

# Prediction of Systemic-to-Pulmonary Artery Shunt Surgery Outcomes Using Administrative Data

Sara Moein<sup>1</sup>, Sanmay Das<sup>2</sup>, Matthew Hall<sup>3</sup>, Hao Yan<sup>2</sup>, Pirooz Egtesady<sup>1</sup>

<sup>1</sup>Division of Pediatric Cardiothoracic Surgery, Washington University School of Medicine, MO, USA

<sup>2</sup>Department of Computer Science and Engineering, Washington University in St. Louis, MO, USA

<sup>3</sup>Children's Hospital Association, Overland Park, KS, USA

Corresponding author: egtesadyp@wudosis.wustl.edu

**Abstract**—Systemic-to-Pulmonary Artery (SPA) shunt surgery, one of the most common cardiac surgical procedures in the newborn period, provides a means to palliate children with limited pulmonary blood flow, such as in Tetralogy of Fallot. Despite the simplicity of the procedure, it is associated with significant morbidity (such as need for extracorporeal membrane oxygenation (ECMO), and long post-operative length of stay (PLOS) in the hospital following surgery) and mortality. These outcomes are known to be impacted by a number of complex factors (including patient specific and procedure specific factors, perioperative related factors, etc.), whose relative importance in clinical decision making remains the domain of clinical judgment. The increasing availability of multi-modal data on patient care and outcomes opens up the opportunity to assess clinical practices from a more data-driven perspective. In this paper, we report results from a study of 1036 patients (from 44 children's hospitals across the US) during 2009-2014 that applies a machine learning approach to predicting post-operative outcomes for patients in the Pediatric Health Information System (PHIS) database. We demonstrate that it is feasible to achieve significant prediction benefits using a standard machine learning approach (random forests) on a carefully constructed dataset, showing the value of applying machine learning even with noisy administrative databases. The methods we describe can be used to identify potential important variables that lead to good clinical judgment as defined by desirable clinical outcomes.

**Keywords**—Systemic-to-Pulmonary Artery (SPA) Shunt Surgery; Prediction; Feature Importance; Random Forest (RF).

## I. INTRODUCTION

In the US, approximately 40,000 infants are born each year with a congenital heart defect [1]. One third of babies born with congenital heart disease are estimated to require intervention within the first year of their life. Management of these infants and children can be associated with significant morbidity and mortality. This results in growing interest in improving the quality of pediatric cardiac care [9].

The Systemic-to-Pulmonary Artery (SPA) shunt is a common procedure, often performed in the neonatal period, used to establish a reliable source of pulmonary blood flow in babies with limited or absent pulmonary blood flow [8, 15]. It is also often an integral part of staged palliative procedures for selected neonates and infants with complex heart defects.

However, despite the relatively “straightforward” nature of the procedure, often not requiring open-heart or use of heart-lung machine as in other pediatric cardiac procedures, the SPA shunt surgery can be associated with significant morbidity and mortality. Frequently, post-surgical intervention or extracorporeal membrane oxygenation (ECMO) is needed as a form of rescue. Infants who end up on ECMO may go on to develop severe neurologic insult(s) despite survival, highlighting the potential morbidity of when ECMO is needed. In addition, some neonates have long post-operative length of stay (PLOS) after SPA shunt, which is undesirable, since this can not only contribute to excess cost but also unanticipated complications (e.g., hospital acquired infections). Therefore, long PLOS, mortality and need for ECMO, can serve as reasonable surrogate markers of poor outcomes for patients who undergo surgical interventions, including neonates that undergo SPA shunt surgery [12].

Our goal in this paper is to understand how a number of features (such as use of certain pharmaceuticals, procedural, and other clinical variables) derived from the large Pediatric Health Information System (PHIS) database are associated with outcomes following SPA shunt surgery. Our long term goal is to apply the knowledge gained as a prediction tool for patient management algorithms. In this manuscript, we employ Random Forest (RF) algorithm [3], which has various desirable properties for a task such as this including robustness to overfitting and a natural method for providing feature importance [7, 11, 19], in addition to the well-known high empirical prediction accuracy [4].

