Predicting death in patients with pneumonia under the age of 21

Kristen McGarry, Ryan Barbaro, & Nigel Michki

Introduction

Pneumonia is a common and morbid pediatric illness. The World Health Organization estimates pneumonia accounts for 1 out of 6 deaths among children under 5 years old and it is the second leading cause of death in this population.[1] In the US pneumonia costs more than 1 billion dollars, leads to more than 150,000 hospitalizations, and causes more than 1,000 child deaths each year.[2-4] In order to further reduce this mortality rate we must better focus efforts on those children most at risk for death. With that in mind we have developed a model to predict death in patients hospitalized for pneumonia under the age of 21.

Methods

The dataset we utilized was the Healthcare Cost and Utilization Project (HCUP) Kids' Inpatient Database (KID) dataset. HCUP KID is a collection of administrative data concerning inpatient hospitalizations for patients under the age of 21.[5] HCUP KID represents the largest nationally publically available all-payer pediatric inpatient care database.[5] We specifically used the HCUP KID data from 2006, 2009, and 2012.

Variables

We used International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnostic codes [ICD-9-CM codes480.0-2,480.8-9,481,482.0,482.30-2,482.41-2, 482.83, 482.89-90, 483, 484.3, 485, 486, and 487.0] to identify children with a principal diagnosis of pneumonia.[6] We defined mechanical ventilation using ICD-9-CM procedure codes (96.70, 96.71, 96.72).[7] Number of organ failures was identified as previously described for septic patients by Angus et al., and comorbid conditions were identified as previously described by Feudtner et al. [7, 8]

Model Evaluation

We chose to evaluate our model using accuracy, which is simply the number of correct classifications over the number of total classifications. We also measured accuracy using F1. F1 is the harmonic mean of recall and precision or a combined measure of the rate of false positives and false negatives. This is because recall measures the true positive rate over the sum of the true positives and false negatives. Alternatively, precision measures the true positive rate over sum of true positives and false positives.

	TEST	TRAIN	p
n	3895	5619	
n_Organ_Failures (mean (sd))	1.59 (0.86)	1.58 (0.86)	0.615
n_Comorbidity (mean (sd))	0.81 (0.93)	0.81 (0.92)	0.897
n_diagnoses (mean (sd))	14.07 (6.54)	14.09 (6.52)	0.892
Operating_Room_Procedure (mean (sd))	0.25 (0.43)	0.27 (0.44)	0.086
Transferred_In (mean (sd))	0.30 (0.46)	0.29 (0.45)	0.113
Admitted_Weekend (mean (sd))	0.26 (0.44)	0.27 (0.44)	0.467
Month_of_Admission (mean (sd))	6.32 (3.74)	6.24 (3.71)	0.343
Childrens_Hospital (mean (sd))	0.32 (0.47)	0.31 (0.46)	0.390
n_Hostpital_Beds (mean (sd))	2.64 (0.61)	2.67 (0.59)	0.040
Mean_Income_of_Zipcode (mean (sd))	2.21 (1.14)	2.22 (1.15)	0.690
Pay_Type (mean (sd))	1.50 (0.65)	1.52 (0.65)	0.177
DIED (mean (sd))	0.07 (0.26)	0.07 (0.26)	0.578
Test_or_Train = TRAIN (%)	0 (0.0)	5619 (100.0)	<0.001

Table 1. Variable Analysis.

For our analysis, we chose 2 models: K-Nearest Neighbor and Random Forest. K-Nearest Neighbor is a “lazy” learner that doesn’t require a model, but is still simple and accurate. The algorithm identifies k other instances based on distance functions, and chooses the majority label (probabilistic view). We chose this model in order to determine what, if any, clusters of factors might give rise to high death rates among patients suffering from pneumonia. Is there one core, central region in the factor-space that can account for high death rates (implying that certain occurrences naturally lend themselves to high death rates), or are there multiple smaller clusters (implying that death is somewhat stochastic in this regard)? While we did not perform this analysis in any great detail, the model does lend itself to such an analysis and would be interesting to pursue further.

