Landscape of Sexually Transmitted Diseases in California

By Debbie Argueta Rufino
2023-12-08

Introduction

This data comprises statistics on the number of cases and infection rates of sexually transmitted diseases (specifically chlamydia, gonorrhea, and early syphilis, encompassing primary, secondary, and early latent syphilis) that have been reported for California residents. The data is categorized by disease type, county, year, and gender.

The data was collected for cases with estimated diagnosis dates spanning from 2001 up to the most recent year available. It was sourced from California Confidential Morbidity Reports and Laboratory Reports, all of which were submitted to the California Department of Public Health (CDPH) by July of the current year. These reports adhered to the surveillance case definition for each respective disease.

After looking at the data, the main question of interest we wanted to investigate was : Which STD has the highest prevalence in California, and how is this disease geographically spread across the state? Further analysis was conducted to look at the year that had the highest STD rates and the difference between infection rates based on sex.

Methods

Data Acquisition

The STD data was retrieved from "https://data.chhs.ca.gov/dataset/stds-in-california-by-disease-county-year-and-sex".

The geographical data was retrieved from "https://public.opendatasoft.com/explore/dataset/us-county-boundaries/export/?disjunctive.statefp&disjunctive.countyfp&disjunctive.name&disjunctive.namelsad&disjunctive.stusab&disjunctive.state_name&refine.stusab=CA".

Data Cleaning Wrangling

The STD data did not include any latitude and longitude coordinates, thus the second data set was introduced to conduct a proper geographic analysis. First, we merged the main data set with the geographic data set.

The combined data set has 11 columns. Among them, columns “Cases” and “Rate” have several missing values because of the “Annotation Code” variable, which prevents them from being publicized. Therefore, these missing values were removed.

The data type of the column “Rate” is *chr (character)*, so we changed it into a numeric format.

The “County” column includes rows called “California”, which is the state not a county, so we delete them. I saved the aggregate “California” data into a new variable “Cali”.

Results

A line plot was generated to examine the trends in STD rates spanning from 2001 to 2020.

Figure 1

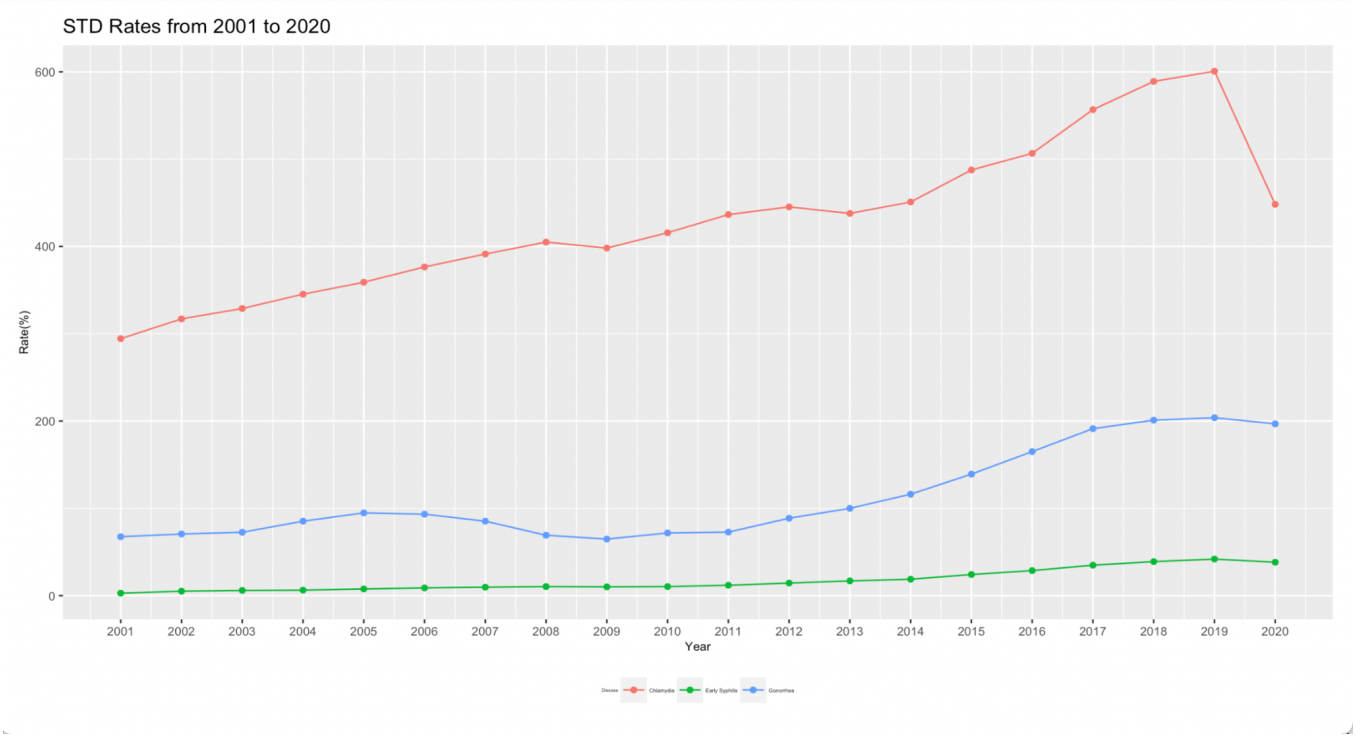

Figure 1. A visual analysis of the graph reveals that Chlamydia consistently maintained the highest infection rate during this entire time frame, with Syphilis and Gonorrhea following closely behind in terms of prevalence.