This paper is organized as follows: In Section II we briefly review the related literature. In section III, we provide an overview of RF. Section IV details our methodology, and Section V describes our experimental results.

## II. RELATED WORK

Machine Learning (ML) algorithms are being increasingly applied to large medical datasets as an approach to identifying factors that can predict outcomes. They have been shown to improve selection of the most important clinical decision making predictors in the medical arena. In the study by Schnieder et. al [22] in 2014, the Naive Bayesian Network

(NBN) and multiple linear regression were used for diagnosis of hyperparathyroidism, and the authors compared the performance of two approaches. They demonstrated that the NBN had better performance than traditional regression. In another study in 2014 [17], Artificial Neural Network (ANN), Bayesian Network (BN) and Logistic Regression (LR) were used to determine the predictors of post-operative delirium in cardiac surgery patients. The authors successfully demonstrated the use of ANN and BN in prediction of delirium. Jalali et. al (2014) [13], developed a decision tree for prediction of Periventricular Leukomalacia (PVL) occurrence in neonates after pediatric heart surgery. Nilsson et. al (2006) [18] used ANN to identify risk factors in cardiac surgery that predict mortality. In 2001, DeGroff et. al [6] used ANN for classification of heart sound data into innocent and pathological classes.

Random Forests specifically have been applied in various bioinformatics studies such as large-scale association studies for complex genetic diseases. Lunetta et al. [15] and Bureau et al. [4] focused on Single Nucleotide Polymorphisms (SNPs). They detected SNP-SNP interactions in the case-control context by computing a random forest variable importance measure for each polymorphism. Diaz-Uriate and Alvarez de Andrés [7] compared the performance of RF and other classification approaches for the analysis of gene expression data. In another study [10], RF was applied to microarray data to match the physiological effect of a novel drug at the cellular level with its clinical relevance. Shi and colleagues [24] used RF for tumor classification after tissue microarray profiling.

Despite the large number of studies in applying ML algorithms in computational biology and bioinformatics, relatively little has been done in some highly complex disciplines, such as pediatric heart surgery, where indeed, patients can present with 40 to 50 different diagnoses, each with subtle variations and associated with additional comorbidities (such as chromosomal problems, other non-cardiac congenital anomalies, prematurity, weight, sex, etc.) that have previously been shown to impact outcomes. The diverse array of various treatment options (surgical, catheter based, medical, etc.) further complicates this milieu. In addition, previous studies that used the hospital administrative data for their analysis have chiefly focused on adults [2, 16, 21, 25], with few studies including data for children [23, 29] and have generally found administrative data to be an inaccurate source of information when compared with clinical data [20]. This study uses RF to overcome the limitations of administrative data.

### III. METHODOLOGY

The prediction process of SPA shunt surgery outcomes is described in detail in this section, including description of the learning algorithm, our pilot dataset, feature extraction, and the experimental parameters.

#### A. RANDOM FORESTS

RF is a collection of classification or regression trees that are fully-grown. A data point's class is predicted by averaging the predictions of individual trees. Each tree is fit from a bootstrapped training set (i.e. sampled with replacement) of the same size ( $N$ ) as the entire training set. While each tree is being grown, for each split, only some random subset of the features (of size  $\sqrt{m}$  where  $m$  is the total number of features) are considered for the split. We choose the split that maximizes the Gini index. Since we are dealing with a highly imbalanced dataset (29% poor outcomes, 71% good outcomes), we reweight the examples of poor outcomes to have 3 times the weight of examples with good outcomes.

Each bootstrap sample includes, on average, 2/3 of the original data points [3]. For each tree, a subset of the entire training data is not used in the construction of that tree. The out-of-bag (OOB) error is estimated by predicting each training data point using only those trees in the ensemble for which that example was out-of-bag in the construction. In the RF, each tree is a weak classifier and combining the trees by averaging the predictions of many weak classifiers results in significant accuracy improvement compared to a single tree [5]. In other words, since the unpruned trees are low-bias, high-variance models, averaging over an ensemble of trees reduces variance while keeping bias low [26]. RFs have previously had significant success in biomedical applications [19, 27].

