Findings from EDA

In our dataset, we have a total of 340,170 rows and 11 columns that provide comprehensive information on sexually transmitted diseases (STDs) across various states in the United States. The variables in our dataset include:

- Disease: This variable represents the type of STD and includes indicators such as prevalence, deaths, and diagnoses. The diseases covered in our dataset are Primary and Secondary Syphilis, HIV diagnoses, Gonorrhea, Early Non-Primary Non-Secondary Syphilis, Chlamydia, AIDS classifications, HIV prevalence, HIV deaths, AIDS prevalence, and AIDS deaths.
- Year: This variable indicates the year during which the data was recorded. Our dataset covers the years 2010 to 2020.
- State: This variable represents the states in the United States, focusing on Maryland, New York, and Georgia for our analysis.
- Race: This variable captures the racial/ethnic background of the individuals in the dataset, including White, Native Hawaiian/Other Pacific Islander, Multiracial, Black/African American, Asian, and American Indian/Alaska Native.
- Sex: This variable differentiates between male and female individuals.
- Age: This variable represents the age groups, categorized as 55+, 45-54, 35-44, 25-34, and <25.
- Cases: This variable represents the number of cases for each disease.
- Rate_per_100000: This variable indicates the rate of cases per 100,000 individuals.
- total_STD: This variable represents the sum of cases for all STD types.
- Population: This variable indicates the population size for each demographic group.
- STD_per_100000: This variable represents the total STD rate per 100,000 individuals.

Our exploratory data analysis (EDA) will primarily focus on assessing the impact of state policies on STDs, particularly:

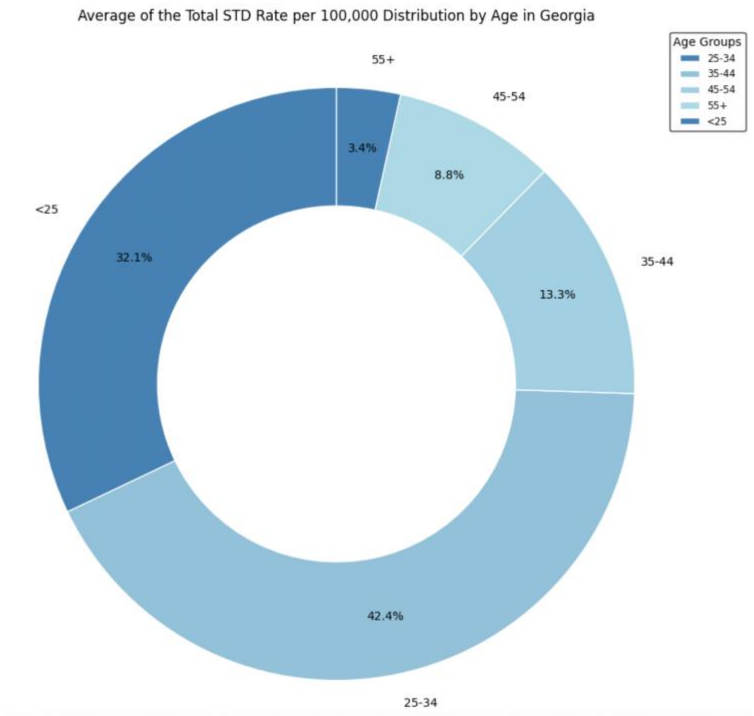
- Differences between males and females.
- Disparities among different racial groups.
- Variations among different age groups.
- Trends and status of STDs, as measured by metrics such as Gonorrhea, Chlamydia, AIDS, HIV, and Syphilis diagnoses, deaths, and prevalence rates.