Given our earlier discovery of Chlamydia having the highest infection rate, we delved deeper by constructing a box plot to investigate this STD's infection rates exclusively.

Figure 2

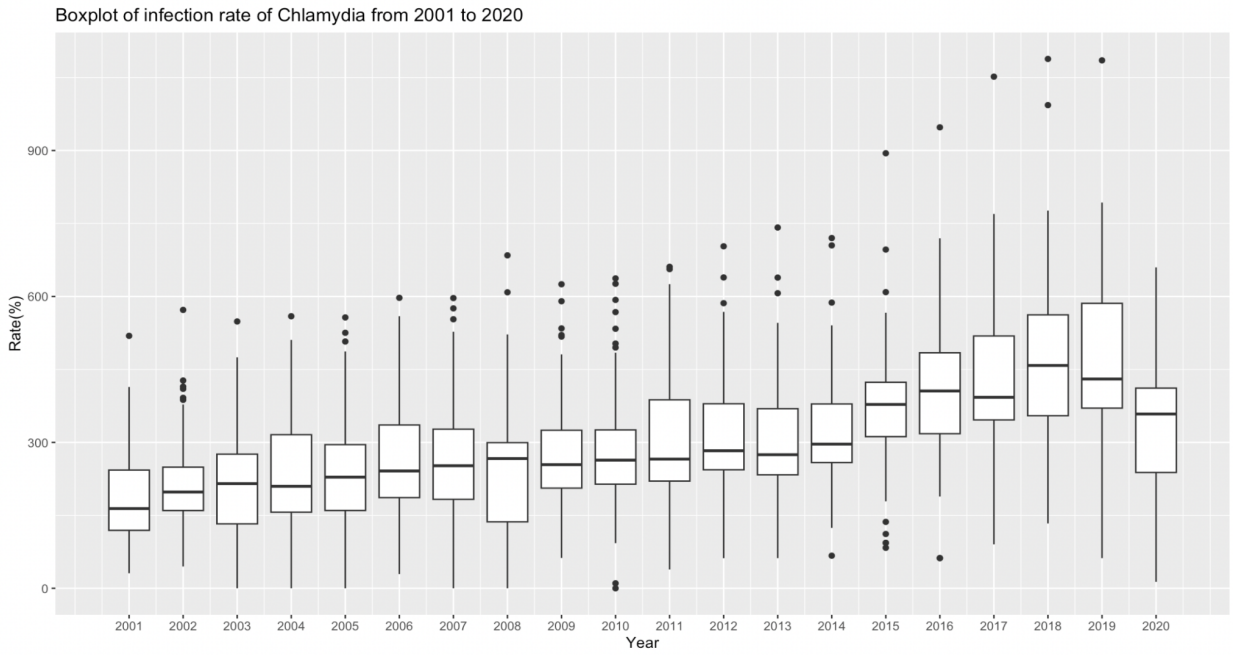

Figure 2. The resulting plot reveals that 2019 stood out as the year with the highest Chlamydia infection rates in California.

To examine infection rates by county for the year 2019, a visual representation was generated in the form of a bar chart.