#### B. DATASET

This study is a retrospective, observational cohort study using the PHIS database (Children's Hospital Association, Overland Park, KS). The PHIS database contains de-identified administrative data, detailing demographic, diagnostic, procedures, and daily billing data from 47 freestanding, tertiary care children's hospitals. This database accounts for approximately 20% of all annual pediatric hospitalizations in the U.S. Data quality is ensured through a joint effort between the Children's Hospital Association and participating hospitals. The protocol was reviewed by the Institutional Review Board of Washington University School of Medicine, Saint Louis, MO, and was determined to be exempt.

The data for 1036 neonates with SPA shunt surgery were collected from 44 children's hospitals in the US (2009 to 2014). For each patient, four types of features were extracted for each day of their hospitalization. The four types were pharmaceutical, laboratory, supply, and imaging features, numbering 326, 310, 345 and 108, respectively. A summary description of the features is shown in Table I. The dichotomous outcome was defined as: positive outcome (+1 value) for neonates who underwent ECMO or had mortality or long PLOS, and negative outcome (0 value) for neonates who had none of the preceding and were discharged home after SPA shunt surgery within 35 days. The 75th percentile (35 days) for the length of stay was used to determine the threshold of long PLOS. Note that, with this convention, for the learning algorithm a "positive" outcome is actually a bad

outcome for the patient, and a negative outcome is a good outcome for the patient.

TABLE I. SUMMARY DESCRIPTION OF FEATURES

Feature type	Number of features	Examples
Pharmaceutical	326	Meperidine HCl, Methadone HCl Morphine sulfate Oxycodone HCl
Laboratory	310	Ketone bodies Lactic acid Citric acid Pyruvic acid
Supply	345	Aspirator suction unit Other respiratory drainage supply Stomach tube for suction and drainage Decompression pump
Imaging	108	Whole body SPECT Brain ultrasound Brain ultrasound real-time Brain ultrasound B-mode
<b>Total number of features</b>	1089	

### C. FEATURE EXTRACTION

For each neonate in the dataset, the pharmaceutical, laboratory, supply and imaging files were used to count the frequency of applying each service and the value of the related features obtained. The mentioned files contain various services related to pharmacy, laboratory, imaging and supply that are used for each patient based on the day of admission. There are some other features describing the clinical profile of the neonates including the gestational age, birth weight, non-heart congenital anomalies, chromosomal anomalies and sex. The last 3 features are categorical features with “yes” or “no” values to show if the neonate has the anomaly or not and if the neonate is male or female. Among 1036 patients, there are 224 patients with at least one non-heart congenital anomaly, for example an anomaly in the nervous or respiratory system or an anomaly in the eye or ear. Also, 205 of the patients have chromosomal anomalies. In a fair test, to count the feature value in the dataset, we calculate the summation of values from the day of surgery until  $k$  days ( $k$  is a constant  $(0,1,2,3,...)$ ) after SPA shunt surgery. All the patients who died or went on ECMO prior to  $k$  days after SPA shunt are excluded from the dataset. This prevents the occurrence of ECMO itself from influencing the prediction algorithm (e.g., the day that the patient goes on ECMO, invariably, heparin, an

anticoagulant is given; this, however, does not mean that heparin leads to the need for ECMO).

### D. RF FOR PREDICTING THE OUTCOMES OF SPA SHUNT SURGERY

We performed three sets of experiments. For the first experiment, we used an ensemble of 100 trees to predict outcomes based on data through 1, 2, and 3 days after surgery respectively. Each of these datasets is different both because some of the training examples are excluded (patients who had ECMO or died in the intervening days are excluded) and because the time averages of the features changes based on the drugs and procedures administered. For the second experiment, we tried different numbers of trees in the ensemble (100, 200 and 400), using the “day 3” data from the first experiment. In the third experiment, we included patient features including gestational age, birth weight, chromosomal anomalies, non-heart congenital anomalies and sex in addition to the feature set described above.

