Evaluating statistical and machine learning methods to predict risk of in-hospital child mortality in Uganda

Grant Nguyen

A thesis
submitted in partial fulfilment of the requirements for the degree of
Master of Public Health

University of Washington
2016

Committee:
Abraham Flaxman, Chair
Herbie Duber
Arthur Mpimbaza

Program Authorized to Offer Degree:
Global Health

University of Washington

Abstract

Evaluating statistical and machine learning methods to predict risk of in-hospital child mortality in Uganda

Grant Nguyen

Chair of the Supervisory Committee:

Assistant Professor Abraham Flaxman

Global Health

Building on existing research on child mortality in Uganda, we used data from a six-hospital malaria surveillance system including signs and symptoms of all admitted patients, testing results, treatments provided, diagnoses at admission and discharge, and patient outcomes. We tested the relative performance of five statistical and machine learning methods to predict in-hospital mortality and extracted variable importance scores relating signs and symptoms, treatment, testing, and diagnosis to in-hospital mortality.

To determine the performance of each method to predict in-hospital child mortality, we applied all of the methods within 10 repetitions of 10-fold cross validation. Based on the predictions of each method on the held-out folds of the dataset, we used Area Under the Curve (AUC) to judge relative performance. We extracted variable importance scores for logistic regression, random forests, and gradient boosting machines, and ranked variable importance using inclusion and model coefficients for conditional inference trees and logistic regression.

Overall, logistic regression, random forests, and gradient boosting machines significantly outperformed decision and conditional inference trees in predicting in-hospital mortality. Using only variables present at admission, logistic regression, random forests, and gradient boosting machines had AUC values of 0.83 (0.80-0.85), 0.82 (0.79-0.84), and 0.80 (0.78-0.83) respectively, compared to AUC values of 0.72 (0.70-0.75) and 0.72 (0.69-0.75) for decision and conditional inference trees. The top-3 methods by AUC were able to correctly categorize 80% of in-hospital deaths while misclassifying 35% or fewer of eventual non-deaths as high-risk.

Considering only variables available at admission, the following variables were important predictors of mortality across four or more methods: treatment at admission with paracetamol, admission with severe malaria, age, inability to sit, inability to drink, hospital site, deep breathing, number of diagnoses at admission, and difficulty breathing. While the top 10-15 variables were highly ranked across multiple methods, many lower-ranked variables were highlighted as important by only one or two methods.

This study highlights the relative strength of logistic regression, random forests, and gradient boosting machines in predicting child mortality using a high-dimensionality dataset. The variable importance scores largely confirm the results of previous studies on symptoms and signs related to mortality, and point to interesting relationships for future investigation and research. However, divergences in variable importance underscore the usefulness of applying multiple methods to identify variables that remain important across various methods. Future directions for this work include applying ensemble models, further exploring key predictor variables, and extending this analytical framework towards other clinical prediction environments.

Contents

Background

Child Mortality in Uganda

Child mortality in Uganda remains a significant public health concern, with 1,733 deaths per 100,000 children under 5 in 2013[1]. Many child deaths occur within 24 hours of admission due to preventable and treatable disease such as pneumonia, diarrhea, malnutrition, and malaria[2]. One way to address child mortality in Uganda is to improve the quality of in-hospital care to reduce preventable in-hospital mortality. Recent studies have found that appropriate case management of significant illnesses in Uganda varies significantly by suspected disease and symptoms and signs at presentation[3]. Another avenue for improvement is through better identification of the distinct characteristics of patients at admission that lead to adverse in-hospital outcomes.

Previous Literature

This study builds on existing work using data collected in conjunction with the Uganda Malaria Surveillance Program (UMSP) in six hospitals in Uganda. A previous study using data from UMSP established a risk score of child mortality based on symptoms and signs available at admission[4]. This risk score, based on 13 of 25 available symptoms and signs, was created using a logistic regression with backwards selection on pooled data from four health facilities. It was intended to be used as a simple guide for in-hospital triage based on the specific patient characteristics and outcomes in the region. Another study used UMSP's data to examine the determinants of appropriate case management of illness at each hospital of interest[3]. We aim to fill the gap between these two papers by using data on symptoms and signs, diagnoses, treatment, testing, and outcomes to see what factors, when considered together, are most important in determining severe outcomes.

Multiple Method Extension

The prior analysis by Mpimbaza et al. used a single backwards-selection logistic regression model and cross-validated the model by comparing its performance across years. Building off of this, we compared the performance of five statistical methods: standard logistic regression, decision and conditional inference trees, and random forests and gradient boosting machines. To analyze the overall performance of each method, we implemented a repeated cross-validation framework. Many studies have analyzed predictors of health outcomes using multi-method analyses, particularly in the cardiovascular field[5,6]. By actively comparing the performance of multiple models and their results, we arrive at a well-rounded understanding of the factors associated with in-hospital mortality, and gain a better picture of the appropriate tools for risk analysis in similar unstructured datasets.

Data

Data Sources

The primary data source was collected as part of the UMSP. The surveillance project collected data from April 2010 to March 2014 across six public hospitals in Uganda: Tororo, Apac, Jinja, Mubende, Kabale, and Kanungu. These hospitals, a mix of general district and regional referral hospitals, were selected as enhanced sentinel sites for malaria surveillance, focused on children. In addition to introducing Standardized Medical Record Forms (MRFs) to catalogue patient information, caregivers were trained on the importance of recording high-quality data and medical records, and applying this data towards improving quality of care.

Although data collection was connected to malaria surveillance programs, the study data includes all patients treated at the six hospitals, regardless of suspected or confirmed malaria status. However, these sites may not be representative of hospitals in the region. Even though study sites were selected to represent different levels

of malaria endemicity, they are still a sub-sample of available hospitals and may have improved their service delivery quality due to other initiatives within the UMSP program. Data was collected using standardized MRFs and symptom and sign checklists incorporating Integrated Management of Childhood Illness (ICMI) terminologies[4]. We restricted this analysis to children under 5 years of age, and dropped all observations where the outcome of in-hospital death was missing.

Variables and Feature Engineering

This data includes measures of symptoms and signs at time of admission, in-hospital test results, treatment provided at admission and during their hospital stay, suspected diagnosis at admission, and final diagnosis at discharge. Malaria testing was a significant emphasis for the UMSP program, resulting in high coverage of malaria testing in suspected cases. In addition, care providers recorded data on child demographics such as age and sex.

Our outcome of interest was in-hospital mortality.

We dropped 41 variables that were only collected after 2012 or 2013. We also dropped variables with missing values in over half of observations, including MUAC, respiration count, immunization status, oxygen treatment, and pulse. There were still many remaining variables that captured similar information to the dropped variables: for example, respiration-related symptoms and signs included intercostal recession, cough, crackles, wheezing, deep breathing, and difficulty breathing. We recoded some variables from numeric to factor variables, encoding missing values as "informative" categorical values to be considered appropriately in the analyses. We generated categorical cutoffs to many numeric variables: for example, we transformed age in months into bins of 0-1 months, 2-3 months, 4-6 months, 7-11 months, and by year afterwards.

We engineered a number of features to explicitly identify important relationships across predictor variables. Each admitted child could be diagnosed with multiple unordered diagnoses at admission and discharge. To examine misdiagnosis at admission, we created a variable for Jaccard distance between diagnoses at admission and discharge.

$$distance = \frac{n_{match}}{n_{admit} + n_{discharge} - n_{match}}$$

We also measured the number of total diagnoses at admission and discharge, and created indicator variables for potential misdiagnosis of the top-10 diagnoses by comparing whether the patient was diagnosed with a condition at admission but not discharge. Conversely, we noted if patients were diagnosed at discharge but not at admission, indicating a potential missed diagnosis or worsening condition. Finally, we created indicator variables for a set of high-prevalence diagnoses to see whether they were provided appropriate treatment given an initial diagnosis of the condition. A list of the diagnosis-treatment maps is listed in the appendix.