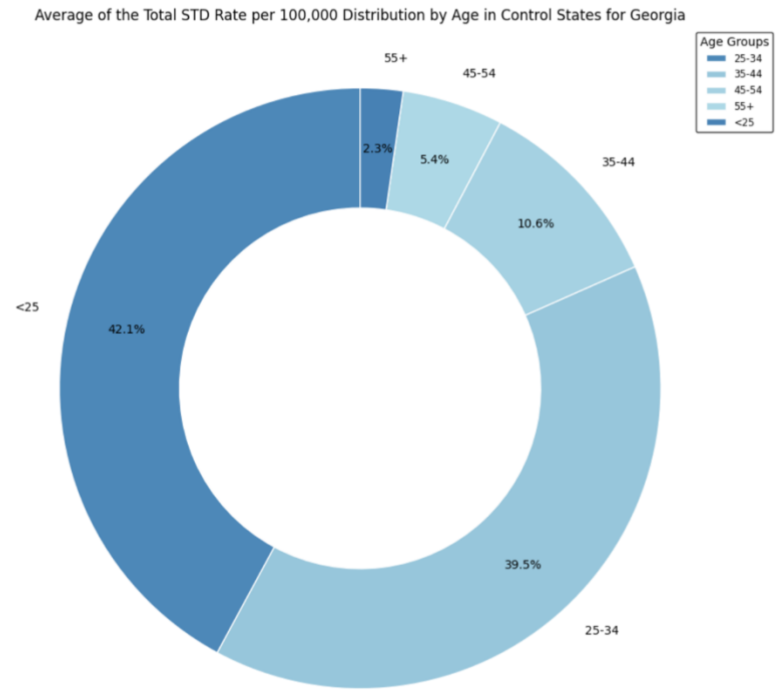
We will identify the most critical aspects among these four areas and further explore the effectiveness of policies to address the sexual health challenges in Maryland, New York, and Georgia. Our EDA will provide valuable insights into the distribution of STDs across demographic groups, enabling us to pinpoint areas in which policy interventions may be most effective in improving sexual health outcomes.

Differences Between Ages

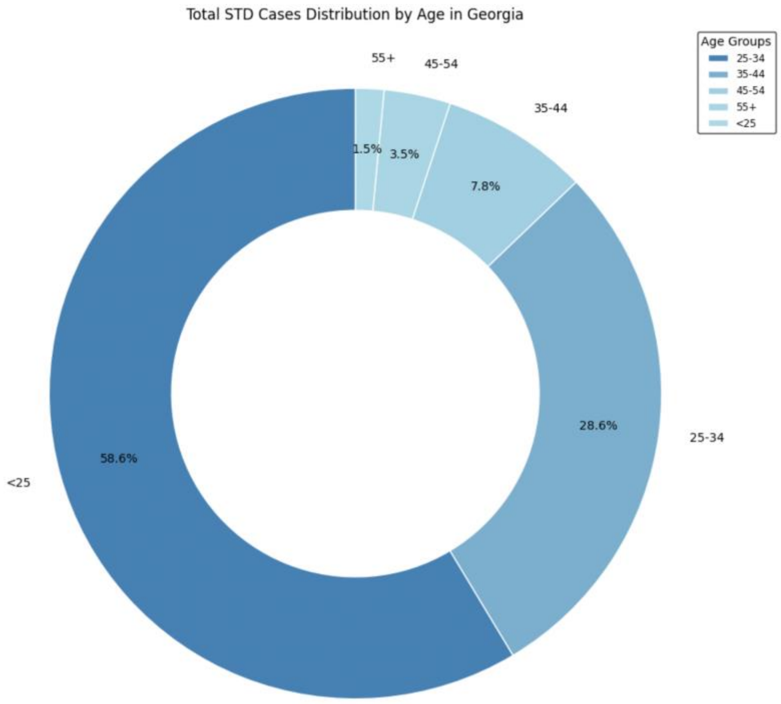
Georgia

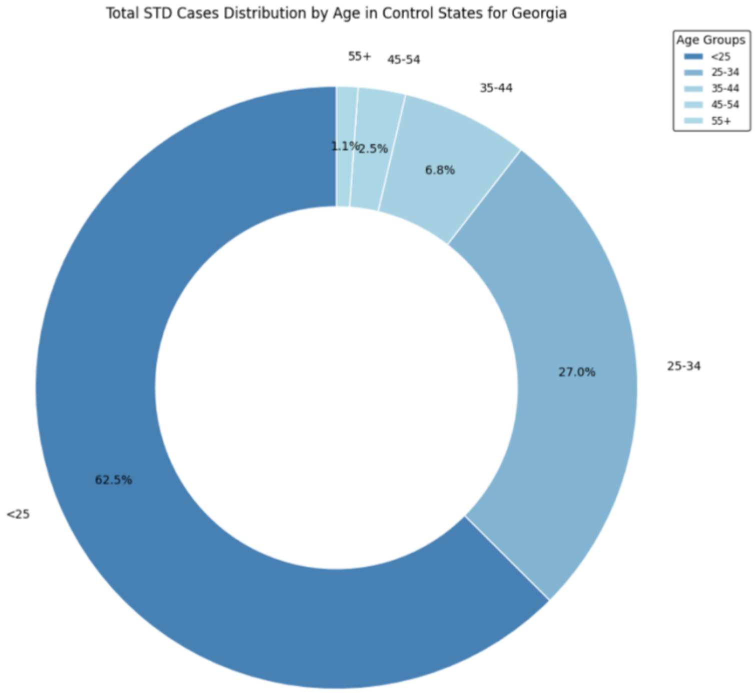
In the chart of the average total STD rate per 100,000 by age distribution in Georgia, it was found that the age group of 25-34 years old had the highest proportion of STD rates, accounting for approximately 42.4% of the average total. The second-highest proportion was observed in individuals aged below 25 years, contributing to about 32.1% of the average total. In comparison, the control states for Georgia exhibited a similar trend, with the two age groups maintaining the highest distribution. However, the age group of less than 25 years old took the lead in the control states, with an average total STD rate of around 42.1%, while the 25-34 years old age group followed closely with a 39.5% share.





