Predicting Pertussis (Whooping Cough) Outbreaks and Vaccination Rates

Jessica Vincent, Jeff Hsu, Shelly Hsu, Heather Feinstein

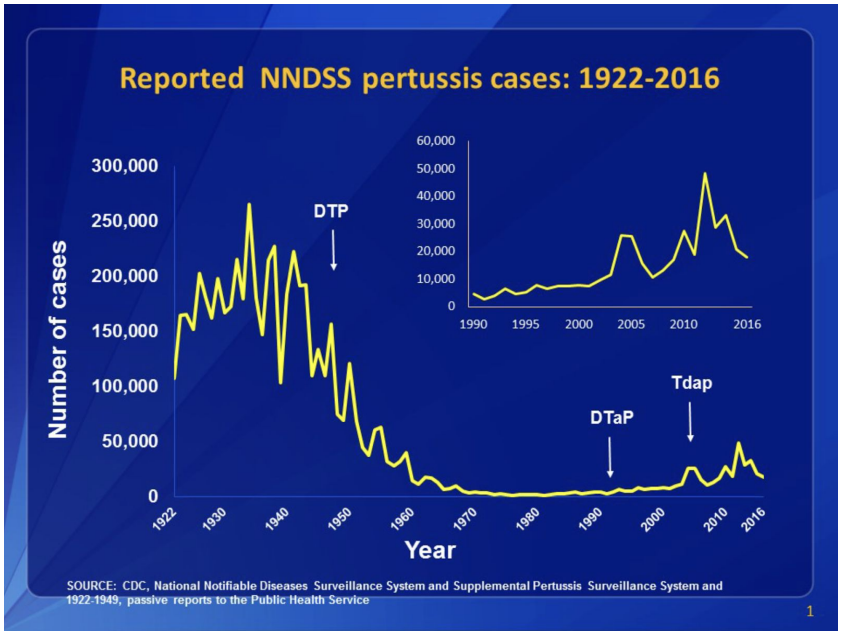
Problem Statement

Describe the problem you are solving. Why is this a compelling and impactful problem to solve? Why is this a big opportunity?

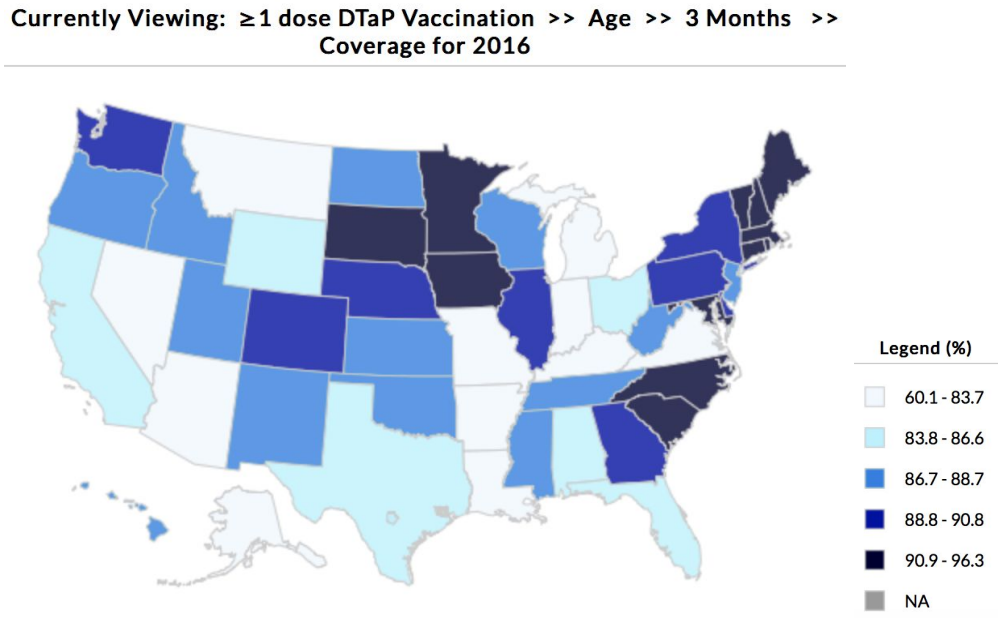
Cases of pertussis (whooping cough) have been increasing since the turn of the millenium. Pertussis causes severe and prolonged coughing which can last months and can cause death in infants. Before pertussis vaccines became available in the 1940s, about 200,000 children contracted it each year in the United States and about 9,000 died as a result. After the introduction and adoption of the DTP (Diphtheria, Tetanus, Pertussis) vaccine in 1948, cases subsequently steadily declined over several decades as the vaccine became widely administered. Cases of pertussis reached their lowest rates in the 1970s and early 1980s, with just 1,000-3,000 cases a year.