We converted treatment and diagnosis variables to binary indicators for each possible treatment or diagnosis. To ease computational burden, we kept the top 20 of 65 total diagnoses for diagnosis at admission and discharge. Full lists of treatment and diagnosis options are listed in the appendix.

Table 1: Percent of cases with diagnosis at admission and discharge

Percent	Diagnosis (Admission)	Percent	Diagnosis (Discharge)
58.11	Malaria - severe	52.92	Malaria - severe
21.69	Respiratory infections (other)	16.80	Respiratory infections (other)
16.36	Anaemia	13.08	Pneumonia
16.29	Pneumonia	12.78	Anaemia
12.19	Diarrhoea - acute	9.98	Diarrhoea - acute
8.27	Malaria - uncomplicated	8.77	Malaria - uncomplicated
7.46	Other	6.42	Septicaemia
6.02	Septicaemia	4.97	Other
2.30	Pyrexia of unknown origin	1.66	Sickle Cell Disease
1.89	Sickle Cell Disease	1.19	Sev. acute malnu. + edema

Percent	Diagnosis (Admission)	Percent	Diagnosis (Discharge)
1.53	Sev. acute malnu. + edema	1.16	Sev. acute malnu. - edema
1.11	Sev. acute malnu. - edema	0.88	Measles
0.85	Measles	0.88	Pyrexia of unknown origin
0.72	Meningitis (other)	0.73	Missing
0.58	Urinary tract infection	0.59	Urinary tract infection
0.54	Diarrhoea - persistent	0.50	Malnutrition - Kwashiorkor
0.49	Malnutrition - Marasm-kwash	0.36	Meningitis (other)
0.49	Malnutrition - Kwashiorkor	0.34	Mod. acute malnutrition
0.49	Mod. acute malnutrition	0.33	Malnutrition - Marasm-kwash
0.39	Dysentery	0.32	Burns

In-hospital mortality varied widely by final diagnosis. Children with a final diagnosis of meningitis or malnutrition were most likely to die, with an in-hospital mortality rate between 7.40 and 18.68%. Although only 2.11% of cases with a discharge diagnosis of severe malaria died, the high prevalence of severe malaria led to 937 total deaths, the highest of any diagnosis.

Table 2: Number of cases and deaths with diagnosis at discharge, and percent of diagnosed cases that died

Diagnosis	Total Cases	Case Deaths	Percent of Cases
Meningitis (other)	273	51	18.68
Sev. acute malnu. - edema	1,060	146	13.77
Sev. acute malnu. + edema	1,090	128	11.74
Missing	493	46	9.33
Malnutrition - Marasm-kwash	311	27	8.68
Malnutrition - Kwashiorkor	473	35	7.40
Septicaemia	5,564	402	7.23
Anaemia	10,717	601	5.61
Pneumonia	11,789	652	5.53
Other	3,537	116	3.28
Burns	248	8	3.23
Sickle Cell Disease	765	21	2.75
Pyrexia of unknown origin	693	17	2.45
Measles	758	17	2.24
Malaria - severe	44,389	937	2.11
Mod. acute malnutrition	323	6	1.86
Diarrhoea - acute	9,493	153	1.61
Urinary tract infection	396	4	1.01
Respiratory infections (other)	14,716	141	0.96
Malaria - uncomplicated	7,503	28	0.37

After variable cleaning and feature engineering, there were 200 predictors of interest. We ran each statistical model twice: once on the full dataset of predictor variables, and once on a subset of variables that were only observable at time of admission. The admission-only variables included symptoms and signs, initial diagnosis, treatment at admission, and covariates. We decided to run the analysis on the admission-only dataset to examine the potential predictive power of each method if it were applied as a diagnostic tool directly following admission. This is the most informative model for applicability towards a clinical triage setting. Meanwhile, the all-variable dataset gives more information around potential post-admission factors that contribute to in-hospital death, including relationships between disease and treatment.

Predictor variables were categorized into the following categories:

Table 3: Total count of predictors, by variable type

Type	Count
Symptom or Sign	27
Initial Diagnosis	21
Discharge Diagnosis	22
Diagnosis Indicators	22
Treatment at Admission	39
In-Hospital Treatment	40
Treatment Indicators	19
Testing	5
Covariate	5
Total	200

Methods

Statistical Models

For this analysis, we used a logistic regression model, two variants of tree-based models (decision and conditional inference trees), and two popular ensemble learning methods (random forests and gradient boosting machines). By running each model 100 times through 10 repetitions of 10-fold cross-validation, we assessed the relative performance quality of each model. To judge performance, we compared methods using area under the curve (AUC), true and false positive rate cutoffs, accuracy, the Hosmer-Lemeshow statistic, and accuracy. In addition, we identified the most important variables according to each method. Data cleaning was performed in Stata 13.0 and statistical analysis was performed in R using the following packages: caret (cross-validation), rpart (decision trees), party (conditional inference trees), randomForest (random forests), doMC (cluster computing), xgboost (gradient boosting machines), and ROCR (AUC calculation)[7–14].

10-times 10-fold Cross-Validation

For each of the 10 repetitions of our analysis, we split the dataset into 10 equal partitions, commonly referred to as folds, assuring that deaths were distributed equally across folds. This follows generally-accepted benchmarks for accurate training and testing of model performance[15]. We then ran the analyses 10 times, once for each unique fold. Each time a different fold was run, we held out the fold and trained each model on the remaining data, then tested the performance of each model on the test dataset. We repeated this process 10 times for a total of 100 different test runs across which we could compare results. By performing this random validation, we arrive at stable estimates of model performance and reliability.

Logistic Regression

Logistic Regression was run using all potential predictors. We relied on the engineered variables as proxies for interaction terms in this regression. Logistic regression has been used widely in the medical field to predict patient risk[16–18]. Due to computational limitations induced by the size of the dataset and number of predictors, we were not able to run backwards selection on logistic regression as in the original study by Mpimbaza et al.[4]. We considered other methods including LASSO logistic regression which perform automated variable selection, but they generally did not offer significant performance improvement compared to the more commonly-used logistic regression[19].

Classification and Regression Trees

We used classification and regression trees, commonly known as decision trees, as one of the two tree-oriented methods for analysis[20]. They can identify valuable interactions and sub-splits within variables of interest that are obscured when looking simply at variable importance. The results of a decision tree analysis include a tree with branches, where each split in a branch represents a variable cutpoint that plays a strong role in determining the outcome. In spite of their wide usage, traditional decision trees have been criticized by some due to biased variable selection[21].

Conditional Inference Trees

Conditional inference trees are another variant of tree-based methods[11]. Differing from decision trees, they implement randomization tests to address problems with variable selection and overfitting[11]. Conditional inference trees produce similarly-formatted results to decision trees, allowing for further analysis of cross-variable interactions and easy interpretation of results. However, both conditional inference and decision trees suffer from the weakness that they only fit one tree on the data, rather than building multiple sets of trees as random forests and gradient boosting machines do.

Random Forests

Random forests are increasingly popular and one of the more widely-used methods of supervised machine learning[22–25]. In this method, decision trees are constructed by taking randomly resampled partitions of the data and using a random subset of potential variables to inform each split of the data. By randomly sampling observations and potential splitting variables, this method introduces an element of randomness that prevents the model from overfitting. After training all trees, they are combined together based on the predictions of each individual tree[26]. Although random forests are tree-based, the ensembled nature of the results mean that they only produce variable importance scores that can be difficult to interpret, as they don't indicate the direction of any predictor towards the outcome.