Upon examining the total STD cases by age distribution in Georgia, the age group of less than 25 years old emerged as having the largest proportion of total STD cases, constituting approximately 58.6% of the total. The age group between 25-34 years old ranked second, with an average share of 28.6% in total STD cases. This pattern was consistent with the findings in the control states for Georgia. In these control states, the age group of less than 25 years old had the largest proportion of total STD cases, with an average of 62.5%, while the age group of 25-34 years old held the second-largest share, with 27% of total cases.

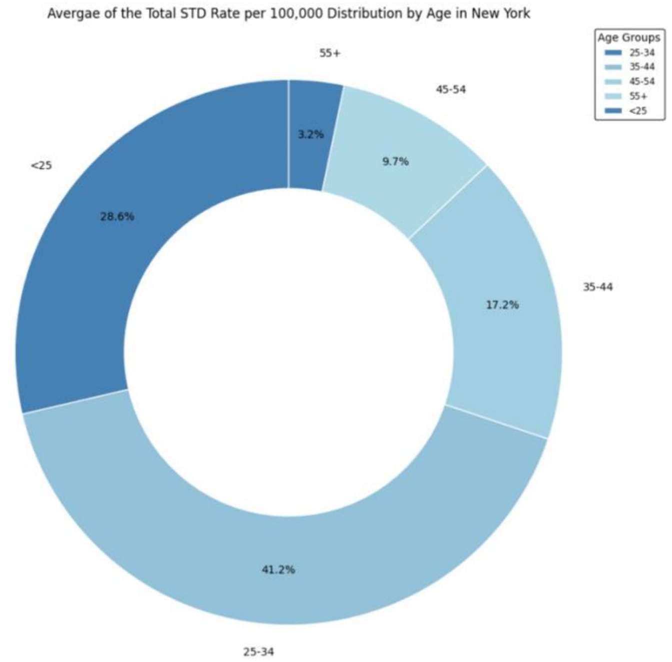


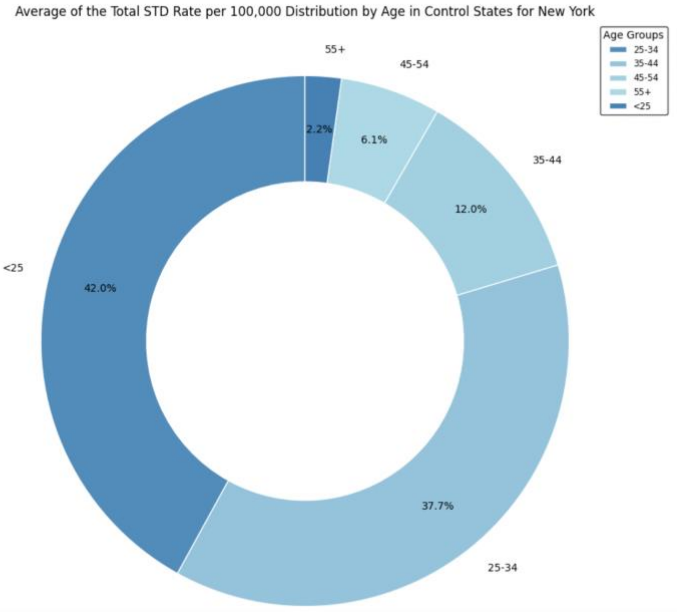


In summary, both the average total STD rate per 100,000 and the total STD cases by age distribution in Georgia and its control states showed that the age groups of 25-34 years old and less than 25 years old consistently had the highest proportions. While the ranking of the age groups varied between the average total STD rate and total STD cases, the two age groups remained the most affected across both metrics.