Cases started to slowly increase again in the late 1980s and 1990s, though during this time there were still fewer than 8,000 cases per year. However, the 2000s brought with them a rapid increase of cases, reaching levels not seen since the 1960s. Since 2003, there have been at least 10,000 annual cases, and more recently closer to 20,000 cases per year. In 2012, there was an estimated 48,277 cases, the highest incidence since 1955.

Pertussis outbreaks have considerable financial impacts on families and the healthcare system. It has been reported that from 1997–2000, 20% of all pertussis cases required hospitalization in the US, and that the mean cost per hospital stay for an infant was $9,580, and slightly less for children and adults.



In the mid-1970s, international controversy began over the safety of the DTP vaccine in Europe, Asia, Australia, and North America, alleging that the vaccination can lead to neurological conditions. In the United States specifically, a documentary *DTP: Vaccination Roulette* was released in 1982, and a book titled *A Shot in the Dark* was published in 1991. Both publications outlined supposed adverse reactions and risks of the DTP vaccine. Opposition and anti-vaccination movements have continued to increase since, claiming links to autism and other developmental and neurological conditions. In recent years, various celebrities have publicly advocated that parents abstain from common childhood vaccines. Currently, 84% of children between 19 and 35 months old have received DTP vaccines on average in the US, with some states having DTP vaccination rates closer to 70%. This timeline of the anti-vaccination movement in the US appears consistent and correlated with the timeline of increased pertussis cases, but is it truly a contributing factor? Some argue yes, while others cite better diagnostic techniques and new immunization-resistant pertussis strains.



The wide variety of conflicting information available regarding the safety of vaccines, the risks associated with contracting pertussis, and the causes of the increased prevalence of whooping cough can leave a parent's head spinning. Parental hesitation to vaccinate their children may delay routine immunizations, and recent outbreaks of vaccine-preventable diseases in the United States have drawn increased attention to this phenomenon. Improved understanding of the association between vaccine refusal and outbreak prevalence is needed, and this transparency is especially needed for parents who are trying to make sense of the various risks and threats they read about as they aim to keep their children safe.

Specifically, unvaccinated children can currently benefit from a concept call herd immunity, which is the resistance to the spread of a contagious disease within a population that results if a sufficiently high proportion of individuals are immune to the disease. However, if enough children in a given population are unvaccinated, the communal immunity could dip below a critical threshold, leading to significantly increased risk for the community at large. Parents in areas where anti-vaccination rhetoric is becoming more common need to be aware of increased risks. Herd immunity for Pertussis occurs when 92-94% of the population is vaccinated.

This project proposes to integrate data on historical vaccination trends, outbreaks, and demographic information to help parents understand the implications of not vaccinating their children against pertussis at a more personal level. It would seek to forecast the probability of future outbreaks and the anticipated vaccination levels in a selected region, also incorporating demographic risk factors, to help parents assess the risk of contracting whooping cough in their community. If time and interest allows, the dashboards could also incorporate regional sentiment analysis based on articles or social media posts related to vaccinations in a given area.

