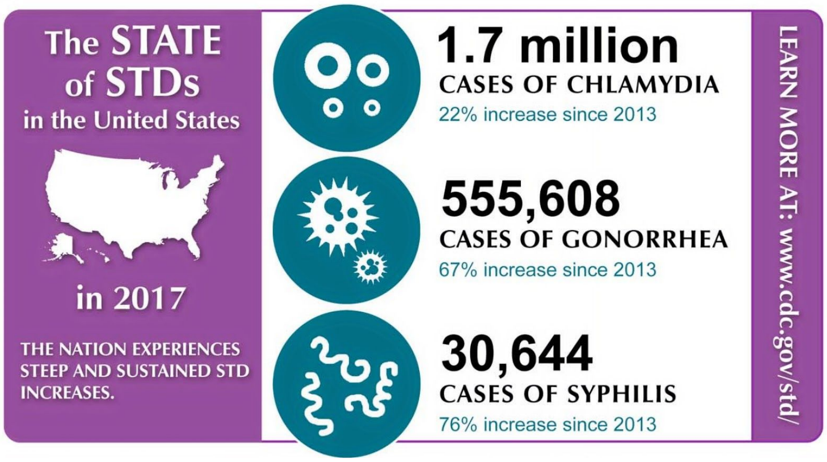
BLUEPRINT
DATATHON

2019 Case: The State of STDs in the United States

Background:

Over the last decade, STD rates have steadily been on the rise, with over 2 million new infections reported each year. In the United States alone, the developing STD epidemic has totaled over $16 billion of medical expenses annually, with younger demographics representing the majority of cases. Students in high school and college acquire 50% of new STDs.



The Center for Disease Control and Prevention has reported an all-time high in STD cases in 2017, despite recent research suggesting an average decrease in sexual activity across the US. Rates of *Gonorrhea* diagnoses have increased by nearly 67% over the past few years, with little to no signs of stopping. *Syphilis* and *Chlamydia* have shown similarly steep increases in overall diagnoses rates.

In some cases, STDs can develop into more serious, lethal infections. Though many of these infections can be cured with antibiotics, many still go undiagnosed and/or untreated. The resulting adverse health effects include ectopic pregnancies, increased HIV risk, and stillbirth. Other serious concerns include the rise of antibiotic resistant STDs, some of which have been shown to resist almost all classes of antibiotics.

Although this epidemic has gained attention from doctors, public health/policy experts, and media over recent years, significant progress has yet to be made. Causes of any of these trends are unclear.

The Research Problem:

Swoop is excited to support SHIFT and Stanford's best data sleuths in this year's Blueprint Datathon. Swoop uses HIPAA-compliant, privacy-preserving AI to improve patient outcomes and the business operations of pharmaceutical companies. We are based in the Boston area. If you are interested in data engineering or ML/AI work that helps improve lives, get in touch with us at interns@swoop.com or careers@swoop.com.

Big trends spanning a great nation typically have root causes, so why haven't researchers been able to find the equivalent of a smoking gun? Are they looking at the right independent variables and interactions? Are they able to think creatively about the key factors influencing the attitudes and behaviors of young adults? Did they consider social media and dating apps?

To answer this conundrum, we've prepared a Real World Evidence (RWE) dataset that combines STD prevalence information with demographics and lifestyle factors for the five year period between 2014 and 2018. We invite you to explore the data sources we have provided to **understand the underlying predictors of the rise in STD prevalence**.

At the core, your final deliverable should be centered around what key factors influence STD prevalence rates. Your analysis should include a thorough examination of the sources provided and use a sound statistical approach to this problem that accounts for various external factors. This can involve predictive models (ie. machine learning, regression-based techniques) to predict STD prevalence based on a defined feature set, or a descriptive modeling to quantify relationships you may find in the data.

Beyond this core question, further research questions that build off of your preliminary insights are open to the creativity of your team. For example, a longitudinal analysis could look into whether there are there any predictors for a spike in STD rates around the country year-to-year? As you play with and investigate the data, we **highly** encourage you to form your own research questions to drive your insights. This will greatly distinguish your team from the rest and will be of interest to the judges!

Deliverables:

The competition will be judged in two parts. First, teams will be required to submit their case by **10 A.M. Sunday, 4/14**. These deliverables should summarize your analysis methods and final conclusions. Judges will choose the top 5 teams based on the submissions. The top 5 teams will be invited to give a 5 minute presentation during our closing ceremony, followed by 5 minutes of questions from the judges. The presentation should expand on the materials presented in the Project Expo pitch.

