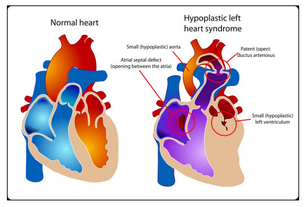
Improving Prediction of Single Ventricle Congenital Heart Disease with Free Text

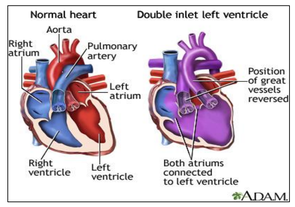
Devin Koehl, Tobias O'Leary
The University of Alabama at Birmingham

Pediatric Heart Transplant Society

- The Pediatric Heart Transplant Society (PHTS) was established in 1993 by a group of physicians who wanted to improve the lives of children who needed a heart transplant.
- The PHTS registry is a national longitudinal observational study enrolling children when they are listed for a heart transplant.
- Patients are followed in the registry while on the waitlist and after transplantation
- PHTS has 56 participating centers and over 9000 listed patients

Single Ventricle Heart Disease



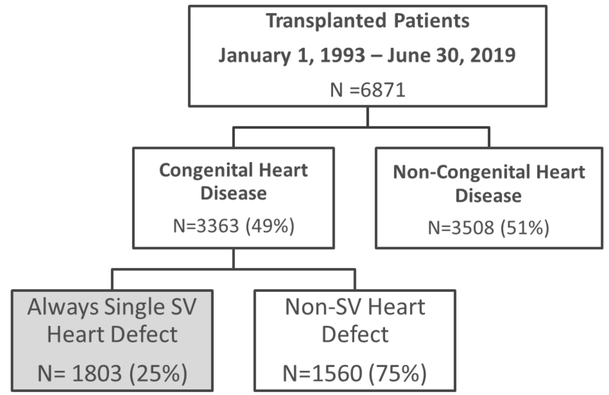


- A single ventricle defect is a type of heart defect that a child is born with. It occurs when one of the two pumping chambers in the heart, called ventricles, isn't large enough or strong enough to work correctly. In some cases, the chamber might be missing a valve.
- Single ventricle defects are rare, affecting only about five out of 100,000 newborns. They are also one of the most complex heart problems, usually requiring at least one surgery.

Materials & Methods

- Pediatric Heart Transplant Society Database was used to identify Single Ventricle (SV) Heart Disease
- Donor, transplant, and listing information for these patients was collected
- Machine Learning was complete using Python 3.5.2 p-value of <0.05 is considered statistically significant and Survival was compared using SAS 9.4.
- Text fields were utilized for modeling purposes

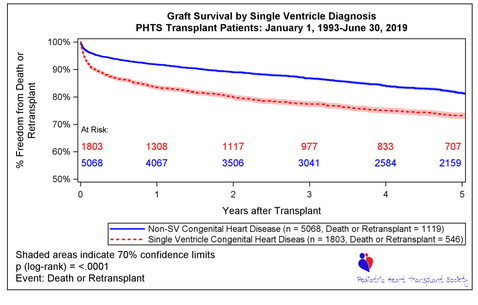
Cohort Identification



- Always Single Ventricle Congenital Heart Disease Includes
  - Hypoplastic Left Heart
  - Tricuspid Atresia
  - Double Inlet Left Ventricle

Single Ventricle Prediction

- Utilizing Patients flagged from our free text fields, SV Status was predicted in our total transplanted patient population (n=6871)
- Physicians must manually review text fields to identify if a patient is SV or Non-SV. Can we train a model to make this prediction?
- Patients that are single ventricle are of very high clinical interest as their risk for death or retransplant is greater as seen in the Kaplan-Meier, p<0.0001

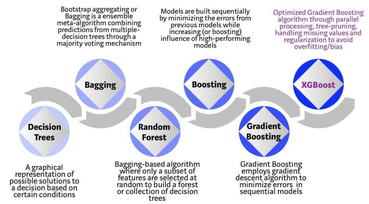


BioWordVec

- BioWordVec is a fastText model for creating biomedical word embeddings with subword information. It was trained using the PubMed corpora and MeSH. The use of subwords allows the model to produce embedding for OOV tokens.
- We used a pre-trained model provided by NCBI to create 200-dimensional Averaged Word Embeddings for text fields in our dataset.

XGBoost

- XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. When it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.



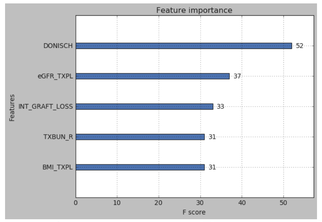
Machine Learning Models

- 80% of the data was used for training and 20% was used for testing
- Features with >25% missing were excluded from the model. Missing values were imputed with the median. Over 100 clinical features were utilized.
- Logistic Regression, SVM, Decision Trees, and XGBoost were all used to analyze 3 separate machine learnings models:
  - All binary and continuous features in PHTS
  - All Features + Text Fields
  - Only Text fields
- Text Fields Included Surgery History, CHD Specify, and Medical History at Listing
- XGBoost performed the best out of all models, with accuracy close to 90% and an AUC of 0.8.

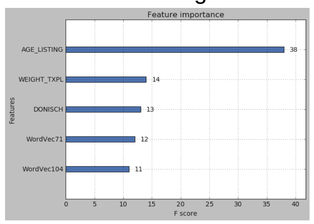
Results

ML Model	F1	AUC
All Features	0.765	0.649
All Features + Text Fields	0.877	0.834
Only Text	0.829	0.743

All PHTS Features



All Features + BioWordVec Embeddings



Importance provides a score that indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance.

Conclusions

- Free text fields provide valuable information that can often be overlooked by medical personnel
- Over 90 patients were added to the cohort from examination of free text fields alone
- These text fields along with other clinical factors can help predict single ventricle status with an accuracy close to 90%
- Text fields alone also provide value to predict single ventricle status with accuracy reaching 85%

WordCloud representing all keywords in the Other Specify Fields

Pediatric Heart Transplant Society

Natural Language Processing: CS 662