## IV. RESULTS AND DISCUSSION

Among 1036 patients, 24 underwent ECMO on the day of SPA shunt surgery. The distribution of the initial day after surgery that neonates went on ECMO is presented in Fig. 1.

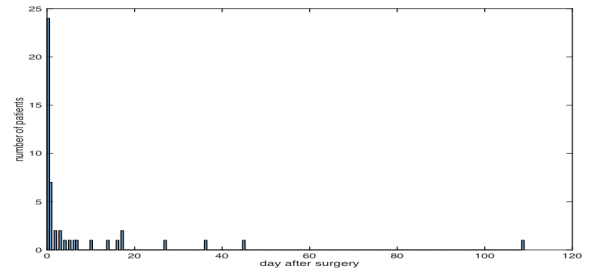


Fig. 1. Distribution of the initial day after surgery that neonates went on ECMO

Using 100 trees in the RF for day 1, 2 and 3 after the SPA shunt provides the Area Under the Curve (AUC) of the receiver operating characteristic (ROC) curves shown in Table II. The results demonstrate that there is some improvement in using data through day 3, despite the loss of several of the (already few) examples of poor outcomes to train on. Therefore, there is value to knowing the histories of patients for the first few days after surgery. For the rest of the experiments, therefore, we focus on the datasets built using 3 days of data after the day of surgery. Table III demonstrates that using 400 trees does offer some prediction benefit over using 100, as can also be seen from the ROC curves (Fig. 2).

Finally, our third experiment was to add clinical features including gestational age, birth weight, chromosomal anomalies, non-heart congenital anomalies and sex of the neonates to the set of 1089 features. The AUC increased to 0.7432 when RF has 400 trees. Table IV and Fig. 3 present the corresponding AUC and ROC curves. This final result demonstrates that one can achieve significant predictive power, with close to 75% chance of ranking a random pair of examples correctly in terms of which of the pairs are more likely to have a poor outcome after SPA shunt surgery.

TABLE II. THE AUC BASED ON DAYS AFTER BT SURGERY; DAY 3 PROVIDES THE AUC MORE THAN DAY 2 AND DAY 1

Day after BT shunt	AUC
1 day	0.6891
2 days	0.6718
3 days	0.7115

TABLE III. THE AUC OF CURVES IN DAY 3 WITH VARIOUS RANDOM NUMBER OF TREES IN RF

Number of trees	AUC
100	0.7115
200	0.7172
400	0.7320

TABLE IV. THE AUC OF THE RF WITH RANDOM NUMBER OF TREES WHEN CLINICAL FEATURES OF NEONATES ARE ADDED TO THE FEATURE SET

Number of trees	AUC
100	0.7267
200	0.7381
400	0.7432

## V. CONCLUSION

There have been a number of reports that have criticized the accuracy and power of administrative data for clinical decision making, especially when compared to clinical data. This paper serves as a proof of concept that ML algorithms may overcome some of these limitations. We have shown that by applying completely standard, off-the-shelf machine learning methods to a carefully constructed dataset with dichotomous “good” and “poor” outcomes we can achieve

good predictive power using such data. The important aspect of our results are not the performance numbers themselves, but the fact that we can get some “lift” in predicting poor outcomes using administrative data. The RFs can then be used to identify critical features of clinical practice that are important in determining patient outcomes. This demonstrates that ML algorithms may have the potential to unravel or flatten critical decisions in a complex system, allowing one to focus on a more limited dataset for deeper investigation.

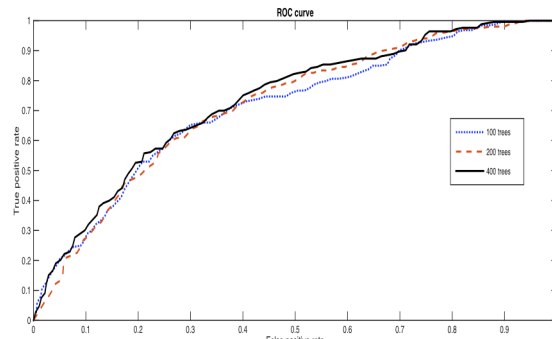


Fig. 2. The ROC curves of RF with random number of trees from day of surgery until day 3