What assumptions are you making about about the problem / opportunity? (these assumptions would directly influence the key elements and features you would build in the MVP)

Attitudes among parents about vaccines can fall into one of three categories:

1) Parents who believe in the efficacy of vaccines and are concerned about the spread of diseases due to unvaccinated children
2) Parents who are on the fence about vaccinating their children because of misinformation
3) Parents who are strongly anti-vaccine

Our target audience falls in group 2 so we are making the assumption that those users are open minded enough to seek out this sort of information. Some people who are sceptical of vaccines are also sceptical of mainstream science so it would be important to reach them in a clear straightforward manner. We may have to target people who are on the fence about vaccinations and find a way to appeal to a user who may not have a strong scientific background.

Will assume derived vaccination rates are reflective of general vaccination rates in the city. (Combine estimates from ChildVaxView, TeenVaxView, and AdultVaxView) and weight by population data derived from American Community Survey/Census derived population estimates.

Assuming that the R0 spread through the population is the same in all areas? (not sure if we're going to do some kind of simulation for how the disease will propagate? If not delete this)

Overlay historical information from past outbreaks (I think this is only available on a county level?)

Value Proposition

What value would your data science based MVP bring to the target customer segment?

Parents will have a better understanding how vaccination affects disease propagation through neighborhoods as well as understand historical outbreak trends within specific counties of California.

How is your solution better and more differentiated than current solution(s) ?

Current solutions display the information on a overall county level or just by school and does not provide a broad overview about how vaccination rates affect a theoretical spread of disease in a specific community. We will try to frame the situation as a story to convey the meaning behind the numbers (feel free to change this)

Why should the market and your target customers care? In other words, why is this compelling to them?

The concept of herd immunity may not be well understood. Even if herd immunity is understood what is not understood is all the assumptions that underlie herd immunity. It may be more compelling to show how much impact a subpopulation of unvaccinated will affect them in a specific neighborhood.

Use Case / Customer Segment

Who are you creating the solution for? Be specific in identifying your targeted customer segment for the MVP.

Parents in California, segmented by county (exploring availability of data on a city/neighborhood level). Our customers would be both parents who are supportive of vaccines and skeptical of them but not parents who are firmly anti-vaccine. As such we do not want to focus on refuting anti-vaccination propaganda but instead focus on spreading science to a wider audience of relatively open minded consumers.

What is the use case enabled by your MVP for this target customer segment?

The use case would involve a parent skeptical of vaccinations who is wondering what the potential impact of not vaccinating their child would be.

When will you be doing initial customer validation? How would you go about to identify potential users / customers for idea and MVP validation?

Will validate model with advice from public health professionals and people who have children. Pilot test some basic mock-ups for visual layouts with people we know. Attempt to identify people who are opposed to vaccinations and get their feedback as well.

MVP

What is the minimal viable product that you are building that will specifically test the fundamental assumptions you have about the problem and the value of your solution?

Dashboard showing:

- Historical pertussis trends in the US
- Drill down into vaccination rates by neighborhood/city within California
    - Historical and point in time
- Simulations modeling of disease outbreak given interactive parameters
- Potential cost modeling

What data science approach and features would you intend to have for the MVP? (this is NOT UI / UX but technical discussion)

Simulation modeling of how disease might spread/likelihood of catching the disease if there are unvaccinated individuals

Data sets

What datasets do you intend to use?

- CDC National Immunization Surveys (NIS) Databases - Conducted annually and used to obtain national, state, and selected local area estimates of vaccination coverage rates for U.S. children 19–35 months (NIS-Child) and for U.S. adolescents 13–17 years (NIS-Teen).
- CDC ChildVaxView - Local, state, and federal health department survey data to estimate vaccination coverage (the proportion of children receiving vaccinations). TeenVaxView and AdultVaxView are also available.
- CDC Pertussis Surveillance and Reporting - Historical annual cases of pertussis from 1922 to 2016, and broken down by age group from 1990 to 2016.
- CDC National Notifiable Infectious Diseases: Weekly Tables Meningococcal disease, all serogroups; Mumps; Pertussis - Weekly cases of selected notifiable diseases (≥ 1,000 cases reported during the preceding year) in United States and U.S. territories.
- CDC Vaccination Coverage for Vaccination Managers - Vaccine coverage levels by age, poverty status, urbanicity, race, facility, other demographics.
- CDC WONDER: Vaccine Adverse Event Reporting System (VAERS) - Counts and percentages of adverse event case reports after vaccination, by symptom, vaccine product, manufacturer, onset interval, outcome category, year and month vaccinated, year and month reported, age, sex and state / territory.
- https://data.hrsa.gov/data/download - Medically underserved areas