Gradient Boosting Machines

Gradient boosting machines are a relatively new form and popular method of machine learning[27]. Although gradient boosting machines are ensemble tree-based models similar to random forests, they start with weak learning trees and build more reliable trees based off of the residuals of the predictions from each prior tree. This differs in approach from random forests, in which trees are independent of one another and rely on the randomness in variable and observation selection to produce unbiased estimates. Similar to random forests, the results can oftentimes be difficult to interpret because they report variable importance scores rather than directional relationships.

Variable Weights

The outcome of interest, death, is a rare event. Of 83,562 total admissions, 2,583 (3.13%) died. To address under-representation of death and to guide the methods towards accurate prediction, we ran the full repetition and fold combinations on the following values of class weights for the training dataset: 1, 5, 10, 20, 30. These class weights were applied using manual resampling with replacement of the patients who died in-hospital to create weighted datasets to train each method. Performance was evaluated using un-weighted testing datasets. We selected the best-performing weights by model using AUC, and report results from the ideally-weighted models throughout the rest of the paper.

Performance Determination

For each repetition/fold/weight combination, we calculate area under the curve (AUC), true and false positive rates, Hosmer-Lemeshow statistic, and accuracy.

AUC is a commonly-used metric to evaluate model performance by estimating the area under a curve generated by using model predictions to estimate the model's positive and negative predictive values[28,29]. Although AUC does not capture performance across at different points of the curve generated by these values, it is widely accepted as a valid approach to evaluate the overall performance of the model particularly when the outcome is very prevalent or rare.

Building off of the Receiver Operator Curve (ROC) results used to calculate the AUC, we also report the False Positive Rate (FPR) at various cut-points of True Positive Rate (TPR). Given the consequences of misclassifying a future in-hospital death, we aim to see how many additional people will be misclassified as potential deaths to achieve incrementally better performance at capturing more deaths.

The Hosmer-Lemeshow test is a test based on the chi-squared statistic that summarizes the accuracy of a model when separated into n probability bins. In general, it compares the expected cases and non-cases based on the predicted probabilities of an event with the actual cases and non-cases that exist within each probability bin. Compared to AUC, the Hosmer-Lemeshow test can evaluate how accurate each method is in predicting the actual probabilities of an event occurring, rather than simply the tradeoffs between TPR and FPR. However, it has been criticized for reporting statistically significant miscalibration in models that are run on large datasets, and is more difficult to interpret than a ROC curve[30].

Finally, accuracy is a widely-used metric in the machine learning field for assessing model performance. To assess accuracy for each method, within each rep/fold combination, we estimated the model accuracy at a number of probability cutoffs (.1, .2, .3, .4, .5, .75, and .9). While this may help with ascertaining optimal cutpoints, accuracy is a crude metric due to the class imbalances across the outcome of death, which it does not handle well[29]. For example, a model that predicted that no child would die would achieve an accuracy score of roughly 96.77% even though it should be considered a poor model due to its inability to detect any deaths.

Results

Method Performance

Judging by AUC, logistic regression, random forests, and gradient boosting machines far outperformed their traditional tree-based counterparts (Table 4). Of the top three performers, logistic regression performed slightly better than random forests and gradient boosting machines. Both decision trees and conditional inference trees struggled with classification: decision trees sometimes failed to create a tree at all, producing simple roots without classification. The table below presents results by model with variables present only at admission and with all variables available. All results in this section are presented only for the optimal death weight results within each method/admission combination, as determined by AUC.

Table 4: Model Performance (AUC) by admission and model type

Method	Death Wt Admit Only	AUC Admit Only	Death Wt All Vars	AUC All Vars
Logistic	20	0.83 (0.8 - 0.85)	10	0.87 (0.85 - 0.89)
Grad. Boost	1	0.8 (0.78 - 0.83)	1	0.87 (0.85 - 0.89)
Rand. Forest	5	0.82 (0.79 - 0.84)	10	0.86 (0.84 - 0.88)
Dec. Tree	30	0.72 (0.7 - 0.75)	30	0.73 (0.68 - 0.77)
CI Tree	20	0.72 (0.69 - 0.75)	20	0.7 (0.67 - 0.72)

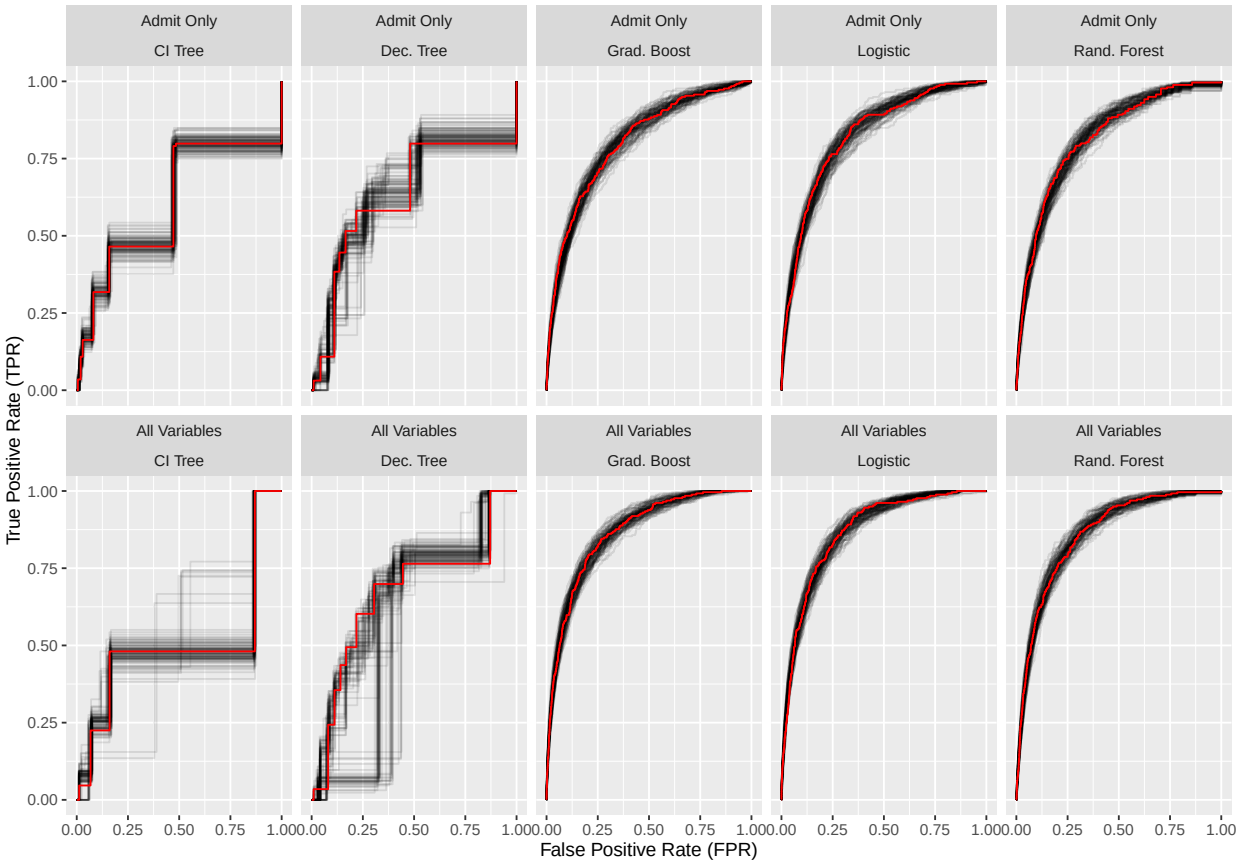

Figure 1: ROC Curves by admission type and method, one line per rep/fold combination. Median curve by AUC in red.