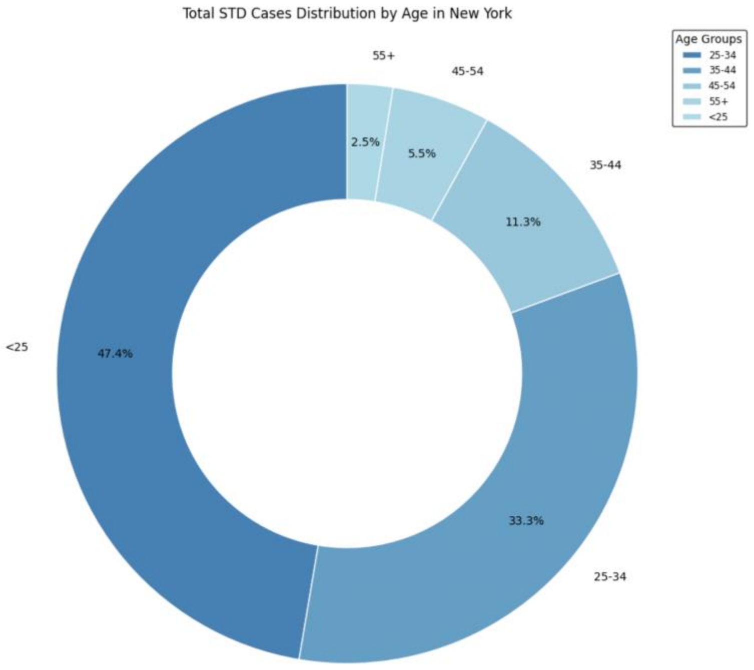
New York

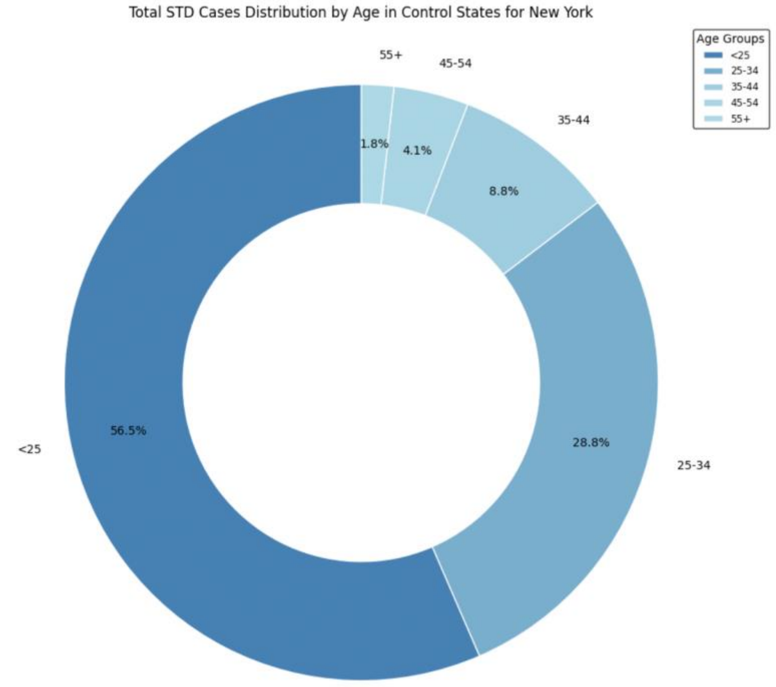
In the chart of the average total STD rate per 100,000 by age distribution in New York, it was found that the age group of 25-34 years old had the highest proportion of STD rates, accounting for approximately 41.2% of the average total. The second-highest proportion was observed in individuals aged below 25 years, contributing to about 28.6% of the average total. In comparison, the control states for New York exhibited a similar trend, with the two age groups maintaining the highest distribution. However, the age group of less than 25 years old took the lead in the control states, with an average total STD rate of around 42%, while the 25-34 years old age group followed closely with a 37.7% share.


