Post-Mortem

of

Capstone Project

By:

Wilson Aleman

Colin Beran

Scott Cummings

Rishi Paul

Allen Tung

ISDS 577 - Capstone Project

Spring 2020

California State University, Fullerton

**Table of Contents**

[**Overview**](#_u0xenss65h1w) **2**

[**Expected Results**](#_9gz66kwajvml) **2**

[**Actual Results**](#_3d1ppqx7oen) **3**

[**What went wrong?**](#_xovesjb2tt3f) **5**

[**Conclusion**](#_f3sl98dh1890) **6**

# 

# Overview

Three questions were answered to decision-makers about how to manage highly contagious diseases: Does a location's economic situation affect the rate of immunization among school children? Does Diphtheria, Tetanus, Pertussis (DTP) vaccine exemptions affect case rates in California? Which California counties and schools have the highest risk of facing a Pertussis outbreak?

# Expected Results

**Question One**:

We anticipated that analysis would reveal that economically prosperous areas would have a higher rate of immunization among its schoolchildren when compared to impoverished and less fortunate areas. The logic behind this notion is that persons living in low-income areas would have less access to health care, be less informed about the benefits of immunization, and in general be more worried about the day to day insecurities that come with poverty than with obtaining preventative health measures that may never bring any immediate and clear benefits. The initial analysis revealed that the relationship between the variables was more complicated than expected. Therefore, we explored the interaction between additional socio-economic measures: population and education level.

**Question Two:**

To answer question 2, our group focused on Pertussis rates because we had case rates for this disease for several years. The first assumption related to exemptions that we had was that urban areas had higher levels of vaccine exemptions than rural areas. Urban areas were expected to be more receptive to Anti-Vaccination beliefs. The next assumption we made was that Urban areas assuming they had higher vaccine exemptions had, as a consequence, had higher Pertussis case rates. The last assumption we had before analyzing the data was that exemption levels show an increase over time.

**Question Three:**

When we proposed question three, we expected to be able to use the data that we found to accurately predict which counties and which schools in those counties would be at the highest risk of a pertussis outbreak. Because the StudentData dataset contained information, including school name and whether the school was a private or public institution, we had expected to be able to make use of that information. We had also expected to find that the more populated counties and schools would be the ones naturally at the highest risk for an outbreak.

# Actual Results

**Question One**:

We did not succeed in demonstrating the relationship between an area’s economic situation and its student immunization rate. Nor were we able to definitively establish the effect of education levels and population on immunization. What we did reveal was that the influences on student immunization rates are multifaceted and likely very dependent upon individual household circumstances and prejudices. Examining the data from the state and county level would not be sufficient to flush out the root causes of adherence or disregard of immunization for school children.

**Question Two:**

The results showed that the first assumption we made urban counties having higher exemption levels than rural counties was wrong. The data proved that if normalizing the exemptions by the number of students in a school, then rural counties have higher levels than in urban counties. Normalizing exemption data was a critical turning point for answering question 2 because urban data consistently overshadowed rural data. The other assumption about exemption levels having increased overtime in California was correct. The graph of exemption counts shows a steady increase from 2000 to 2013 that only decreased in 2014. We did not anticipate the decline in exemptions due to the 2014 Pertussis outbreak. The data, however, did not show an apparent negative effect of high vaccine exemptions on case rates. Counties with an average percentage of exemption below 10% were affected by the outbreak based on factors other than exemptions. Nevertheless, we did discover that counties with a high average percentage of exemption, like Napa (27%), had appalling case rates. We were able to conclude that low exemptions help manage diseases, but it is not a single solution.

**Question Three:**

The results of the regression model did not meet expectations. Based solely on the data that we had at our disposal, we were correct in our assumption that the county’s population of kindergarten students would make the most substantial impact in determining the number of cases. The counties of Los Angeles and San Diego consistently having the highest number of documented pertussis cases confirms these results. However, a few outlier counties featuring high numbers of cases despite low populations. Most notable of which being Fresno county in 2010.

# What went wrong?

**Question One:**

We were only able to find complete and publicly accessible socio-economic data at the county level; this was a hindrance as the analysis we performed would have likely worked much better at a finer granularity - either at the city/township or zip code level. Examining variables at the county level can obscure important details, particularly in diverse areas such as Los Angeles County and Alameda County that have dozens if not hundreds of subcommunities and enclaves within their borders. Additionally, regression models typically perform better with larger sample sizes.

**Question Two:**

The student data from Kaggle had immunization and school data, but no case rate at the school. We were able to obtain the case rate at the county-level for Pertussis only from another dataset. It would have been preferable if the schools reported the number of pertussis cases themselves. Another issue with the student data and county-level case rate data from Kaggle was that some counties with low population had missing data. We concluded that high exemptions, like for the Napa county, cause a higher degree of case rates. However, at the same time, we did not see that counties with low exemptions had, on average better case rates. It seems like exemptions are only one of the variables that explain how a disease might spread.

**Question Three:**

Our best regression model featured an R-squared value of only .527, which was far lower than we had expected. We had concluded that population size made the most significant impact on case numbers, which was, for the most part, correct, however having outliers left us to believe that other significant factors were not previously measured. Additionally, because our data on the number of cases did not specify what schools the students attended, we were unable to make any use of the private vs. public institution factor, which could have proved to be relevant.

# Conclusion

**Question One:**

In retrospect, the scope of performing a student immunization analysis on the entire state was too ambitious. Performing a similar review on one unusually large county or cluster of locations that have more accessible and granular sub-community-level data may have been a more manageable endeavor. A reduced scope and a more focused approach with better data would have probably improved the fit of our regression models and resulted in more precise patterns in the visualizations. Using these enhanced models would have likely resulted in getting closer to understanding the root causes of low student immunization.

**Question Two:**

Investigating the variables that affect disease is a complicated task. A single variable like exemptions cannot explain by itself if a county has higher or lower case rates. The main difficulty we experienced when answering question two was that one dataset was for infant children the other for schools, and the third was a county-level pertussis case rate dataset. Working with these datasets that covered different segments of the population made the task of answering question 2 more challenging. It would have been necessary to find a dataset that would combine both school data and case rates to facilitate this task. For example, county hospitals that keep a record of the age of the children, school name, and vaccination records, then it would have been more natural to find any correlation between exemptions and case rates. Nevertheless, we did find that exemptions at a higher percentage, like in the case of Napa (27%), were cause for appalling case rates.

**Question Three:**

Given the opportunity to improve on the gathered data, we would have first and foremost broken down the number of cases by the school they originated from rather than by county. With this information, we would have been able to find out if the institution is a private school or a public school would have been a statistically significant factor. Initially, we also wanted to find out which schools were more at risk for a pertussis outbreak, having the cases broken down by school would have made finding this information possible. Gathering more information from each school likely would have improved the data as well. Relevant information could include the precautionary measures taken when students began to show signs of infection. Having access to more information on each county and all of the schools likely would have improved the fit of our regression model and led to a far more accurate analysis.