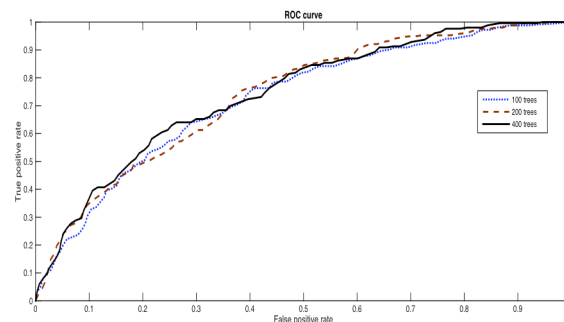


Fig. 3. The ROC curves of RF with 100, 200 and 400 trees when clinical features of the patients are added to the feature set; the curve with 400 trees is above the curves with 100 and 200 trees

## ACKNOWLEDGMENT

This work is supported by University Research Strategic Alliance program at Washington University in St. Louis and Children's Surgical Sciences Research Institute in St. Louis Children's Hospital.

## REFERENCES

- [1] American Heart Association, Understand your risk for congenital heart defects, [www.heart.org/HEARTORG](http://www.heart.org/HEARTORG), retrieved June, 2015
- [2] D. Aronsky, P. J. Haug, C. Lagor, N. C. Dean, “Accuracy of administrative data for identifying patients with pneumonia” , *Am J Med Qual*, vol. 20(6), pp. 319-328, 2005
- [3] L. Breiman, Random forests, *Machine Learning*, (2001) 45, pp. 5-32; <http://www.sagenet.org>.