Avergae of the Total STD Rate per 100,000 Distribution by Age in New York


Average of the Total STD Rate per 100,000 Distribution by Age in Control States for New York

Upon examining the total STD cases by age distribution in New York, the age group of less than 25 years old emerged as having the largest proportion of total STD cases, constituting approximately 47.4% of the total. The age group between 25-34 years old ranked second, with an average share of 33.3% in total STD cases. This pattern was consistent with the findings in the control states for New York. In these control states, the age group of less than 25 years old had the largest proportion of total STD cases, with an average of 56.5%, while the age group of 25-34 years old held the second-largest share, with 28.8% of total cases.

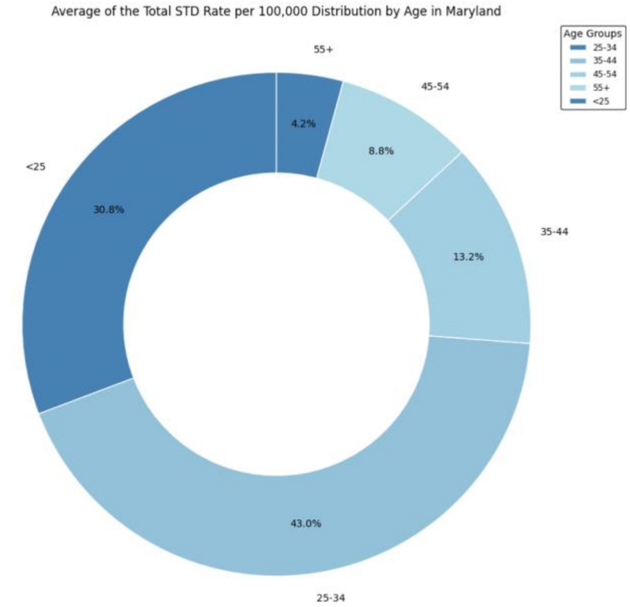

Total STD Cases Distribution by Age in New York

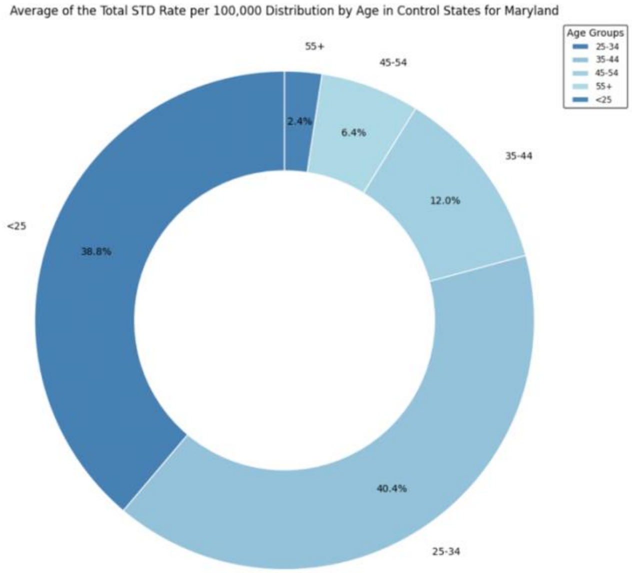

Total STD Cases Distribution by Age in Control States for New York

Overall, both the average total STD cases per 100,000 and the total STD cases by age in New York and its control states showed that the 25-34 and under 25 age groups consistently had the highest rates. Although the ranking of the age groups varied between the average total number of STD cases and the total number of STD cases, both age groups remained the highest in both measures.