- https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/Immunization/pertussis.aspx California DPH General Website with surveillance report by county within CA
- http://www.shotsforschool.org/child-care/how-doing/ Vaccination coverage
- https://www.cdph.ca.gov/Programs/CID/DCDC/CDPH%20Document%20Library/Immunization/PertussisQuicksheet.pdf General information about pertussis that might be useful to include in information/introduction section
- School vaccination data for California State
    - Childcare facilities http://www.shotsforschool.org/child-care/reporting-data/ (2011-12 to 2016-17 School year)
    - Schools with kindergartens http://www.shotsforschool.org/k-12/reporting-data/ (2001-2 to 2017-18 School year)
    - Schools with 7th grade http://www.shotsforschool.org/7th-grade/reporting-data/ (2011-12 to 2017-18 School year)
- Census Data weights for california cities
    - Census 2010 Total Population Counts https://factfinder.census.gov/bkmk/table/1.0/en/DEC/10_DP/DPDP1/0400000US06.16000
        - Can also get 2000 numbers if needed
    - ACS Population 5-year Estimates (2012-2016) https://factfinder.census.gov/bkmk/table/1.0/en/ACS/16_5YR/DP05/0400000US06.16000
        - There is data available from as far back as 2006-2010 for ACS 5-year estimates

Are the datasets public?

Yes.

What are the datasets attributes / metadata that would could make the exploratory data analysis easier / harder?

Geographic areas might not be defined in the same way so data processing will be needed to make the geographic areas referenced similar/comparable. Will need to derive actual vaccination coverage rates by city and weight according to the population distribution.

Project Management

In your team, who will be the project lead and project manager to ensure everything will be delivered on time and at expectation by each team member and collectively as a team?

Heather

What is the role of each member (who will do what specifically)? How would the team work together? What's the operating rhythm?

The team will have weekly meetings each Saturday at 1pm PST to discuss progress, issues, feedback and determine next steps. Our roles are as follows:

> **Jessica**: Happy to do whatever. Let's discuss in person. I'm comfortable with PM roles, R, Tableau, but also open and willing to learn new tools or technologies as needed!
>
> **Jeff**: I'm open to do data processing, modelling, web development and project management. In terms of programming languages or tools, I am more familiar with python than R, I use Tableau for visualisation at work and am pretty adept at setting up database & writing ETL pipelines. I am weaker at subject matter know-how and industry, so it'd be good if someone can help out on user surveys & the domain matter intuition inputs.
>
> **Shelly**: I'm open to doing visualization, spatial analysis, and data processing and management. I can provide subject matter expertise about geography, demographics, and epidemiology data. I can also test the the prototype mock-ups and final product among public health professionals. I'm fine with helping with the python coding as well as R although I'm slightly more comfortable with python coding than R. I can create more sophisticated spatial maps in ArcMap if D3, plotly, and/or tableau aren't sufficient.
>
> **Heather:** I'm open to doing Project Mgmt, Visualization, R Statistical analysis. Weaker on the Python/data pipeline side.

How would you plan to implement agile methods, and what tools will you use?

We plan to iterate over multiple versions based on user testing and feedback.

Technical Approach

What methodologies would you use for initial data exploratory analysis to ensure your datasets are sufficient and meaningful?

Python and Tableau EDA

What data science algorithms are you intending to develop and build for the project? What challenges do you potentially foresee?

Regression, Time Series analysis

What help do you need?

Website design expertise.