Figure 3

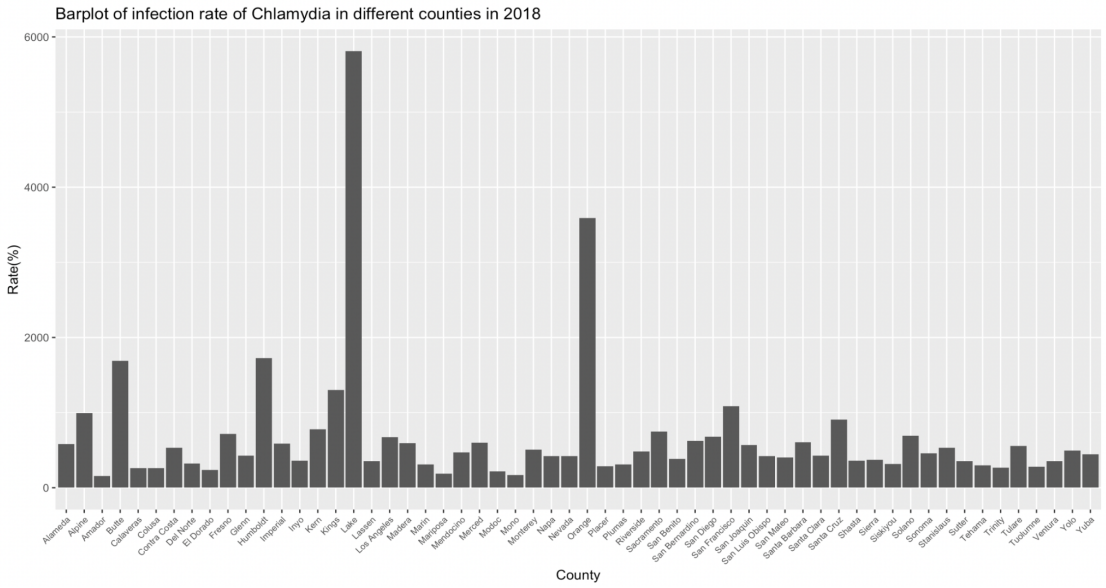

Figure 3. This figure demonstrates that Lake County exhibited the highest infection rate for that specific year, while Amador County documented the lowest infection rate.

In order to assess geographical variations in infection rates, we designed a map for a visual analysis

Figure 4

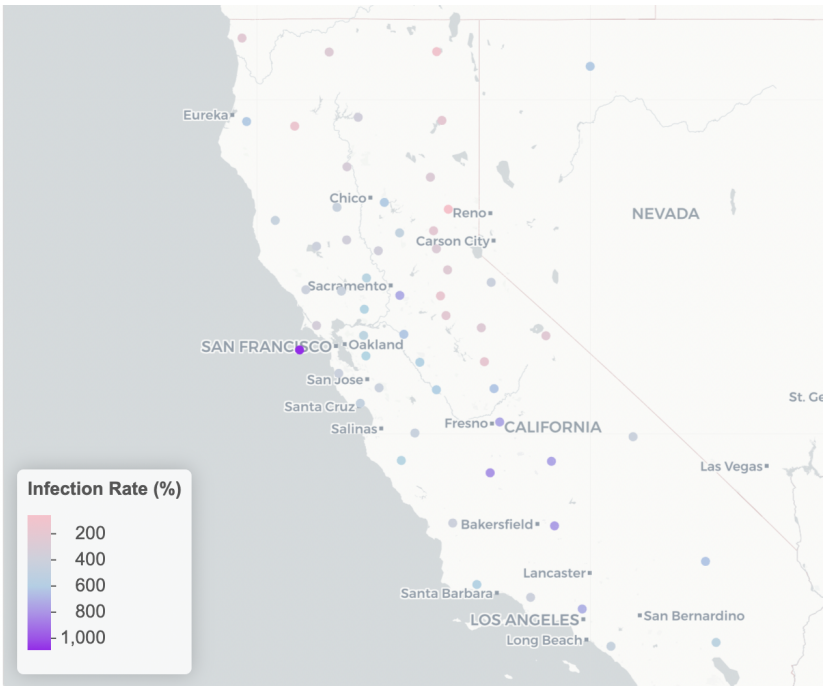

Figure 4. Our observation from this map indicates that the central valley of California, along with a selection of counties in the Bay area, exhibited the highest infection rates. Notably, the counties in closer proximity to Nevada demonstrated the lowest infection rates.

A table was created to evaluate the differences between Chlamydia rates in men and women.

Table 1.

Sex	Cases	Population	Rate
Female	2292	407916	0.5618804
Male	1032	409452	0.2520442
Total	3324	817368	0.4066712

Table 1. The data indicates that females are twice as likely to acquire infections compared to males.

Conclusion

Chlamydia held its position as the most prevalent STD in California from 2001 to 2020. The year 2019 witnessed the highest infection rates statewide, with Lake County bearing the brunt of this issue.

An apparent geographic pattern emerged, with the central valley reporting the highest infection rates and a gradual decrease towards the Nevada border.

Additionally, a notable gender discrepancy was observed in Lake County in 2019, where females reported twice as many infections as males, highlighting the importance of tailored interventions and awareness initiatives.

Discussion

This analysis gives a preliminary overview of STD infection rates in California. The observed variations in the decline of syphilis and gonorrhea rates in contrast to chlamydia rates in 2020 raise intriguing questions. While chlamydia exhibits a more pronounced decrease, syphilis and gonorrhea seem less affected. Investigating the potential factors influencing these distinct patterns could be a valuable

avenue for further research. Possible contributors may include variations in testing practices, public health interventions, or the unique nature of each infection.

The stark gender disparity identified in Lake County in 2019 prompts an exploration into whether similar patterns exist across the broader dataset. Unraveling the factors contributing to such differences could provide valuable insights into the targeted development of interventions. Understanding whether these disparities are consistent or vary across different regions and demographics would be pivotal in tailoring public health strategies to address the specific needs of affected populations.

It's crucial to acknowledge the potential influence of external factors on reporting accuracy. Changes in healthcare access, awareness campaigns, or even disruptions caused by external events (such as the COVID-19 pandemic) could impact testing rates and subsequently affect the reported data. A comprehensive analysis should consider these elements to provide a nuanced interpretation of the observed trends.