For all deliverables, teams should:

- Describe the methods used
- Interpret results, concentrating on what you learned through the Datathon
- Emphasize challenges in carrying out the analysis
- Illustrate the originality and novelty of your approach
- Reference any external sources you used to help you complete the task

Judging:

Judging will broadly be based upon the following:

- Soundness of the approach taken (including, but not limited to, statistical significance, auROC)
- Potential scientific, societal, and policy impacts of the results
- Originality and novelty of the approach
- Quality of the description of the data and tools used, especially reproducibility
- Quality of the 3-minute pitch during the Project Expo

The Data:

The puzzle of the driving factors behind the rise of STDs presents an excellent opportunity to apply data science techniques to address a pertinent health crisis. We invite you to find creative solutions regarding the aforementioned research questions, but we also encourage you to ask and begin finding answers to any other questions to you may find relevant.

You can download the data as a compressed CSV file here. (72.7Mb compressed, 1.54Gb uncompressed, 1,366,808 rows).

To make the lawyers happy: The above data is meant exclusively for use by the Stanford Blueprint Datathon teams for the duration of the datathon. It should not be shared outside of this context without prior written approval by Swoop.

The dataset consists of four types of columns:

- **`date`:** a date rounded to the start of a quarter between 2014-2018.
- **aggregation columns:** demographics attributes that determine the aggregation level.
- **propensities:** summary statistics about the lifestyles of the people at a certain aggregation level.
- **prevalences:** ratios of the number of STD-related diagnoses to the total number of patients in the data.

In the entire dataset, `null` means unknown.

Aggregation Columns:

The provided data is aggregated into groups based on core demographics attributes:

- age
- gender
- state
- income
- education

The primary key for the dataset is `(date, age, gender, state, income, education)`. Each row of the dataset summarizes the attributes of patients matching the demographics attributes for any given quarter.

Each group described by the aggregation columns may have multiple rows at different dates. There is no guarantee that the same values for a group will be present for all dates.

Propensity Columns:

The following columns summarize lifestyle propensities. Some propensities are encoded as categorical variables (`low`, `medium`, `high`) while others are encoded as real values between zero and one.

- technology_and_connectivity__online_gamers
- technology_and_connectivity__stream_music

- technology_and_connectivity__science_&_new_tech_enthusiasts
- technology_and_connectivity__owns_apple_products
- technology_and_connectivity__smart_phone_owners
- technology_and_connectivity__download_videos
- technology_and_connectivity__stream_videos
- technology_and_connectivity__electronics_enthusiast
- wearable_technology_and_connectivity_users
- smart_phone_users
- restaurant_app_users
- services_software_and_online_services__genealogy_research_service_paid_users
- services_software_and_online_services__online_review_services_paid_users
- services_software_and_online_services__graphic_design_software_paid_users
- services_software_and_online_services__online_dating_paid_users
- services_software_and_online_services_frequency
- services_software_and_online_services_spend
- restaurants_genres__ice_cream_restaurant_customers
- restaurants_genres__italian_restaurant_customers
- restaurants_genres__steakhouse_customers
- restaurants_genres__pizza_restaurant_customers
- restaurants_genres__coffee_shop_customers
- restaurants_genres__sandwich_&_sub_restaurant_customers
- restaurants_genres__donut_shop_customers
- restaurants_genres__bbq_restaurant_customers
- restaurants_genres__chicken_restaurant_customers
- restaurants_genres__asian_restaurant_customers
- restaurants_genres__pasta_restaurant_customers
- restaurants_genres__dessert_restaurant_customers
- restaurants_genres__seafood_restaurant_customers
- restaurants_genres__lobster_restaurant_customers
- restaurants_genres__french_restaurant_customers
- restaurants_genres__american_restaurant_customers
- restaurants_genres__mexican_restaurant_customers
- restaurants_genres__mediterranean_restaurant_customers
- restaurants_genres__delivery_pizza_restaurant_customers
- restaurants_genres__chicken_wing_restaurant_customers
- restaurants_genres__cafe_restaurant_customers
- restaurants_genres__bakery_&_pastry_restaurant_customers
- restaurants_genres__breakfast_food_customers

- restaurants_genres__burger_restaurant_customers
- restaurants_sit_down_casual_dining__arcade_&_entertainment_restaurant_customers
- restaurants_sit_down_casual_dining__lunch_&_dinner_focused_restaurant_customers
- restaurants_sit_down_casual_dining__upscale_casual_restaurant_customers
- restaurants_sit_down_casual_dining_frequency
- restaurants_sit_down_casual_dining_spend
- restaurants_sit_down_upscale_dining__upscale_asian_fusion_restaurant_customers
- restaurants_sit_down_upscale_dining__upscale_seafood_restaurant_customers
- restaurants_sit_down_upscale_dining__upscale_steakhouse_customers
- restaurants_sit_down_upscale_dining_frequency
- restaurants_sit_down_upscale_dining_spend
- restaurants_fast_casual_dining_frequency
- restaurants_fast_casual_dining_spend
- entertainment_movies__movie_renters
- entertainment_movies__small_&_independent_theater_customers
- entertainment_movies__national_brand_theater_customers
- entertainment_movies__online_movie_ticket_buyers
- entertainment_movies_frequency
- entertainment_movies_spend
- services_ride_sharing_spend
- services_ride_sharing_frequency
- coffee_enthusiasts

Prevalence Columns:

The prevalence columns are ratios of diagnosis counts to number of patients. For example, if the value for hpv is 1.2 this could be because the group contains 10 people who were diagnosed with HPV a total of 12 times. This could also be caused by 20 people who were diagnosed with HPV a total of 24 times. Keep in mind that the same person can be diagnosed with a condition multiple times.