Most methods produced surprisingly low uncertainty surrounding the median estimate, indicating that method performance was stable irrespective of the random samples of data that informed each run (Figure 1). The median ROC curves underscore the AUC results, demonstrating the clear gap between the three primary methods and the conditional inference and decision trees (Figure 2).

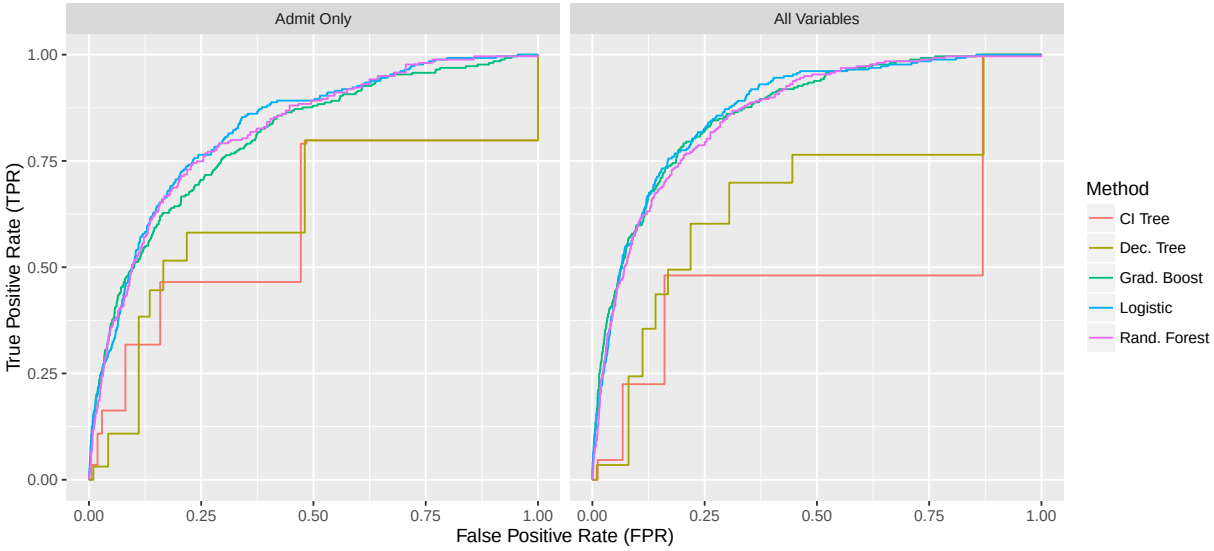

Figure 2: ROC curves of median AUC rep/fold combination by admission type and method

Based off of the ROC curves, we compared model performance based off of various cutoffs in True Positive Rate (TPR): .7, .8, and .9 (Tables 5 and 6). These cutoffs illustrate the poor performance of the decision and conditional inference trees compared to the other models. The top 3 methods correctly classified over 70% of in-hospital deaths while mis-classifying 25% of non-deaths using admission-only variables, dropping to 16% mis-classification when using all available variables. The False Positive Rate (FPR) rose to 35% and 25% respectively to capture 80% of in-hospital deaths, and 55% and 40% for a 90% TPR.

Table 5: False Positive Rate at cutoffs of Total Positive Rate, Admit Only

Method	FPR at TPR 0.7	FPR at TPR 0.8	FPR at TPR 0.9
CI Tree	0.47 (0.46-0.49)	0.77 (0.46-1)	1 (1-1)
Dec. Tree	0.48 (0.31-0.54)	0.66 (0.48-1)	1 (1-1)
Grad. Boost	0.23 (0.19-0.29)	0.35 (0.29-0.43)	0.55 (0.46-0.65)
Logistic	0.2 (0.17-0.25)	0.3 (0.25-0.38)	0.48 (0.39-0.58)
Rand. Forest	0.22 (0.18-0.26)	0.31 (0.26-0.38)	0.5 (0.42-0.6)

Table 6: False Positive Rate at cutoffs of Total Positive Rate, All Variables

Method	FPR at TPR 0.7	FPR at TPR 0.8	FPR at TPR 0.9
CI Tree	0.85 (0.51-0.87)	0.87 (0.86-0.87)	0.87 (0.86-0.87)
Dec. Tree	0.38 (0.3-0.47)	0.69 (0.4-0.87)	0.85 (0.81-0.87)
Grad. Boost	0.15 (0.12-0.18)	0.23 (0.19-0.28)	0.39 (0.3-0.46)
Logistic	0.15 (0.12-0.19)	0.22 (0.18-0.28)	0.35 (0.29-0.42)
Rand. Forest	0.16 (0.12-0.19)	0.24 (0.19-0.29)	0.38 (0.31-0.44)

The importance of developing sensitive metrics is underscored by Figure 3, which shows the distribution of predicted probabilities between non-deaths and deaths in one test-run of the logistic regression. The high number of non-deaths underscores the challenges associated with prediction. Moving further down the predicted probabilities, the number of included non-deaths increases drastically while distribution of included deaths is relatively stable.

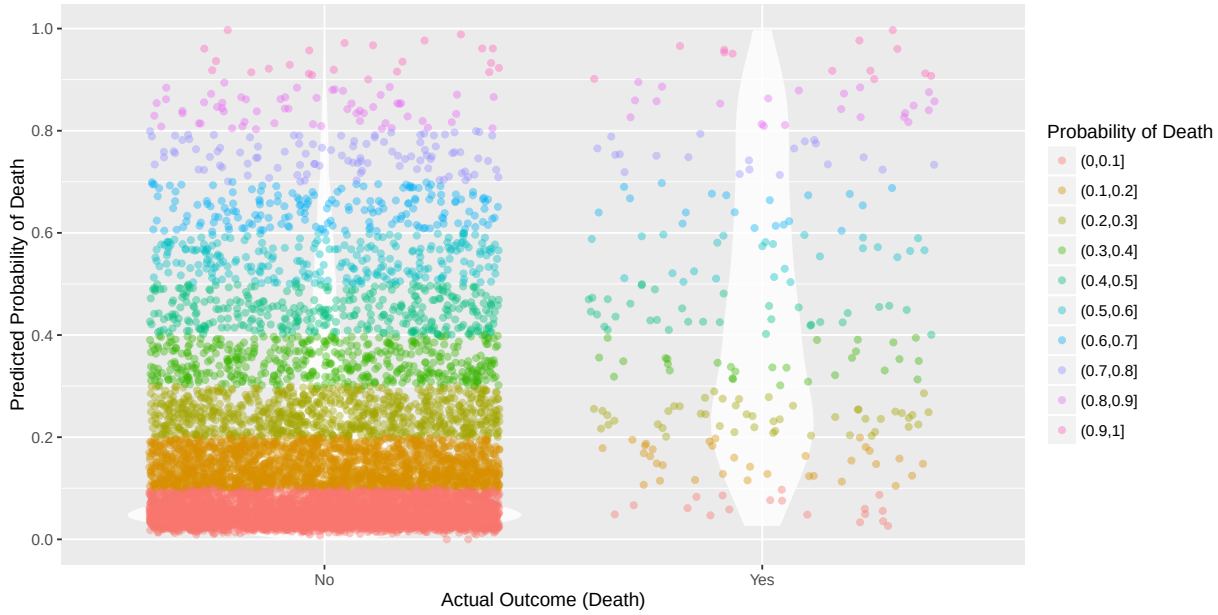

Figure 3: Predicted probability of death from logistic regression for each patient in a test dataset, separated by actual outcome

The Hosmer-Lemeshow statistics show that gradient boosting machines and random forests perform fairly well. Surprisingly, logistic regression performs poorly, with a high HL score. Logistic regression predicts higher probabilities of death than actually happen, particularly for high-likelihood cases. However, its good performance on AUC indicates that although the probability predictions from the logistic regression may be biased, it can reliably predict true positives while avoiding false positives. Decision and conditional inference trees performed poorly by this metric, while gradient boosting machines and random forests performed well. However, all models achieved a significant mean p-value for their statistic, indicating generally-biased prediction. Complete Hosmer-Lemeshow results can be found in the appendix.