Random forests generate multiple uncorrelated classification trees. The chosen training set for each classification tree is chosen based on sampling with replacement.[9] Once multiple classification trees have been generated, random forest counts the “votes” for the classifications, the probability of each outcome, and chooses the prediction with the highest probability.[9] Some advantages of random forest and the reasons why we chose this model include that it avoids overfitting, it is fast with large datasets, and it contains inherent cross-validation.[9] Furthermore the concept of ‘splitting on factors’ is analogous to the process by which physicians incorporate evidence into their own diagnoses, and

while random forests do not generate any one single decision tree, the decision making logic is similar and natural in this context.

Missing Values

After subsetting the data to only include those variables in which we were interested, we discovered that some observations still contained missing values. However preliminary analysis (table 2) showed that only a small fraction of observations where patients died contained any missing values, and so it was determined that exclusion was the most straightforward and justified manner in which to deal with these cases.

	Percent (number) of cases with deaths
Filtered data with NAs *included*	7.36% (703 out of 9514)
Filtered data with NAs *omitted*	7.32% (696 out of 9514)

Table 2. Missing values analysis.

Data splitting

In order to generate 'fair' training and testing data sets from our bulk, we chose to split our data as depicted in figure 1. In essence, observations in which patients died (a rare event) were selected out of the bulk, and then both these and those observations in which patients lived were split into ⅔ train, ⅓ test sets. The alive-train and dead-train sets were then combined, and likewise for the test sets. This ensured that during the testing phase observations in which patients died would included, and so performance metrics such as F1 would have meaning (as without this slightly complicated division of observations, the test set would often have so few observations of patients that had died that the true-positive rate would be 0).

Furthermore, in order to judge how well our models might perform should the relative rarity of observations in which patients died be increased, we trained and tested our model on split data sets which had 'living' observations removed during both phases. This subsetting was done in the hope of improving performance, but our approach was flawed in that *testing* on such a subsetted dataset is not valid (what good is a model that is capable of classifying data that is not as rare as in the real world?).

In order to estimate out-of-sample performance, we opted to perform 10-fold cross validation on the *training* dataset and measure the average performance of those models before performing a similar analysis on the *test* dataset.

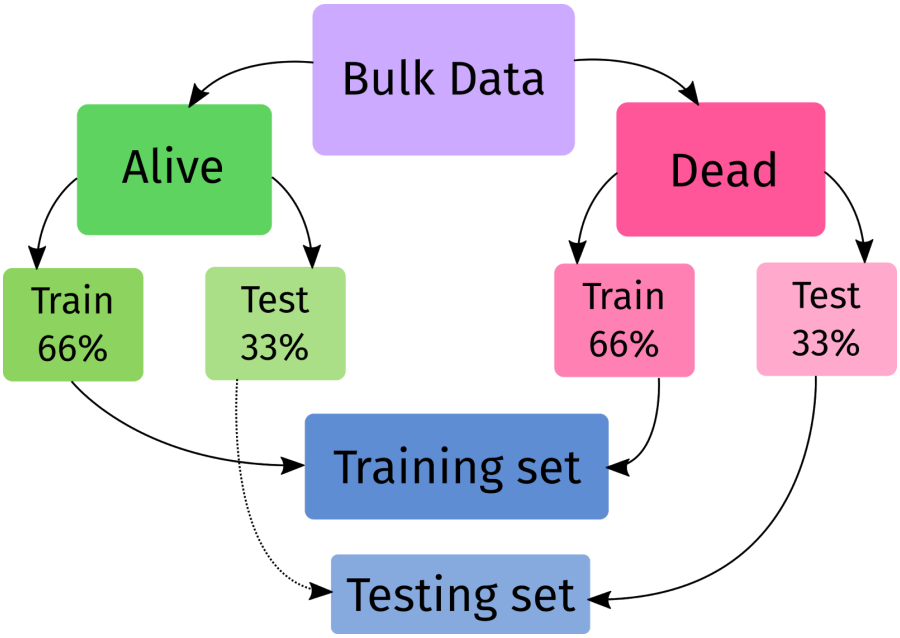

Figure 1. Distribution of dataset into training and testing set.

Results

As stated previously, our choice to subset *both* the train and test data was flawed, and thus our performance metrics for all 'sample fractions' other that 1/1 are not reflective of real world performance.

