CART Analysis: Difference in Symptom Manifestation in Child and Adult Dengue Patients

Saran Kunaprayoon

Lehigh University

Follow this and additional works at: https://preserve.lehigh.edu/undergrad-scholarship-eckardt

Part of the Virus Diseases Commons

Recommended Citation
https://preserve.lehigh.edu/undergrad-scholarship-eckardt/16

This Article is brought to you for free and open access by the Undergraduate scholarship at Lehigh Preserve. It has been accepted for inclusion in Eckardt Scholars Projects by an authorized administrator of Lehigh Preserve. For more information, please contact preserve@lehigh.edu.
CART Analysis: Difference in Symptom Manifestation in Child and Adult Dengue Patients
Saran Kunaprayoon
Eckardt Scholars Honors Thesis
Advisor: Jill Schneider
Department of Biological Sciences
Lehigh University
Bethlehem, PA 18015
May 19, 2017
Introduction

Dengue is one of the many mosquito born diseases, belonging to a class of flaviviruses. Other viruses that also belong to this class are Zika and Japanese Encephalitis (JE). *Aedes aegypti* female mosquitoes serve as vectors that transmit these viruses from individual to individual. As expected Dengue fever is endemic in many warm places such as Southeast Asia, South America, and warmer parts of Africa. Recently, dengue has spread to other warmer regions in North America such Florida, Texas, and Hawaii. In 2012, dengue rose by 70% in the United States, most likely attributed to rising global temperatures (*Thanks, Global Warming--Mosquito-Borne Diseases Are on the Uptick*, 2013).

In addition to dengue’s spread throughout the world, the demographics of people affected with dengue have also changed. Original known as a pediatric disease, in the past two decades, dengue has penetrated the older population. The result has been an increase in more severe dengue secondary infections. Primary infections (first contractions) are usually less severe, even asymptomatic at times. Secondary infections (second time contracting dengue) are usually much more severe because of antibody cross reactivity. Given that dengue has four serotypes, there is a relatively high likelihood of contracting other dengue serotypes even though one may have contracted a primary infection and subsequently became immune to that specific serotype (Hasan, Shamimul et al., 2016). However, it is important to note that a dengue secondary infection does not constitute that the patient has contracted dengue for the second time. It can also mean that the individual has had other flaviviruses infected them before. For
instance, if an individual had the JE vaccine as a child and later contracted dengue for the first time, serology tests would suggest that he/she had a secondary infection.

As dengue becomes more prevalent in adults, studies show that adult and child dengue patients exhibit different symptoms. There has been research done to identify unique symptoms that disproportionately manifest in each age group. Hanufasa S., et al., (2008) conducted a study in Rayong, Thailand found that headache and myalgia were more common in adults while coughing, vomiting, abdominal pain and rash were more common in children. With almost identical findings, a similar study in Brazil found that myalgia, retro-orbital pain, nausea, arthralgia were more common in adults while vomiting and skin rash were more common in children (Souza, L. J., et al, 2013). Both studies found similar patterns of symptoms that manifested in adults and children. They both found myalgia to be more common in adults and vomiting and skin rash to be more common in children. Perhaps much of this could be related to the fact that adults develop secondary infections while children develop symptoms from primary infections. Despite identifying these symptoms as more common in a certain age group, there has not been a model to predict whether a set of symptoms can clearly identify the age group of a dengue patient.

My research aims to develop a satisfactory model that predicts whether the dengue patient is an adult or child based on a set of symptoms. This model can aid doctors in differentiating symptoms between adult dengue from child dengue patients.

There are many types of models that one can build. Specifically for this analysis, I built a decision tree. The reason I used a decision tree is so that doctors can input the
symptoms of a patient and determine visually based on the model whether the patient is a child or an adult.

The main question here is if adult and child dengue patients consistently manifest different symptoms and if the same set of symptoms appear in each class is consistent enough for a model to accurately predict whether a dengue patient is a child or adult based on their symptoms. My hypothesis is that the model that I develop will be able to accurately predict whether a dengue patient is a child or an adult since adults and children tend to exhibit different symptoms.