Finally, accuracy across a number of bins was relatively similar across methods. Accuracy climbs drastically across the first three cutpoints specified (up to probability of 0.3) before leveling off for most methods. This is mostly due to the low probabilities of death within each major predictive group and the relatively low occurrence of death in this data. Given the aforementioned limitations of accuracy, particularly given the low prevalence and high importance of in-hospital deaths, we focused more AUC and the TPR/FPR analyses for model evaluation. All accuracy results can be found in the appendix.

Variable Importances

Within each fold/repetition run, variable importances were extracted from three methods: conditional inference trees, random forests, and gradient boosting machines. Random forests produces two importance measures: accuracy and gini. Accuracy refers to the variable's effect on prediction accuracy, while gini refers to the variable's effect on the impurity of a tree through the contribution of the variable's node towards splitting the dataset into groups. These variable importances were averaged across repetition and fold to create average importance scores across all model runs.

For logistic regression, we kept each variable that was significant across all repetition and fold combinations, then ranked them by the absolute value of the mean beta of each variable. For decision trees, we extracted the number of times that a variable served as a node within a single tree (e.g. if a single tree contained three nodes which used the value of cv_age to split the data, it would be noted three times).

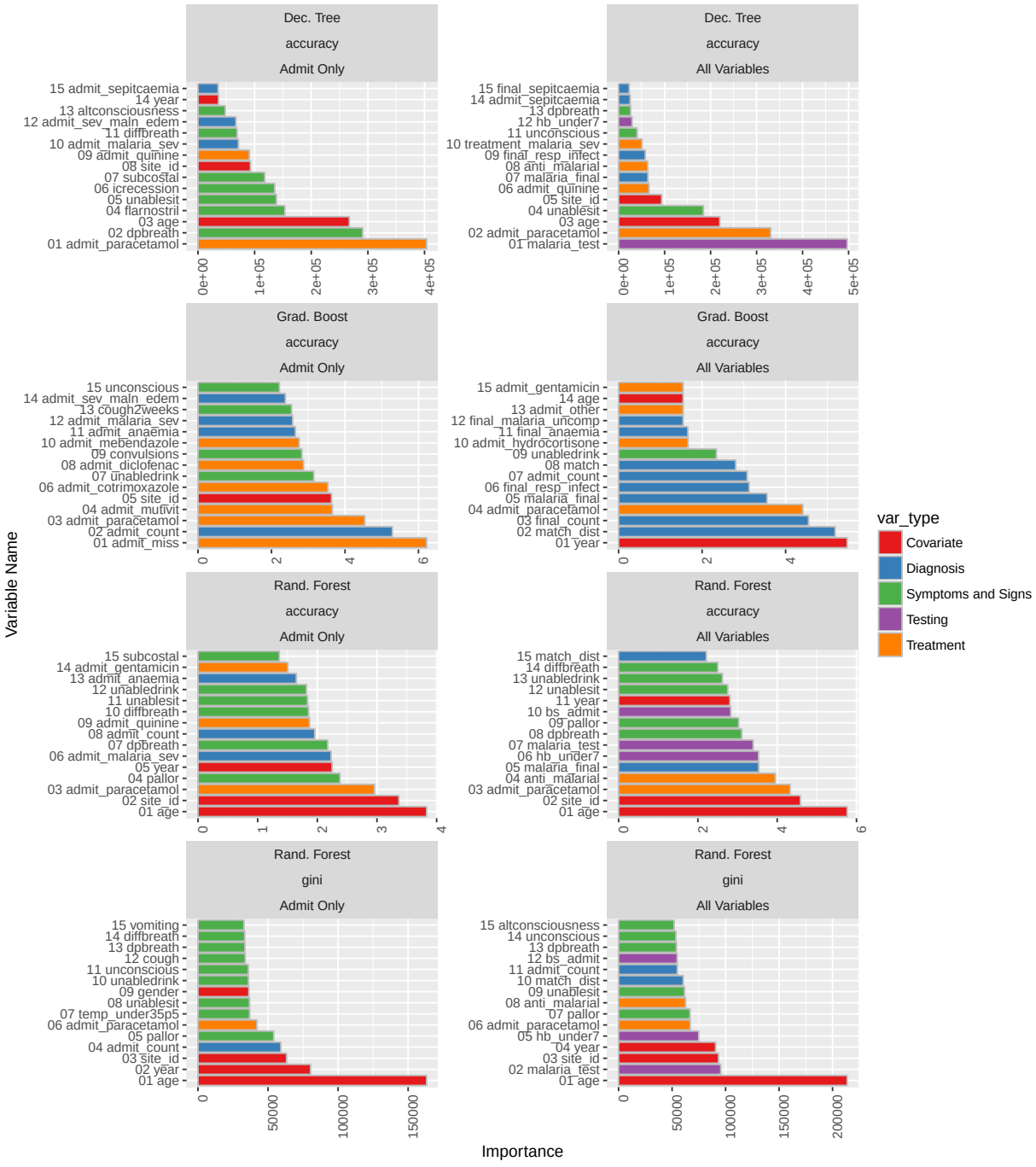

Figure 4: Variable Importances by method and admission type

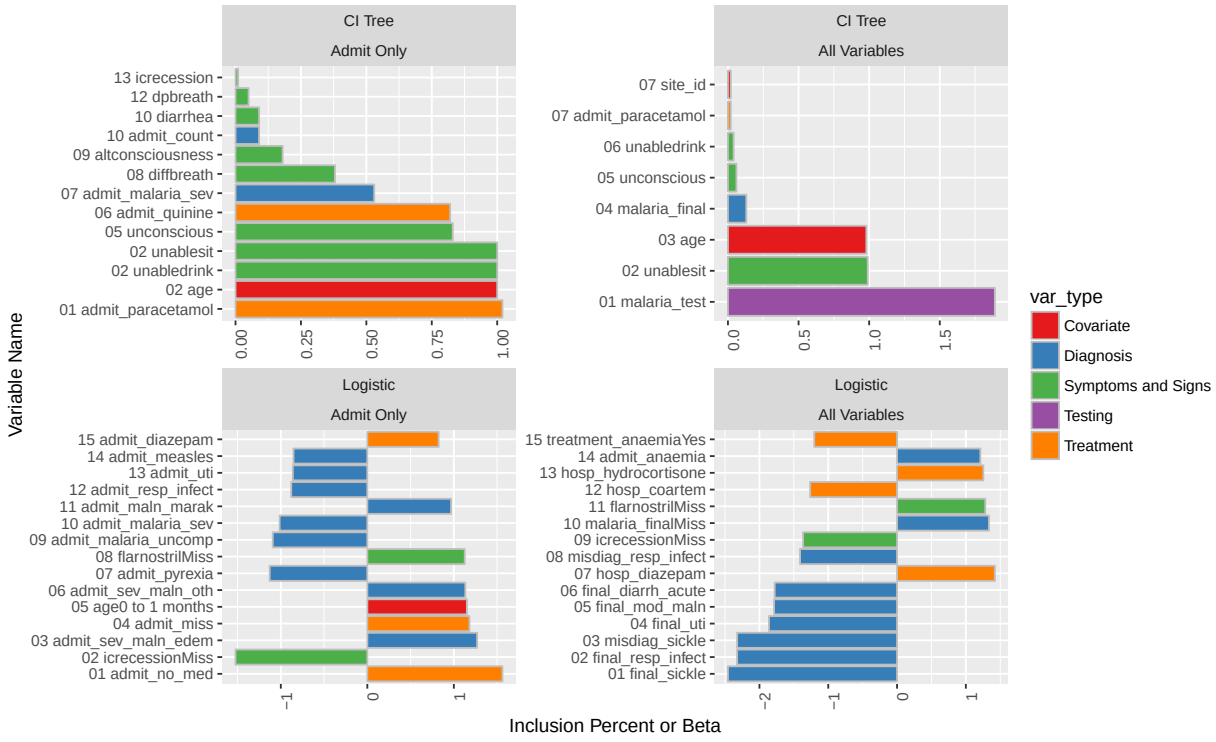