The medical meaning of the prevalence columns is as follows:

- **chlamydia**: includes aggregated set of diagnosis codes pertaining to sexually transmitted diseases due to chlamydia trachomatis. Chlamydia trachomatis is a bacterial infection and one of the most common STIs (sexually transmitted infections). The infection is spread by coming into contact with infected secretions.

- **gential_warts**: includes aggregated set of diagnosis codes pertaining to sexually transmitted diseases due to Condyloma acuminatum (genitial warts). Gential warts refers to an epidermal manifestation attributed to the epidermotropic human papillomavirus (HPV). The disease is spread by skin to skin contact.
- **gonorrhea**: includes aggregated set of diagnosis codes pertaining to sexually transmitted diseases due to Neisseria gonorrhoeae bacterium. Gonorrhea is a bacterial infection. The infection is spread by coming into contact with infected secretions.
- **herpes**: includes aggregated set of diagnosis codes and treatment codes pertaining to sexually transmitted diseases due to viruses called herpes simplex virus type 1 (HSV-1) and herpes simplex virus type 2 (HSV-2). Both types of herpes simplex virus are transmitted through skin to skin contact with open sores or non-visible infected skin, saliva, and infected secretions.
- **hpv**: includes aggregated set of diagnosis codes pertaining to sexually transmitted diseases due to human papillomavirus (HPV). The code set **does not** include the diagnosis codes used to define genital warts, a subtype of HPV, nor does it include diagnosis codes pertaining to preventative treatment or testing. There are over 100 types of human papillomavirus infections with the various subtypes linked to types of cervical cancer. It is spread through skin to skin contact with an infected source.
- **other_std**: includes aggregated set of diagnosis codes pertaining to sexually transmitted diseases that are not specifically enumerated in the description. Commonly referred to as “catch all” codes, the codes indicate a confirmation of a sexually transmitted disease that is not enumerated at a granular level or classified elsewhere. This group also includes diagnosis codes pertaining to chancroid and granuloma inguinale . Chancriod is an infection of the genital skin or mucous membranes caused by Haemophilus ducreyi. Granuloma inguinale is a bacterial disease caused by Klebsiella granulomatis (formerly known as Calymmatobacterium granulomatis).
- **parasitic**: includes aggregated set of diagnosis codes pertaining to sexually transmitted diseases for Phthirus pubis (pubic louse) and scabies. Pubic lice and scabies are both parasitic infestations. Both are highly contagious and transmitted by skin to skin contact or even contact to a contaminated source.
- **std_screen**: includes aggregated set of diagnosis codes pertaining to the screening/testing of sexually transmitted diseases. These codes are not definitive confirmation of a disease, but are used when there are investigational efforts to determine if an STD or STI is present. There are multiple causes for STD

screening, ranging from determining the exact cause of present symptoms to a safety check for sexually active people.

- **syphilis**: includes aggregated set of diagnosis codes pertaining to sexually transmitted diseases for syphilis infections. The codes include all codes pertaining to various stages of syphilis: primary, secondary, latent, and tertiary, and excludes all codes related to congenital and neonatal related syphilis. Syphilis is spread by coming into direct contact with a syphilis sore.
- **trich**: includes aggregated set of diagnosis codes pertaining to sexually transmitted diseases for Trichomoniasis, an infection caused by a protozoan parasite called Trichomonas vaginalis. Trichomoniasis is a parasitic infection and is transmitted by sexual contact with an infected source.

BRFSS Data:

Here are additional state level data sets that you might find useful in your analysis of the case. The BRFSS data sets contain information regarding U.S. residents health related risk behaviors, chronic health conditions, and use of preventative services. You may find these data sets as a good complement to the one given above to determine new trends between certain behavioral tendencies and the rise of STDs.

2013 Data

2017 Data

The data dictionaries for these files is available here:

2013 Data Dictionary
2017 Data Dictionary

More information about the survey methodology and variable descriptors can be found on the BRFSS website. Choose the year the data set was collected and look at the codebook to help interpret the meaning behind variables in the data set.

QUESTIONS:

If you have any questions regarding the case or event logistics, make sure to send your question in the Slack group. If you haven't already, join the slack group here.

One more thing... Some years ago, Sim Simeonov, our CTO and a longtime TechStars mentor, kicked off a different type of hackathon, a Startup Weekend, with some winning tips. 50-75% of the tips apply to the Blueprint Datathon so you may want to check them out. 😉

We can't wait to see the results of your work. Here's to the crazy ones!