelligence (AI) - machine learning (ML)-based non-linear methodology of Optimal Classification Trees (OCTs) can be used to accurately and interpretably predict risk after CHS even at individual patient level, taking into account all relevant recorded risk factors, automatically identifying important ones without any \textit{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 of which are performed quite infrequently even in large centers, we limited this analysis to the ten so called “benchmark procedure groups”, a practice also adopted by the STS CHSD\textsuperscript{4,10}.

\textbf{Material and Methods}

\textbf{The data}, provided by the 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. 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 ten “benchmark procedure groups” (see TABLE 1) from 1/1/2010 to 12/31/2018. 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”, i.e., 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 was applied. The risk factors entered in the model, all being preoperative features, are shown in TABLE 2. The risk of postoperative adverse outcomes previously defined (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. 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.

Results.

a) 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 two 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 four pathways leading to four distinct cohorts:

Cohort 1 (Mortality Risk: 7.5%): This cohort underwent AV Septal Defect Repair, Arterial Switch Operation, Arterial Switch Operation/VSD Repair, Off-Bypass Coarctation Repair, Tetralogy of Fallot Repair, and VSD Repair, and it has been <1,569 days since the previous admission.

Cohort 2 (Mortality Risk: 2.0%): This cohort underwent the same set of
procedures as Cohort 1, but has never had a previous admission, or it has been at least 1,569 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 one of the following 3 procedure groups: Fontan, Glenn Hemi-Fontan, and Truncus Repair.

In this non-linear 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 one 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 one and only one terminal cohort.

Furthermore, 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 high risk (Cohort 3); some have low risk (Cohort 2). It is evident that some nodes have a high percentage of patients (Cohort 2); some do not (Cohort 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.

**b) 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 the virtual hospital’s performance environment: This is
a two-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. The calculation of these metrics is illustrated with an example 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 divide 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", i.e., cohorts for which the observed index hospital’s measured performance is statistically better than that of the same patient cohort’s calculated risk in the virtual hospital (“expected rate”), and, second, "areas of opportunity", i.e., cohorts of greater observed risk in the index hospital compared to the “expected rate”, i.e., 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 direct comparison of identical terminal nodes of the pathways of the index and virtual hospitals, i.e., **direct comparison of outcomes of patient cohorts with similar risk characteristics**.

c) 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 3A, 3B, and 3C for mortality, prolonged MSVT, and prolonged LOS, respectively. In Figure 4, the **observed vs expected mortality (O/E) ratio** of all hospitals is shown, sorted from lowest to highest. For mortality,
17.0% of hospitals perform better and 26.4% worse than expected. For 11.3% of hospitals, the \textbf{O/E mortality} ratio is above 2. We observe 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 1,254 patients included in this study. The summary assessment of this hospital is shown in \textbf{TABLE 3}, with \textbf{the actual observed versus the expected values} for HM, prolonged MVST, and prolonged LOS rates being 7.2% vs 7.2% \textbf{(p>0.05)}, 22.7% vs 19.7% \textbf{(p=0.011)}, and 25.7% vs 21.6% \textbf{(p=0.002)}.

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 \textbf{Figure 5}, comparison of the risk stratified cohorts of Hospital A and the virtual hospital, it is evident Hospital A’s case-mix is higher risk. Accordingly, when we calculate Hospital A’s \textit{expected rate}, which is, 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 performance is not statistically significantly different from the virtual hospital. Indeed, in \textbf{Figure 4}, where the O/E ratio of all hospitals is shown, O/E for Hospital A is 1.0, the error bar indicating 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 over-performance (green) or under-performance (red), compared to expected, the case adjusted benchmark. Despite overall performance as expected, this full tree (\textbf{Figure 6}), 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 opportunity for improvement (worse than expected performance). \textbf{Figure 7A} shows the characteristics of these cohorts, and \textbf{Figure 7B}
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 non-linear Artificial Intelligence (AI) – Machine Learning (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, in order 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 preoperative factors (e.g., diagnosis, age, weight, prematurity, history of prior cardiac procedures, presence of other non-cardiac anomalies such as genetic and chromosomal defects), and other preoperative factors indicating clinical status (e.g., 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.

It should be noted that, alternatively, we could assess the reverse, i.e., 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, e.g., 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 similar risk. The criterion for assigning a patient to a cohort is the combination of shared characteristics and statistically similar risk 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
sub-cohorts 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 7B). Therefore, a detailed risk-adjusted view of the participating hospital is provided, analyzed both by risk group and by 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 over- 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 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 which 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), yet these are generally not clinically tracked. We also know that there are specific features relevant only to
certain procedures (e.g., coronary anatomy for transposition of 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 achievement of even more accurate predictions in the future. Furthermore, we emphasize that our models reflect the current status of the field. On-going 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.