To evaluate this model, I will use evaluation measures, including an ROC curve, area under the ROC curve, accuracy, precision, f-measure, and recall.

Methods

I acquired a dataset used in this study through Armed Forces Research Institute of Medical Sciences (AFRIMS). The dataset has a collection of approximately 7,000 potential dengue patients in Kamphaeng Phet, Thailand from 2007-2015. Included in the dataset are PCR results, serology test results, age, location of blood collection, and various symptoms manifested in those patients.

To analyze all this data, I used RStudio. First I determined which patient had dengue. PCR results were the main indicator of whether a patient had dengue. If PCR was positive, then there would be no reason to consider serology. If PCR was negative, then serology results were used to determine whether the patient had a dengue infection. If the patient had a dengue positive serology result, then I classified the patient as dengue
positive. Dengue positive patients were grouped as 1 while dengue negative patients were grouped as 0.

Removal of “N/A”s was done only rows with really sparse data. Another important classification is to determine whether a patient was a child or adult. Adults were classified as >15 years of age. The flow chart below describes how I classified patients in the dataset.

To analyze the data, I generated a boxplot, showing the average age of dengue patients from 2007-2015. I then generated a histogram to display the proportion of dengue serotypes by year. To analyze the differences in symptoms between adults and children, I filtered the dataset for dengue patients only; then I made histograms showing the proportion of children with certain symptoms in relation to adults. This was done for cough, muscle pain, eye pain, headache, skin rash, nausea, vomiting, diarrhea, abdominal pain, and joint pain. I analyzed these symptoms specifically because Souza, L. J., et al., (2013) and Hanufasa S., et al., (2008) found age bias in these symptoms.
After preliminary symptom analysis, to predict the difference in symptoms between adults and children, I used a classifier called a decision tree. I also used 10 fold cross-validation to ensure generalizability of the model. Once this was done I evaluated the model with confusion matrices. With the confusion matrix, I made Receiving Operating Characteristic (ROC) curves, calculated accuracy, precision, f-measure, recall and area under the ROC curve.

Results

![Box plot showing age of diagnosis vs year](image)

*Figure 1-Median age of all dengue patients each year from 2007-2015. The median age is stable throughout.*

In Figure 1 and 2, we observe a median age range of 13.5-17 from 2007-2015.
Although the median age is stable across the years of analysis, the standard deviation has increased from 8.15 to 12.1.

<table>
<thead>
<tr>
<th>Year</th>
<th>Age</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>17</td>
<td>8.151346</td>
</tr>
<tr>
<td>2008</td>
<td>13.5</td>
<td>8.127164</td>
</tr>
<tr>
<td>2009</td>
<td>14</td>
<td>10.35704</td>
</tr>
<tr>
<td>2010</td>
<td>16</td>
<td>9.28804</td>
</tr>
<tr>
<td>2011</td>
<td>14</td>
<td>9.836988</td>
</tr>
<tr>
<td>2012</td>
<td>16</td>
<td>12.17346</td>
</tr>
<tr>
<td>2013</td>
<td>15</td>
<td>10.21516</td>
</tr>
<tr>
<td>2014</td>
<td>14</td>
<td>11.74059</td>
</tr>
<tr>
<td>2015</td>
<td>16</td>
<td>12.10503</td>
</tr>
</tbody>
</table>

Figure 2-Median age of dengue patients each year. Though the median age is stable throughout 2007-2015, the spread has been increasing.

Figure 3-There are 4 serotypes in this histogram: Dengue-1,2,3 and 4. The histogram displays the proportion of each serotype by year.
Figure 3 displays the proportion of each Dengue serotype by year. Dengue-1 was the predominant strain affecting dengue patients in 2007. From 2008-2012, we see a general increase in Dengue-2. Starting 2013, we observe a sudden increase in Dengue-3. Dengue-3 levels continue to rise until 2015 at which point we see a sudden increase in Dengue-4.
In Figure 4, I developed histograms to visualize the proportion of adult and dengue patients with certain symptoms. In this figure, I included the following symptoms: joint pain, diarrhea, cough, skin rash, nausea, abdominal pain, headache, vomiting, muscle pain, and eye pain. Based on the histograms above, joint pain, skin rash, nausea, headache, muscle pain, and eye pain affected more adults than children. Diarrhea and cough affected both children and adults equally. Lastly, abdominal pain and vomiting affected more children than adults, although not by much.