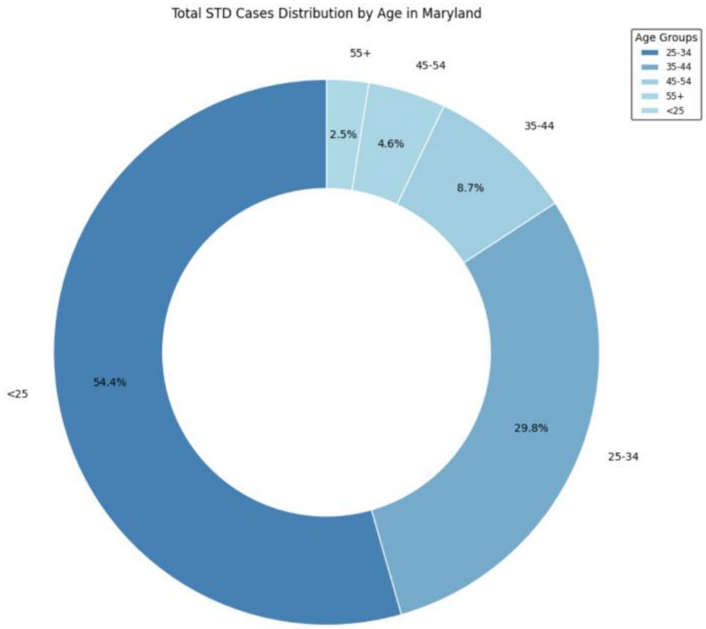
Maryland

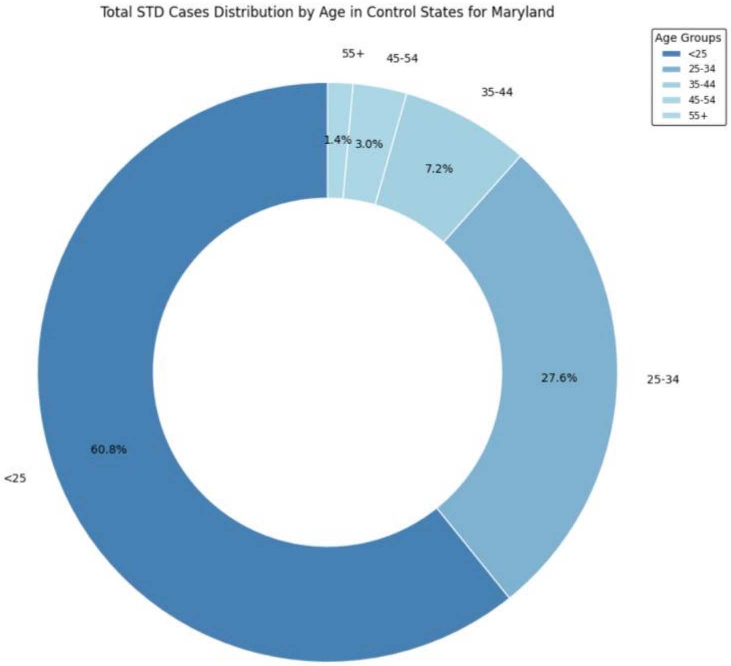
In the chart of the average total STD rate per 100,000 by age distribution in Maryland, it was found that the age group of 25-34 years old had the highest proportion of STD rates, accounting for approximately 43% of the average total. The second-highest proportion was observed in individuals aged below 25 years, contributing to about 30.8% of the average total. In comparison, the control states for Maryland exhibited a similar trend, with the two age groups maintaining the highest distribution. Similarly, the age group of 25-34 years old took the lead in the control states, with an average total STD rate of around 40.4%, while the less than 25 years old age group followed closely with a 38.8% share.





Upon examining the total STD cases by age distribution in Maryland, the age group of less than 25 years old emerged as having the largest proportion of total STD cases, constituting approximately 54.4% of the total. The age group between 25-34 years old ranked second, with an average share of 29.8% in total STD cases. This pattern was consistent with the findings in the control states for Maryland. In these control states, the age group of less than 25 years old had the largest proportion of total STD cases, with an average of 60.8%, while the age group of 25-34 years old held the second-largest share, with 27.6% of total cases.