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

**Funding:** The author(s) received no financial support for the research, authorship, and/or publication of this article. **Disclosures:** Conflicts of interest: none
References


583.

# TABLE 1. The ten benchmark procedures studied.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of procedures</td>
<td>Mortality</td>
</tr>
<tr>
<td>Off Bypass Coarctation Repair</td>
<td>3,519</td>
<td>2.2%</td>
</tr>
<tr>
<td>Fontan Procedure</td>
<td>1,264</td>
<td>4.7%</td>
</tr>
<tr>
<td>Glenn or Hemi-Fontan Procedure</td>
<td>832</td>
<td>7.5%</td>
</tr>
<tr>
<td>Arterial Switch Operation with VSD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repair</td>
<td>895</td>
<td>8.2%</td>
</tr>
<tr>
<td>Arterial Switch Operation</td>
<td>2,175</td>
<td>4.0%</td>
</tr>
<tr>
<td>Complete Atrioventricular Canal Repair</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repair</td>
<td>2,157</td>
<td>4.7%</td>
</tr>
<tr>
<td>Tetralogy of Fallot Repair</td>
<td>728</td>
<td>1.5%</td>
</tr>
<tr>
<td>VSD Repair</td>
<td>7,257</td>
<td>0.9%</td>
</tr>
<tr>
<td>Norwood Procedure</td>
<td>1,288</td>
<td>31.4%</td>
</tr>
<tr>
<td>Truncus Repair</td>
<td>3,265</td>
<td>4.0%</td>
</tr>
<tr>
<td>Overall</td>
<td>23,380</td>
<td>4.6%</td>
</tr>
</tbody>
</table>
TABLE 2. Preoperative risk features (potential risk factors) analyzed and the features’ relative importance for mortality prediction

<table>
<thead>
<tr>
<th>Predictive Variable</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedure</td>
<td>71.0%</td>
</tr>
<tr>
<td>Days since previous admission, if any</td>
<td>11.9%</td>
</tr>
<tr>
<td>Weight</td>
<td>9.2%</td>
</tr>
<tr>
<td>Age (months)</td>
<td>3.7%</td>
</tr>
<tr>
<td>Number of preoperative diagnoses</td>
<td>3.1%</td>
</tr>
<tr>
<td>Any general preoperative risk factor present</td>
<td>1.0%</td>
</tr>
<tr>
<td>Case category (CPB vs non-CPB)</td>
<td>0.0%</td>
</tr>
<tr>
<td>Gender</td>
<td>0.0%</td>
</tr>
<tr>
<td>Number of concomitant procedures performed</td>
<td>0.0%</td>
</tr>
<tr>
<td>Any non-cardiac abnormality present</td>
<td>0.0%</td>
</tr>
<tr>
<td>Any prior admission</td>
<td>0.0%</td>
</tr>
<tr>
<td>Year of procedure</td>
<td>0.0%</td>
</tr>
</tbody>
</table>
TABLE 3: Example “Hospital A” Performance Summary

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Actual Rate</th>
<th>Average Rate of Virtual Hospital</th>
<th>Expected Rate</th>
<th>Performance difference referenced to expected rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortality</td>
<td>7.2%</td>
<td>4.4%</td>
<td>7.2%</td>
<td>Not statistically significant (p=1.0)</td>
</tr>
<tr>
<td>Prolonged</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MVST</td>
<td>22.7%</td>
<td>15.3%</td>
<td>19.7%</td>
<td>Worse (p=0.011)</td>
</tr>
<tr>
<td>Prolonged LOS</td>
<td>25.7%</td>
<td>15.5%</td>
<td>21.6%</td>
<td>Worse (p=0.002)</td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

Figure 1:
The mortality model OCT tree for the virtual hospital (A), with a zoomed-in version of the top two 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 two, 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 2,801 procedures in 4 categories, with mortality 19.8%. The left branch is further split based on the variable *days since the previous admission 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 cohort 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.

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 risk, higher risk 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).
**Figure 3A:** Comparison of mortality for all hospitals. The observed over expected (O/E) mortality ratio (y-axis) is plotted against the expected mortality for each hospital. Dot size reflects hospital size (i.e., number of patients). The vertical solid line represents the mortality of the virtual hospital. The horizontal line corresponds to *mortality of hospital equals mortality of virtual hospital*. These two lines divide the x-y plane into 4 quadrants: "easy cases" and underperforming hospitals, "easy cases" and overperforming hospitals, "hard cases" and underperforming hospitals, and "hard cases" and overperforming hospitals. Red indicates statistically significant underperformance, green indicates overperformance, and white no statistically significant performance difference.

