Tristan Almada

Final Project Design Write-Up

Identifying Local-Level Barriers to HPV Vaccination

General Assembly Data Science Course

### Project Problem and Hypothesis

\* What's the project about? What problem are you solving?

The Human Papillomavirus, or HPV, causes 5% of all cancers, with rates increasing every year. Meanwhile, the standard of care for advanced HPV-associated cancers treatments remains the same as it was in the 1970s, meaning harsh outcomes for those that are diagnosed. Ten years ago the first HPV vaccine was introduced to cover the majority of cancer-causing HPV strains. Yet, despite of this monumental opportunity to rid the world of a ubiquitous cancer-causing virus, HPV vaccine rates in the USA remain the lowest in the developed world at 45% for girls and 28% for boys (compare with Australia at 80% and 75% respectively). What are the specific barriers that stand in the way of higher vaccination against HPV? What county and local-level data point predict vaccination rates? What counties have adjusted policies to improve vaccination rates, providing a replicable model for the rest of the country?

\* Where does this seem to reside as a machine learning problem? Are you predicting some continuous number, or predicting a binary value?

I aim to predict vaccination (1 = vaccination, 0 = not vaccination).

\* What kind of impact do you think it could have?

The life expectancy on Earth was 30 in the year 1900. Today, it is close to 70, thanks largely to improvements in sanitation and biomedical innovation (anti-biotics, vaccines). Why is there resistance to HPV vaccine, especially when HPV is found in 80% of sexually-active people (i.e. it is relevant to the majority of the population)?

Anti-vaccine sentiment may not be logical but it is understandable. By definition, there is little to no media coverage of the lives that are saved by vaccines (think small pox, measles, rubella, diphtheria, typhoid, yellow fever, polio – household names a century ago) on a given year. On the other hand, rare, isolated, unsubstantial case reports highlighting HPV vaccine side effects (often without scientific causality or evidence) make their way into sensationalist, mainstream press, causing parents to re-evaluate the risk/reward balance.

The impacts of universal vaccination against HPV are substantial. Over 20 million cases of cancer a year could be prevented, along with 100s of millions of cases of “pre-cancer” (requiring invasive diagnostics and treatments) and anogenital warts. The United States spends 94% of its healthcare budget on care versus only 6% on prevention. A small increase in prevention spending could have a massive impact on the overall healthcare budget and help would-be sufferers live longer-more product lives. It is hard to see how universal prevention against cancer is not in everyone’s interest.

My hope is that this study will identify specific strategies that have succeeded in improving vaccination rates on a local, state, and federal level. If the variables are credible, I will present the study to CDC / ACIP with the goal of emphasizing “what works” in improving vaccination rates in areas which still have resistance.

\* What do you think will have the most impact in predicting the value you are interested in solving for?

Provider recommendation should be a primary determinant of parental consent to HPV vaccination.

### Datasets

\* Description of data set available, at the field level (see table)

**Data Dictionary:**

\* If from an API, include a sample return (this is usually included in API documentation!) (if doing this in markdown, use the javacription code tag)

\* What experience do you already have around this area?

In 2010, my mother passed away from HPV-associated cancer at the young age of 53. She lived for just two years after diagnosis. My sisters and I set up a foundation called The HPV and Anal Cancer Foundation with the stated aim of solving the problem – eliminating 5% of all cancers – and make ourselves obsolete in the process. In 2011, my sisters and I presented a compelling case to the CDC to expand the vaccination program to include boys in the USA. They agreed and starting in 2013 young adolescent boys were included in the USA program. Vaccine rates are now 23% in boys (from 0% when we started) and rising.

\* Does it relate or help inform the project in any way?

Yes, it is as pertinent as it gets.

\* What other research efforts exist?

The CDC has done this work on a national level and updates their analysis once a year. This will provide me a framework for analyzing county-specific data. Furthermore, numerous epidemiological academics publish data on trends in HPV-associated cancer rates and also present hypothesis for why vaccine rates are low. Much of the analysis is done on sample data (phone surveys on a state level). However, more counties are publishing their own “full” data sets which is where I want to focus my research.

### Project Concerns

\* What questions do you have about your project? What are you not sure you quite yet understand? (The more honest you are about this, the easier your instructors can help).

I am concerned about categorizing some of the important variables, such as “provider recommendation.” I am concerned that the county-level data (which has the output variable, vaccine / no vaccine) will not provide enough data breadth to test all of my various hypotheses.

\* What are the assumptions and caveats to the problem?

Variables: socioeconomic status, demographics, public-education requirements for school, clinical recommendation, patient outreach/engagement strategies, mass media campaign, social media?

\* What is already implied about the observations in your data set? For example, if your primary data set is twitter data, it may not be representative of the whole sample (say, predicting who would win an election)

Clinician recommendation, state-level mandates appear to be major drivers

\* What are the risks to the project?

Risks are that I will spend too much time cleaning data. I may decide to focus on just ONE county or state that has the highest resolution data. Another risk is that I miss one of the many “unknown unknowns” that go into the mosaic of parental mistrust of pharmaceutical companies and their products.

\* Is any of the data incorrect? Could it be incorrect?

There appears to be a large error in the sampled data that the CDC conducts (roughly 200 phone calls per state).

### Outcomes

\* What do you expect the output to look like?

I expect it to be very messy and the “fit” to be poor amongst the variables I choose. However, I hope to see a table of odds ratios showing how “clinician recommendation”, for example, improves the likelihood of HPV vaccination.

\* What does your target audience expect the output to look like?

Epidemological tables.

\* What gain do you expect from your most important feature on its own?

Broadly, a strong recommendation from a provider would increase the chance of vaccination by 50%.

\* How complicated does your model have to be?

It could be fairly complicated – perhaps modeling parental decision to vaccinate as a decision tree could be a decent way to map out the behavior, even if ultimately I choose another modeling methodology.

\* How successful does your project have to be in order to be considered a "success"?

Any fit would be impressive.

\* What will you do if the project is a bust (this happens! but it shouldn't here)?

Examine the causes of the “bust” – I am fairly certain having all the data will be the major barrier. However, if I have all the data and I still cannot identify a compelling fit, it would point to other factors that are nevertheless worth thinking about (for example, a national mass media campaign).

\* Convert ‘goal metric’ to something non-data people can understand.

Cost-benefit analysis is a classic problem in evaluating interventional health-care strategies. While it is likely beyond the scope of this project, one idea would be to evaluate how much “outreach” is required per each new vaccination for a non-vaxer and weigh this against the benefits of a child never dealing with the conseqences of HPV vaccination in their lifetimes.