- [4] A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, P. V. Eerdewegh, "Identifying SNPs predictive of phenotype using random forests", *Genetic Epidemiology*, vol. 28, pp. 171-182, 2005.
- [5] R. Caruana, A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms", In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2006.
- [6] C. G. DeGroof, S. Bhatkar, J. Hertzberg, R. Shandas, L. Valdes-Cruz, R. L. Mahajan, "Artificial neural network-based method of screening heart murmurs in children", *Circulation*, vol. 103, pp. 2711-2716, 2001.
- [7] R. Diaz-Uriarte, S. Alvarez, "Gene selection and classification of microarray data using random forest", *BMC Bioinformatics*, vol. 7, 2006.
- [8] R. M. Freedom, S. J. Yoo, J. Russell, D. Perrin, W.G. Williams, "Designing therapeutic strategies for patients with a dominant left ventricle, discordant ventriculo-arterial connections and unobstructed flow of blood to the lungs", *Cardiol Young*, vol. 14, pp. 630-653, 2004.
- [9] R. F. Gillum, "Epidemiology of congenital heart disease in the United States", *Am Heart J*, vol. 127, pp. 919-927, 1994.
- [10] E. C. Gunther, D. J. Stone, R. W. Gerwien, P. Bento, M. P. Heyes, "Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro, Proceedings of the National Academy of Sciences", vol. 100, pp. 9608-9613, 2003.
- [11] I. Guyon, A. Elisseeff, "An introduction to variable and feature selection", *The Journal of Machine Learning Research*, vol. 3, 2003.
- [12] J. Hill, "Adult emergency cardiopulmonary support systems", In: *ECMO - Extracorporeal cardiopulmonary support in critical care*, J. B. Zwischenberger, R. H. Bartlett, Eds. Extracorporeal Life Support Organization, 1995, Chapter 28, pp. 491-510.
- [13] A. Jalali, E. M. Buckley, J. M. Lynch, P. J. Schwab, D. J. Licht, C. Nataraj, "Prediction of periventricular leukomalacia occurrence in neonates after heart surgery", *IEEE Journal of Biomedical and Health Informatics*, vol. 18(4), pp. 1453-1460, 2014.
- [14] P. Lang, W. I. Norwood, "Hemodynamic assessment after palliative surgery for hypoplastic left heart syndrome", *Circulation*, vol. 68, pp. 104-108, 1983.
- [15] K. L. Lunetta, L. B. Hayward, J. Segal, P. V. Eerdewegh, "Screening Large-Scale Association Study Data: Exploiting Interactions Using Random Forests", *BMC Genetics*, vol. 5 (32), 2004.
- [16] J. Meddings, S. Saint, L. F. Jr. McMahon, Hospital-acquired catheter-associated urinary tract infection, "documentation and coding issues may reduce financial impact of Medicare's new payment policy", *Infect Control Hosp Epidemiol*, vol. 31(6), pp. 627-633, 2010.
- [17] H. N. Mufti, S. Abidi, S. R. Abidi, G. M. Hirsch, "Predictors of post-operative delirium in cardiac surgery patients", a machine learning approach, *Canadian Journal of Cardiology*, vol. 30, pp. 237-238, 2014.
- [18] J. Nilsson, M. Ohlsson, L. Thulin, P. Hoglund, S. A. Nashef, J. Brandt, "Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks", *J Thorac Cardiovasc Surg*, vol. 132, pp. 12-19, 2006.
- [19] O. Okun, H. Priisalu, "Random forest for gene expression based cancer classification: overlooked issues", *Lecture Notes in Computer Science*, vol. 4478, pp. 483-490, 2007.
- [20] S. W. Patrick, M. M. Davis, A. B. Sedman, J. A. Meddings, S. Hieber, G. M. Lee, et al., "Accuracy of hospital administrative data in reporting central line-associated bloodstream infections in newborns", *Pediatrics*, vol. 131(Suppl 1), pp. S75-S80, 2013.
- [21] M. K. Schaefer, K. Ellingson, C. Conover, A. E. Genisca, D. Currie, T. Esposito, et al., "Evaluation of International Classification of Diseases", Ninth Revision, Clinical Modification Codes for reporting methicillin-resistant *Staphylococcus aureus* infections at a hospital in Illinois, *Infect Control Hosp Epidemiol*, vol. 31(5), pp. 463-468, 2010.
- [22] D. F. Schneider, M. Craven, P. E. Sonderman, S. A. Weber, R. S. Sippel, C. C. Greenberg, H. Chen, "Machine learning is superior to traditional modeling: a parathyroid hormone calculator example", *Journal of American College of Surgeons*, vol. 219, pp. 146-147, 2014.
- [23] E. R. Sherman, K. H. Heydon, S. K. H. John, E. Teszner, S.L. Retting, S. k. Alexander, et al., "Administrative data fail to accurately identify cases of healthcare-associated infection", *Infect Control Hosp Epidemiol*, vol. 27(4), pp. 332-337, 2006.
- [24] T. Shi, D. Seligson, A. S. Belldegrun, A. Palotie, S. Horvath, "Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma", *Modern Pathology*, vol. 18, pp. 547-557, 2005.
- [25] K. B. Stevenson, Y. Khan, J. Dickman, T. Gillenwater, P. Kulich, C. Myers, et al., "Administrative coding data, compared with CDC/NHSN criteria, are poor indicators of health care-associated infections", *Am J Infect Control*, vol. 36(3), pp. 155-164, 2008.
- [26] C. Strobl, A. L. Boulesteix, A. Zeileis, T. Hothorn, "Bias in random forest variable importance measures: illustrations, sources and a solution", *BMC Bioinformatics*, vol. 8 (25), 2007.
- [27] V. Svetnik, A. Liaw, C. Tong, T. Wang, "Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules", In: F. Roli, J. Kittler, T. Windeatt, Eds. *LNC3*, Springer, Heidelberg, 2004, pp. 334-343.
- [28] J. S. Tieder, M. Hall, K. A. Auger, P. D. Hain, K. E. Jerardi, A. E. Myers, et al., "Accuracy of administrative billing codes to detect urinary tract infection hospitalizations", *Pediatrics*, vol. 128(2), pp. 323-330, 2011.
- [29] S. B. Wright, W. C. Huskins, R. S. Dokholyan, D. A. Goldmann, R. Platt, "Administrative databases provide inaccurate data for surveillance of long-term central venous catheter-associated infections", *Infect Control Hosp Epidemiol*, vol. 24(12), pp. 946-949, 2003.