Figure 4- Histograms of symptoms exhibited in adult vs. dengue patients.

In Figure 5, a decision tree is used to determine whether the dengue patient is an adult or child based on symptoms. TRUE represents the proportion of adults in the prediction and FALSE represents the proportion of children in the prediction.

Figure 5- Decision Tree to determine whether the dengue patient is an adult or child based on symptoms.
Figure 5 displays one of the decision trees generated through 10 fold cross validation. “TRUE” and “FALSE” at the end of the nodes represent “Adults” and “Children” respectively. This CART model was developed based on the training data, nine out of the ten segmented portions of the dataset. The sequence of symptoms demonstrates the proportion of adults and children with that specific set of symptoms.

<table>
<thead>
<tr>
<th>prediction</th>
<th>FALSE</th>
<th>TRUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>FALSE</td>
<td>194</td>
<td>99</td>
</tr>
<tr>
<td>TRUE</td>
<td>48</td>
<td>106</td>
</tr>
</tbody>
</table>

Figure 6 is an example of a confusion matrix taken from the first cross-validation. “Prediction” is the model’s predictions whether the patient is a child or adult based on their symptoms. These predictions were matched against true classes of child or adults in the test set. In this confusion matrix, we observe many false negatives; true values that were predicted as false. For instance, there are 106 true predictions, 99 false negatives, 194 true negatives, and 48 false positives.

Figure 7 - The Receiving Operator Characteristic (ROC) curve plots the true positive predictions against the false positive predictions.
The ROC (Receiving Operator Characteristic) Curve plots true positive predictions against false positive predictions. The ratio of true positives to false positive is approximately 2.

<table>
<thead>
<tr>
<th>Evaluation Measures</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.66</td>
</tr>
<tr>
<td>Precision</td>
<td>0.67</td>
</tr>
<tr>
<td>Recall</td>
<td>0.72</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.69</td>
</tr>
<tr>
<td>Area Under the Curve</td>
<td>0.70</td>
</tr>
</tbody>
</table>

*Figure 8-This table displays the average accuracy, precision, recall, f-measure, and the area under the curve of the AUC curves. This average is obtained by doing a 10-fold cross validation.*

With the confusion matrix, numerous evaluation measures were employed to determine the predictive ability of the model. Figure 8 displays the average of all evaluation measures. The average accuracy, precision, recall, f-measure, and area under the ROC curve are 0.66, 0.67, 0.72, 0.69, and 0.70. All of these evaluation measures were calculated from the average of all confusion matrices.

**Discussion**

Based on my data analysis, median age of infection from 2007-2015 hovered around 14-17, although the standard deviation has increased from ~8 to ~12 overtime.
Simmons, C. P., & Farrar, J., (2009) found the median age of dengue infections in Thailand increased from 13-17 from 1999-2005. Since 2005, my analysis suggests that the median age has stayed in the same range. However, the spread increased throughout the years, suggesting that a greater proportion of adults are contracting the disease. What was known as a pediatric disease has become a more widespread phenomenon among adults.

As more adults become infected, there is a greater likelihood of developing dengue secondary infections by nature of probability. There are two types of dengue fever: apparent and inapparent. Inapparent dengue patients have the virus, but exhibit no symptoms. Apparent dengue fever is the opposite of inapparent. Thus when one has inapparent dengue, the immune system still produces antibodies against the dengue virus. As antibody levels build up, the body develops resistance against the serotype that infected the person. When the individual again contracts another strain of the Dengue virus, they have higher chances of developing secondary infections. When a secondary infection occurs, the body overreacts and produces excess antibodies, enhancing the inflammatory response in the body, describing a process called Antibody-Dependent Enhancement (ADE). During this antibody overdrive, the body produces non-neutralizing antibodies, which bind to the virus RNA without deactivating them (Goncalvez et al., 2007). As a result, these non-neutralizing antibodies carry these infectious particles throughout the body, into different tissues. The end result is that there is increased inflammation around tissues in which Dengue viruses were transported.