Figure 5: Inclusion in decision tree and model betas of significant (p<.05) predictors in logistic regression

Overall, there was a striking difference across predictive methods related to variable importance. Although some variables ranked in the top-15 across a number of models, many variables were only predictive in one or two models (Figures 6 and 7). In particular, the logistic regression and gradient boosting machines appeared to generate variable importance scores that differed strongly from other methods.

Using only admission variables, the following variables were important predictors of mortality across four or more methods: treatment at admission with paracetamol, admission with severe malaria, age, inability to sit, inability to drink, site, deep breathing, number of diagnoses at admission, and difficulty breathing. Using all variables in the full dataset, the following variables were important across three or more models: age, paracetamol at admission, malaria test result, inability to sit, site, final diagnosis of malaria, year, Jaccard distance of diagnoses, final diagnosis of respiratory infection, anti-malarial treatment, unconsciousness, hb under 7 mg, inability to drink, and deep breathing.

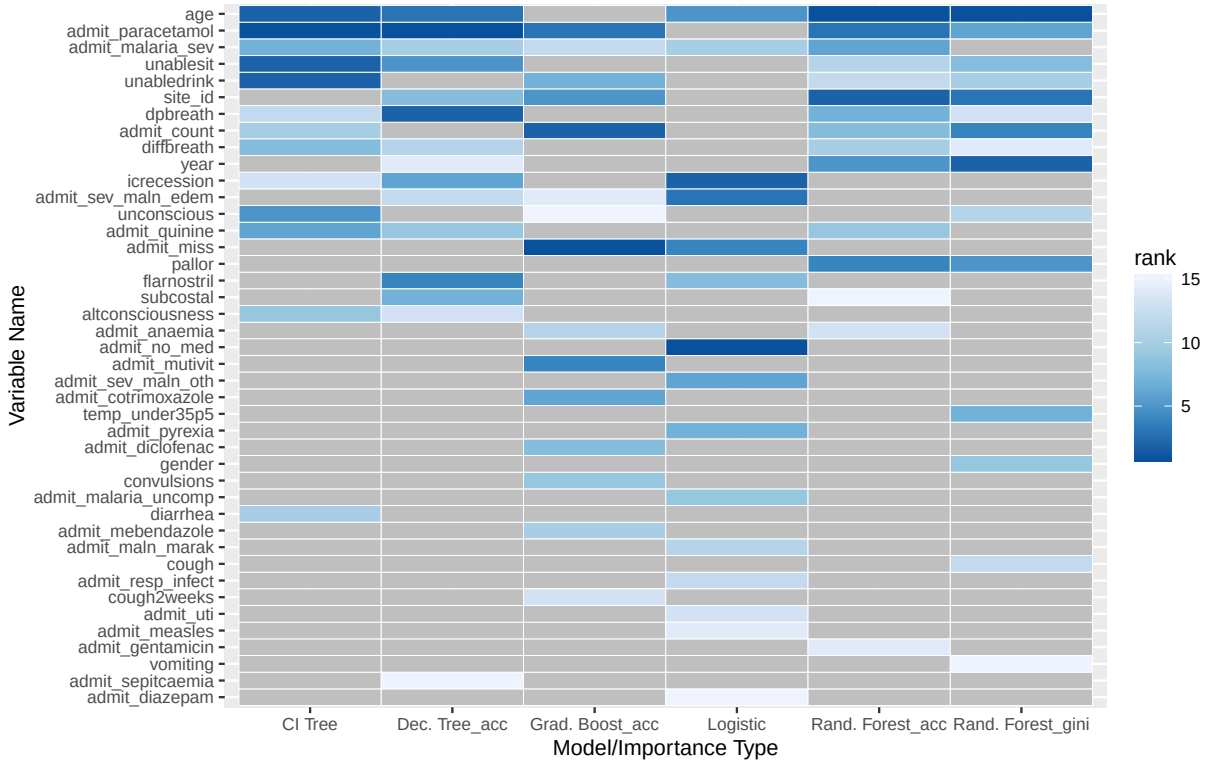

Figure 6: Comparison of variable importance and inclusion across method types for Admit Only variables

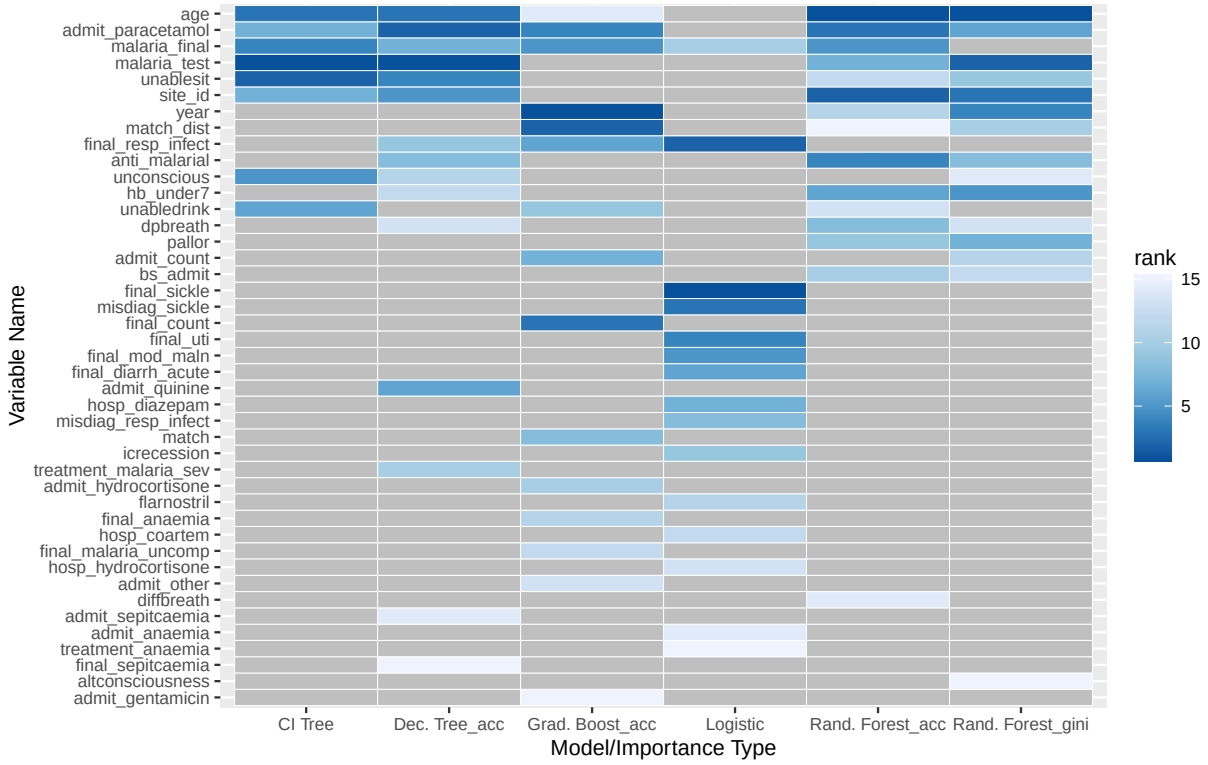

Figure 7: Comparison of variable importance and inclusion across method types for All variables

Discussion

Focusing on in-hospital child mortality in six hospitals in Uganda, this project accomplished two primary goals: to ascertain the relative performance of various statistical methods to determine risk of mortality, and to highlight key predictors of in-hospital mortality across multiple methods.