Sample fraction (alive)	**% Dead**	**Accuracy [Train, CV, Test]**	**F1 [Train, CV, Test]**
1/1	7.3	94.8, 93.0, **91.6**	44.8, 9.9, **14.4**
1/2	13.6	92.0, 85.1, **84.9**	58.8, 21.9, **21.1**
1/4	24.0	88.7, 75.2, **77.6**	71.4, 34.5, **41.7**
1/10	44.1	87.8, 63.4, **66.6**	85.3, 55.5, **60.8**

Figure 2. K-Nearest Neighbor Results.

Sample fraction (alive)	**% Dead**	**Accuracy [Train, CV, Test]**	**F1 [Train, CV, Test]**
1/1	7.3	99.1, 92.2, **91.6**	98.1, 5.1, **5.6**
1/2	13.6	99.7, 84.9, **85.2**	99.0, 15.9, **25.2**
1/4	24.0	99.8, 77.6, **77.1**	99.7, 44.3, **43.5**
1/10	44.1	99.7, 68.5, **68.7**	99.7, 62.4, **64.8**

Figure 3. Random Forest Results.

It is clear, however, that in this very basic model training exercise using default parameters for the so-called 'learners', the *random forest model underperforms the k-nearest neighbors model at the full (original) sample fraction*, having similar accuracies on the test set but an improved F1 score, implying that the recall and sensitivity of this model is superior. However to qualify that statement the 10-fold cross-validation F1 score for the k-nearest neighbors model does not accurately estimate the model's eventual performance on the testing data, which is in stark contrast to the cross-validation estimate for the random forest model. Noting this, it is reasonable to presume that *perhaps the k-nearest neighbors approach is less stable than the random forest*, and may merit further study. For example changing the number of decision trees that grow in the forest and likewise varying k, the number of neighbors required for weighting our model's decision, will serve as ways to better understand these models' performance on such a vast real-world dataset.

Conclusion

The most jarring result of this analysis was discovering how poorly *both* models performed on this data set in order to predict the rare outcome of death among patients with pneumonia. Why the F1 scores of these models are so low is a question worth studying further; we propose a few possible hypotheses:

1. The predictors selected for this model were not sufficient for segmenting which patients might die
2. *Too many* predictors were used in order to determine which patients might die
3. The occurrence of death amongst patients with pneumonia is semi-random: *any* given set of factors is insufficient for determining whether or not a patient might die during their hospital stay.

The first two of these are fairly straightforward, though tedious, to test. Performing a rigorous regression analysis on all of the available predictors will better inform us to select correlated but non-redundant predictors for model training. The last hypothesis, that these deaths are semi-random, is more challenging to test. Pneumonia deaths are

thankfully a rarity in the United States, and so it may well be the case that randomness plays a significant role in (preventing) the prediction of death among hospital patients that contract the disease.

References

1. Liu L, Oza S, Hogan D, et al.: Global, regional, and national causes of under-5 mortality in 2000-15: an updated systematic analysis with implications for the Sustainable Development Goals. Lancet, 2016
2. Jain S, Williams DJ, Arnold SR, et al.: Community-acquired pneumonia requiring hospitalization among U.S. children. N Engl J Med 372:835-845, 2015
3. Lee GE, Lorch SA, Sheffler-Collins S, et al.: National hospitalization trends for pediatric pneumonia and associated complications. Pediatrics 126:204-213, 2010
4. Kochanek KD, Murphy SL, Xu J, et al.: Deaths: Final Data for 2014. Natl Vital Stat Rep 65:1-122, 2016
5. Healthcare Cost and Utilization Project (HCUP) Kids' Inpatient Database (KID). Rockville, MD, Agency for Healthcare Research and Quality, 2012
6. Leyenaar JK, Shieh MS, Lagu T, et al.: Variation and outcomes associated with direct hospital admission among children with pneumonia in the United States. JAMA Pediatr 168:829-836, 2014
7. Angus DC, Linde-Zwirble WT, Lidicker J, et al.: Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit Care Med 29:1303-1310, 2001
8. Feudtner C, Christakis DA, Connell FA: Pediatric deaths attributable to complex chronic conditions: a population-based study of Washington State, 1980-1997. Pediatrics 106:205-209, 2000
9. Brieman L, Cutler A. Random Forests. Random forests. https://www.stat.berkeley.edu/~brieman/RandomForests/cc_home.htm. Accessed December 13, 2016.