Also equally important is the shift in serotype trends from 2007-2015. In 2007, we observe predominantly Dengue-1 presence. However, we observe a quick shift to
Dengue-2 from 2008-2012. Once again we see another serotype shift to Dengue-3 from 2013-15 with some rising Dengue-4 in 2015. Fried, Jessica R, et al., (2010) found Dengue 2 and 3 to be associated with Dengue Hemorrhagic Fever, which is a severe form of Dengue Fever. However, Dengue-1 is associated with milder forms of Dengue. We might suspect here that the severity of dengue infections has increased throughout the years in our sample. At the same time, the number of adult dengue patients has also increased, suggesting potential link between age and serotype. We suspect that Dengue-2 and 3 mutate much more quickly than other strains, making them much more virulent and likely to infect people who may have had inapparent dengue as children. This area is research largely unknown and genomic analyses of dengue serotypes need to be done. The trend towards greater presence of Dengue-4 in 2015 is intriguing and is worth following to see whether Dengue-4 will develop into a virulent strain as well in the coming years.

With increasing age diversity in dengue patients, adults and children manifest different symptoms. Studies conducted in Brazil and Thailand found headache, myalgia, retro-orbital pain, nausea and arthralgia to be more common in adults. Symptoms predominately found in children are cough, vomiting, abdominal pain, and skin rash. (Hanafusa S., et al., 2008 & Souza L. J., et al., 2013). My analysis, however, only partially supports these findings. Not only did joint pain, nausea, headache, muscle pain, and eye pain affected more adults than children, but skin rash was also more prevalent in adults than in children. Yet based on Hanafusa S., et al., 2008 & Souza L. J., et al., (2013), skin rash should have been the predominant symptom in children. In terms of typical child dengue symptoms, in accord with literature, my analysis suggests that
abdominal pain and vomiting are more common in children. Coughing, which was supposed to be a pediatric symptom affected children and adults equally in my analysis. This could be the result of regional differences.

To further determine the difference in symptoms between children and adults, I made a decision tree to predict whether the patient is an adult or child based on their symptom profile. I then evaluated the model and made a confusion matrix, shown in Figure 6. Since there are 100 true predictions to 99 false positive predictions, which is not ideal, we know that there is some overfitting in the model. In other words, the model in Figure 5 is not able to make generalized predictions on the test set, and incorrectly classifies an adult more than it should have.

Figure 7 further evaluates the model by plotting true positives against false positives. Ideally, we want a steep curve, suggesting a high ratio of positive rate to false positive rate. However, we observe here a relative flat curve. Looking at the evaluation measures more closely, Figure 8 displays the accuracy, recall, precision, f-measure, and AUC. Our accuracy is 0.66 which is rather low, as well as our precision 0.67. However, our recall is 0.72, which should not be discounted. A higher recall suggests that model is better at minimizing false negatives than false positives. A false negative in this case would be if the actual class were adult, but the model predicted that the patient was a child; while a false positive would be the opposite. Having a high recall works to our benefit because we are more interested in adults in the first place. Our AUC value was 0.7, which was not bad either. AUC values can range from 0-1. Ideally for us to definitively use a set of symptoms to classify adults or children, we would need an AUC value of 0.95 or more.
This project is far from complete, and in the future, I would use other classifiers to predict the classes. Other classifiers that may be useful for predicting age group are Naïve Bayes and linear regression. If those two classifiers can better predict our classes, then it most likely means that our model is not the best model for this classification task, and that there was overfitting in our model. However, if other models did not perform better than our current decision tree, then perhaps the difference in symptoms between adults and children was not that significant. After all, there could be too much variability in how symptoms manifest in adult and child dengue patients.