Model Comparison Results

Using AUC as the primary benchmark to determine model accuracy, we found that logistic regression, random forests, and gradient boosting machines performed relatively well, particularly compared to conditional inference and decision trees. Although gradient boosting and random forests are generally considered to perform well with large, unstructured datasets, logistic regression slightly outperformed both methods.

Using only variables available at admission, the top-3 methods could predict over 80% of cases while misclassifying 35% or fewer of eventual non-deaths as high-risk. Using all information available at both admission and discharge, this improves to a misclassification rate of 35% or less. While the methods are fairly discerning in predicting in-hospital deaths, even marginal improvements in prediction quality could make these models much more useful in an applied clinical setting.

Each of the high-performing methods has unique strengths. Logistic regression has been a standard and well-understood model, particularly in clinical settings where one would eventually need to apply their results. Both random forests and gradient boosting machines have the advantage of taking into account cross-variable interactions that would need to be explicitly specified in a logistic regression – a difficult task with datasets including many potential predictors. Of the machine learning methods, random forests are relatively more established and well-known, and are easily applied using cluster computing. Meanwhile, gradient boosting

machines are growing in popularity, particularly for predictive analytics, and are computable much quicker than random forests in many settings.

Given the differences between underlying methodologies in the models, we anticipated some divergence across the model results. Accordingly, there were distinct differences in many of the major predictors in the models. Our results indicate that researchers who attempt to examine predictor-outcome relationships across many variables of interest should strongly consider analyzing their data across a number of different methods rather than basing their conclusions on a single method. Using only a single method to examine outcomes, as the results in this paper show, may lead to conclusions about variable importance that conflict with the results of other similarly appropriate models. An extension of this work could include an ensemble modeling technique, blending the results of multiple models together to create a combined model that takes into account the various elements of the data that each model has picked out as important.

Predictors of Interest

Even given the divergence in the variable importance results, we found a number of variables that were considered important across a variety of methods. For these variables, a multi-faceted approach to modeling and variable selection enables more confident confirmation and exploration of variable associations with the outcome of death.

Of the top-15 variables from each method, many symptoms and signs highlighted here confirm the results found in Mpimbaza et al.[4]. Inability to sit or drink, deep breathing, difficulty breathing, age, unconsciousness, and pallor were all symptoms and signs selected in the original analysis that were still significant in this analysis when including many additional predictors of interest. Intercostal recession was one factor that was important in a number of methods but not in the original analysis, along with flaring of the nostrils which was not analyzed in the original study. The importance of symptoms and signs faded in the all-variable models; however, some symptoms and signs such as inability to sit, age, unconsciousness, inability to drink, pallor, and deep breathing were still important predictors across multiple models.

Provision of paracetamol at admission was marked as important across all methods except for logistic regression, in both the admission-only and all-variable datasets. Paracetamol is primarily given for fever and pain reduction, rather than for more severe illnesses. 59.3% of all patients were given paracetamol at admission, of whom 1.9% died compared to 4.9% of those who didn't recieve paracetamol. It is likely that, rather than paracetamol itself preventing death, paracetamol provision is associated with certain diseases or symptoms that are less likely to cause in-hospital mortality. For example, patients were given paracetamol in only 28% of cases of severe acute malnutrition with or without edema, which had two of the top 3 case fatality rates among diagnoses. Meanwhile, 67.5% of patients with severe malaria were given paracetamol at admission.

In both the admit-only and all-variable models, malaria status was an important predictor. For the admit-only model, an admission diagnosis of severe malaria was important in all models except for gradient boosting machines and logistic regression. In the all-variable model, the results of malaria testing, a final diagnosis of malaria, and treatment with an anti-malarial were all important factors across multiple methods. Overall, mortality rates for admissions of uncomplicated malaria were much lower than others, at .37%. One future direction of this analysis could be to look at the determinants of in-hospital mortality specifically among patients who have suspected malaria, to focus more on malaria-specific characteristics and outcomes.

The hospital site, number of diagnoses at admission, and year were also important predictors in the admit-only models. With the all-variable models, hospital site and year remained important, along with differences between admission and final diagnoses, hemoglobin below 70 grams per liter, and treatment with an anti-malarial drug. The importance of hospital site and year are difficult to infer from, as these associations could be due to differences in disease dynamics and severity or quality of care across hospitals or time. Differences between admission and final diagnoses, as measured by the Jaccard distance, illustrate the impact of extreme differences between the diagnoses listed at admission and discharge, whether due to misdiagnosis or worsening condition. Finally, test results of hemoglobin levels under 70 grams per liter, recommended by the WHO as an indicator for severe anaemia, was an important factor – in fact, mortality was lowest in those who

did not receive a test (2.5% of non-testers) and highest in those with hemoglobin above 70 g/l (6.3% of patients testing above 70 g/l)[31]. Those with over 70 g/l could have been in poor enough condition to merit a hemoglobin test and also have high-risk factors beyond hemoglobin level that worsened their ultimate outcomes.

In both the admit-only and all-variable datasets, the logistic regression highlighted very different variables as important as compared to other methods. This could be due to the way in which variable importance was determined for logistic regression. Each variable must be significant across all 100 repetition/fold combinations before being ranked based on the betas in the logistic regression. It is possible that some variables may be considered important across most regressions, but not significant in some repetition/fold combinations. Of the highest-ranked logistic regression results, the high rankings of a final diagnosis of sickle cell disease, urinary tract infection, and misdiagnosis of sickle cell are particularly surprising. Both urinary tract infection and sickle cell disease have low rates of in-hospital mortality per case, at 2.75% and 1.01% respectively. Patients with a diagnosis of sickle cell disease at admission but not discharge had lower rates of mortality, at 2.5%. However, it seems as though the logistic regression is potentially over-influenced by small sample sizes: there were only 8 cases of misdiagnosis of sickle cell that ended in death, and only 4 final diagnoses of urinary tract infections ending in death. While these results are thought-provoking, the lack of concordance with the other high-performing methods indicates that more investigation should be done to examine the root causes of these divergences.

Translation and Generalizability

This analysis aims to inform clinical decision-making and evaluation by identifying factors at admission or in-hospital factors that may influence a child's risk of in-hospital death. Although the results of this analysis are not as directly interpretable as Mpimbaza et al., they gain strength by using multiple models for variable inference and expanded cross-validation of the results[4].

The analytical approach of this study can be generalized to a number of different risk prediction settings. In any risk prediction setting, a similar modeling approach could be used to test multiple models, compare performance, and analyze variable importances across methods. Although we used cluster computing to speed up processing time, this analysis could easily be adapted to a setting with fewer resources by allowing for longer computation time to run all requisite analyses. All code used for this analysis is publically available (source code available at https://github.com/gnguy/mph_thesis_ml).