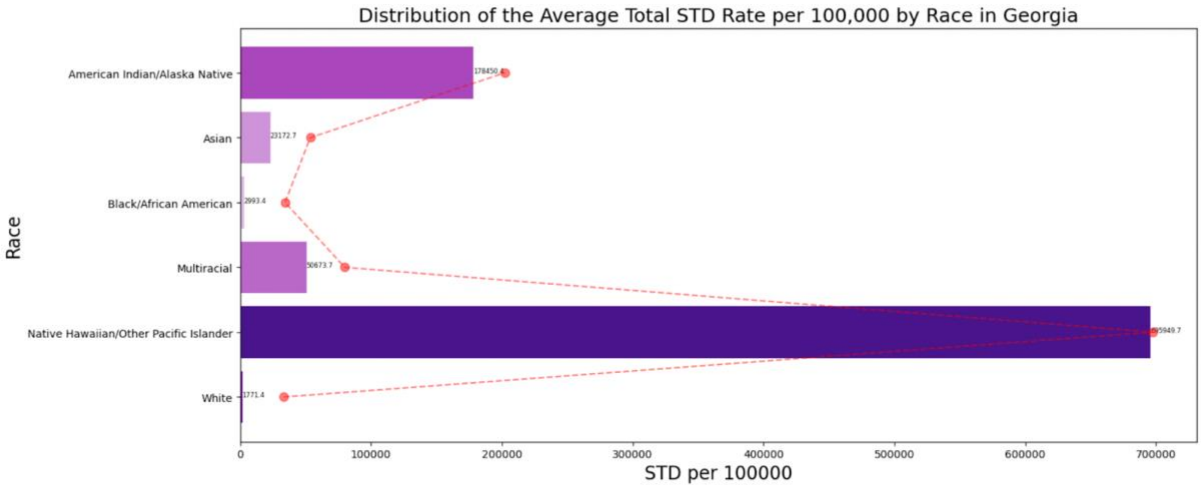


Generally, the 25-34 and under 25 age groups continuously had the greatest rates, according to both the average total STD cases per 100,000 and the total STD cases by age in Maryland and its control states. Both age groups continued to be the top in both measures, despite the fact that the rankings of the age groups changed depending on the average total number of STD cases and the overall number of STD cases.

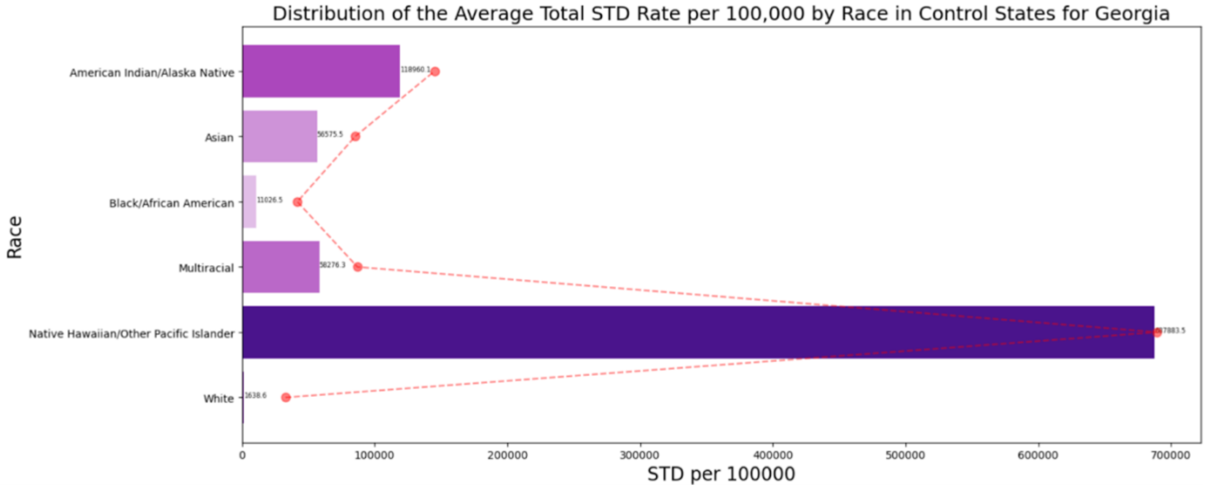
Differences Between Race

Georgia

In the histogram illustrating the average total STD rate per 100,000 by race in Georgia, it is evident that the Native Hawaiian/Other Pacific Islander community exhibits the highest average total STD rate with 1,784,504 cases per capita. Following closely in second place, the American Indian/Alaska Native community experiences a rate of 695,949.7 cases per capita.