**Figure 3B:** Comparison of prolonged MVST for all hospitals.

**Figure 3C:** Comparison of prolonged LOS for all hospitals.

**Figure 4A:** 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 O/E ratio. The red dotted line is at ratio 1.0 for observed equal to expected mortality. The O/E ratio of the example hospital, highlighted in red, is at unity, no different from expected.

**Figure 4B:** The distribution of hospitals by the O/E ratios, with the ratio of 1.0 for no difference, in red. Of all hospitals, 11.3% have O/E ratios above 2.

**Figure 5:** 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 (i.e., expected mortality rate) of 7.1% compared to the virtual hospital’s 4.4%, hence it is seeing overall “harder” patients.
**Figure 6:** 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.

**Figure 7A:** Characteristics of cohorts with higher than expected mortality.

**Figure 7B:** Cohort summary for mortality, showing performance for each cohort broken down by benchmark procedure.
The mortality model OCT tree for the virtual hospital (A), with a zoomed-in version of the top two 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 two, 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 2,801 procedures in 4 categories, with mortality 19.8%. The left branch is further split based on the variable days since the previous admission 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 cohort 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 mortality model OCT tree for the virtual hospital (A), with a zoomed-in version of the top two 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 two, 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 2,801 procedures in 4 categories, with mortality 19.8%. The left branch is further split based on the variable days since the previous admission 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 cohort 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.
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 risk, higher risk 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).

211x123mm (150 x 150 DPI)
Comparison of mortality for all hospitals. The observed over expected (O/E) mortality ratio (y-axis) is plotted against the expected mortality for each hospital. Dot size reflects hospital size (i.e., number of patients). The vertical solid line represents the mortality of the virtual hospital. The horizontal line corresponds to mortality of hospital equals mortality of virtual hospital. These two lines divide the x-y plane into 4 quadrants: "easy cases" and underperforming hospitals, "easy cases" and overperforming hospitals, "hard cases" and underperforming hospitals, and "hard cases" and overperforming hospitals. Red indicates statistically significant underperformance, green indicates overperformance, and white no statistically significant performance difference.
Comparison of prolonged MVST for all hospitals.

211x211mm (150 x 150 DPI)
Comparison of prolonged LOS for all hospitals.

211x211mm (150 x 150 DPI)
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 O/E ratio. The red dotted line is at ratio 1.0 for observed equal to expected mortality. The O/E ratio of the example hospital, highlighted in red, is at unity, no different from expected.

211x141mm (150 x 150 DPI)
The distribution of hospitals by the O/E ratios, with the ratio of 1.0 for no difference, in red. Of all hospitals, 11.3% have O/E ratios above 2.
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 (i.e., expected mortality rate) of 7.1% compared to the virtual hospital’s 4.4%, hence it is seeing overall “harder” patients.
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.
Characteristics of cohorts with higher than expected mortality.

189x132mm (150 x 150 DPI)

<table>
<thead>
<tr>
<th>Cohort 3</th>
<th>Virtual Hospital Mortality 1.0%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>This Hospital Mortality 3.2%</td>
</tr>
<tr>
<td></td>
<td>(p = 0.085)</td>
</tr>
</tbody>
</table>

1) Procedure is AVC (AV/Canal 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 ≥ 1669.5 or no previous operation
3) Age (months) ≤ 0.3167
4) Weight ≥ 3.005 or missing
5) Number of diagnosis ≤ 1.5

<table>
<thead>
<tr>
<th>Cohort 5</th>
<th>Virtual Hospital Mortality 2.6%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>This Hospital Mortality 6.2%</td>
</tr>
<tr>
<td></td>
<td>(p = ≤ 0.001)</td>
</tr>
</tbody>
</table>

1) Days since previous admission ≥ 1669.5 or no previous operation
2) Age (months) ≥ 0.3167
3) Procedure is AVC (AV/Canal or AVSeptal Defect Repair), Arterial Switch Op, Arterial Switch Op with VSD Repair, Tetralogy of Fallot Repair

<table>
<thead>
<tr>
<th>Cohort 8</th>
<th>Virtual Hospital Mortality 8.7%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>This Hospital Mortality 15.8%</td>
</tr>
<tr>
<td></td>
<td>(p = 0.009)</td>
</tr>
</tbody>
</table>

1) Procedure is Fontan, Glenn Hemi-Fontan, Truncus Repair
Cohort summary for mortality, showing performance for each cohort broken down by benchmark procedure.

524x273mm (150 x 150 DPI)