Limitations

As mentioned previously, the results of this study are somewhat more difficult to interpret than previous analyses due to the study's increased methodological complexity. The variable importances from certain methods do not indicate direction, making it more difficult to infer relationships based on importance scores alone. In addition, the 10-repetition, 10-fold cross-validation framework requires variable importance and inclusion to be aggregated across run and fold combinations. However, this complexity allows for much deeper examination and cross-validation of the factors that influence in-hospital mortality in this case.

Additionally, extended parameter tuning could not be performed on a number of methods due to computational limitations. Although tuning the parameters of each model could improve their performance, the baseline models performed relatively well, and tests of parameter tuning resulted in minimal performance gain and significant computational strain.

Finally, we cannot currently apply these results to perform real-time risk prediction at admission. The results at the current stage have highlighted certain avenues to explore more in-depth, but more work remains to explore the cross-relationships between predictors of interest. Also, given the computational and infrastructural limitations in Ugandan hospitals, it will be difficult to make real-time predictions using these models. However, further testing and translation of the current models and variable importances may inform

practice-based changes in treatment and triage to be applied in a clinical setting. In addition, we have laid the framework for predictive modeling in higher-resource settings where real-time risk calculation is possible.

Conclusions

We measured the performance of five popular classification models to predict in-hospital child mortality in six Ugandan health facilities, and compared variable importance and inclusion across all methods. As a result, we have established the high performance of logistic regression, random forests, and gradient boosting machines compared to decision trees and conditional inference trees. Using only variables available at admission, logistic regression, random forests, and gradient boosting machines all had mean AUC values of over 0.8. Accordingly, these three high-performing models could predict over 80% of deaths given a tradeoff of miscategorizing 30-35% of non-deaths as high-risk cases.

In partnership with clinicians in Uganda, we can explore the relationships between death and symptoms and signs at admission, treatment, testing, and diagnosis. The variables available only at admission can help clinicians determine which children are at highest risk of mortality, sooner in the process. Meanwhile, the results of the full-variable analysis can highlight potential areas of concern or further exploration related to the accuracy of initial diagnoses, testing procedures, and quality of in-hospital treatment. Overall, this work confirms the findings of Mpimbaza et al., with many of the symptoms and signs at admission highlighted in that paper also marked as important in this analysis[4]. In addition, we discovered that the relationship of paracetamol treatment with mortality may be due to its association with low-mortality diseases and that the high mortality of patients with high hemoglobin test results may be due to the lack of guidance from a negative result combined with overall poor health of children who were tested. Finally, we identified the strong importance of hospital site and year, indicating distinct differences for in-hospital mortality across time and site that are not captured by the symptom, sign, disease, and treatment variables available in this dataset.

This analysis highlights the drawbacks of relying on single-model analyses of large, unstructured datasets. Given the high variation in variable importance across models, particularly with logistic regression, it is important that researchers consider using multiple modeling techniques to check for variable importance across multiple methods. By doing so, they can build confidence in their final results and identify new predictors that may have been missed by a single method of analysis.

References

1 GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: A systematic analysis for the Global Burden of Disease Study 2013. *Lancet (London, England)* 2015; **385**: 117–71.

2 Afolabi BM, Clement CO, Ekundayo A, Dolapo D. A hospital-based estimate of major causes of death among under-five children from a health facility in Lagos, Southwest Nigeria: Possible indicators of health inequality. *International Journal for Equity in Health* 2012; **11**: 39.

3 Sears D, Kigozi R, Mpimbaza A *et al.* Anti-malarial prescription practices among outpatients with laboratory-confirmed malaria in the setting of a health facility-based sentinel site surveillance system in Uganda. *Malaria Journal* 2013; **12**: 252.

4 Mpimbaza A, Sears D, Sserwanga A *et al.* Admission Risk Score to Predict Inpatient Pediatric Mortality at Four Public Hospitals in Uganda. *PLOS ONE* 2015; **10**: e0133950.

5 Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statistics in Medicine* 2007; **26**: 2937–57.

6 Au AG, McAlister FA, Bakal JA, Ezekowitz J, Kaul P, Walraven C van. Predicting the risk of unplanned readmission or death within 30 days of discharge after a heart failure hospitalization. *American Heart Journal* 2012; **164**: 365–72.

7 StataCorp. Stata statistical software: Release 13.

8 R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2015 https://www.R-project.org/.

9 Jed Wing MKC from, Weston S, Williams A *et al.* Caret: Classification and regression training. 2016 https://CRAN.R-project.org/package=caret.

10 Therneau T, Atkinson B, Ripley B. Rpart: Recursive partitioning and regression trees. 2015 https://CRAN.R-project.org/package=rpart.

11 Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 2006; **15**: 651–74.

12 Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002; **2**: 18–22.

13 Chen T, He T, Benesty M. Xgboost: Extreme gradient boosting. 2016 https://CRAN.R-project.org/package=xgboost.

14 Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: Visualizing classifier performance in r. *Bioinformatics* 2005; **21**: 7881.

15 Kim J-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis* 2009; **53**: 3735–45.

16 Morrow DA, Antman EM, Charlesworth A *et al.* TIMI Risk Score for ST-Elevation Myocardial Infarction: A Convenient, Bedside, Clinical Score for Risk Assessment at Presentation An Intravenous nPA for Treatment of Infarcting Myocardium Early II Trial Substudy. *Circulation* 2000; **102**: 2031–7.

17 Lindström J, Tuomilehto J. The Diabetes Risk Score. *Diabetes Care* 2003; **26**: 725–31.

18 Blatchford O, Murray WR, Blatchford M. A risk score to predict need for treatment for upper-gastrointestinal haemorrhage. *Lancet (London, England)* 2000; **356**: 1318–21.

19 Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B* 1994; **58**: 267–88.

20 Breiman L, Friedman J, Stone CJ, Olshen R. Classification and Regression Trees. 1984 https://www.crcpress.com/Classification-and-Regression-Trees/Breiman-Friedman-Stone-Olshen/p/

book/9780412048418 (accessed June 25, 2016).

21 Loh W, Shih Y. Split selection methods for classification trees. *Statistica Sinica* 1997; **7**. http://www3.stat.sinica.edu.tw/statistica/j7n4/j7n41/j7n41.htm (accessed June 28, 2016).

22 Flaxman AD, Vahdatpour A, Green S, James SL, Murray CJ. Random forests for verbal autopsy analysis: Multisite validation study using clinical diagnostic gold standards. *Population Health Metrics* 2011; **9**: 29.

23 Riddick G, Song H, Ahn S *et al.* Predicting in vitro drug sensitivity using Random Forests. *Bioinformatics* 2011; **27**: 220–4.

24 Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: Exploiting interactions using random forests. *BMC Genetics* 2004; **5**: 32.

25 Gray KR, Aljabar P, Heckemann RA, Hammers A, Rueckert D. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroImage* 2013; **65**: 167–75.

26 Breiman L. Random Forests. *Machine Learning* 2001; **45**: 5–32.

27 Friedman JH. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 2001; **29**: 1189–232.

28 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**: 29–36.

29 Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 1997; **30**: 1145–59.

30 Feudtner C, Hexem KR, Shabbout M, Feinstein JA, Sochalski J, Silber JH. Prediction of Pediatric Death in the Year after Hospitalization: A Population-Level Retrospective Cohort Study. *Journal of Palliative Medicine* 2009; **12**: 160–9.

31 Organization WH. Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity. 2011. http://www.who.int/vmnis/indicators/haemoglobin/en/ (accessed Aug 10, 2016).