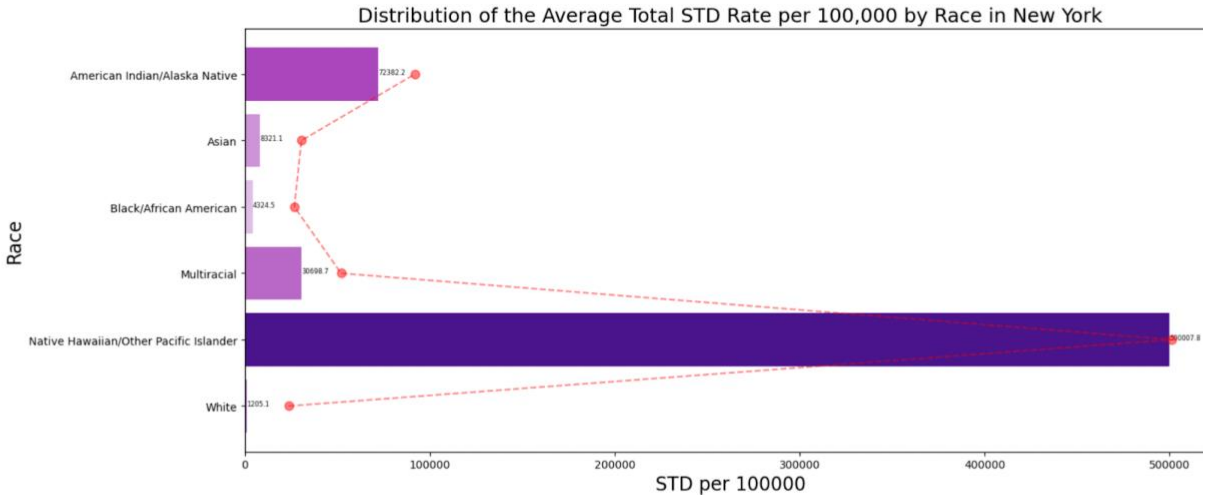
Upon comparison with control states for Georgia, the findings reveal a similar pattern. The Native Hawaiian/Other Pacific Islander population retains the highest average total STD rate, amounting to 687884 cases per capita. Meanwhile, the American Indian/Alaska Native community holds the second-highest position with an average total STD rate of 118960 cases per capita.



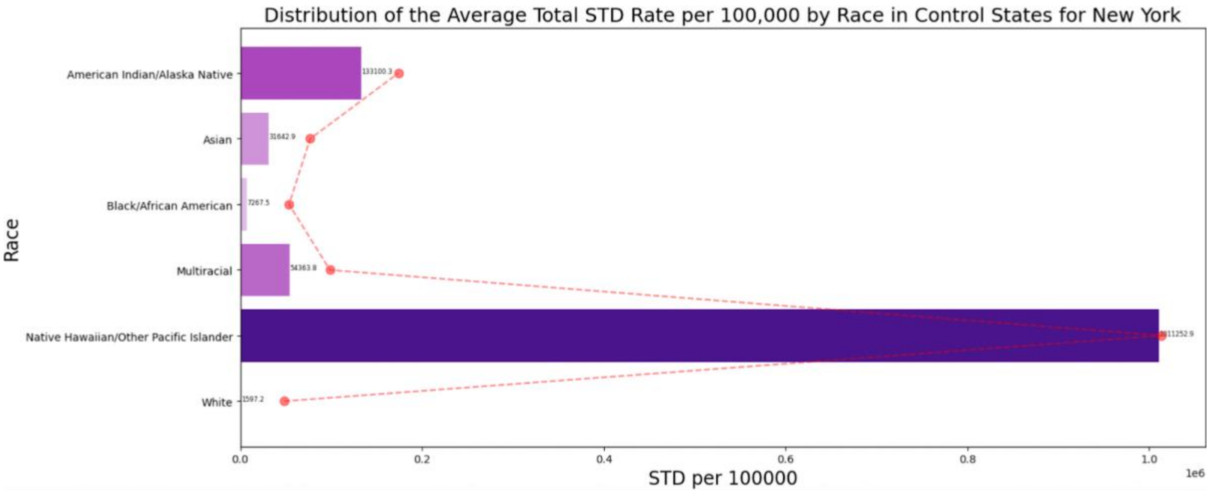
These results demonstrate a clear disparity in STD rates among different racial groups in Georgia and its control states. The rankings indicate that the Native Hawaiian/Other Pacific Islander community is disproportionately affected by STDs, consistently ranking first in both Georgia and the control states. Similarly, the American Indian/Alaska Native population consistently ranks second in both instances, further highlighting the disparities in public health outcomes among racial groups.

New York

In the histogram illustrating the average total STD rate per 100,000 by race in New York, it is evident that the Native Hawaiian/Other Pacific Islander community exhibits the highest average total STD rate with 500007.8 cases per capita. Following closely in second place, the American Indian/Alaska Native community experiences a rate of 72382 cases per capita.



Upon comparison with control states for New York, the findings reveal a similar pattern. The Native Hawaiian/Other Pacific Islander population retains the highest average total STD rate, amounting to 1011253 cases per capita. Meanwhile, the American Indian/Alaska Native community holds the second-highest position with an average total STD rate of 133100 cases per capita.

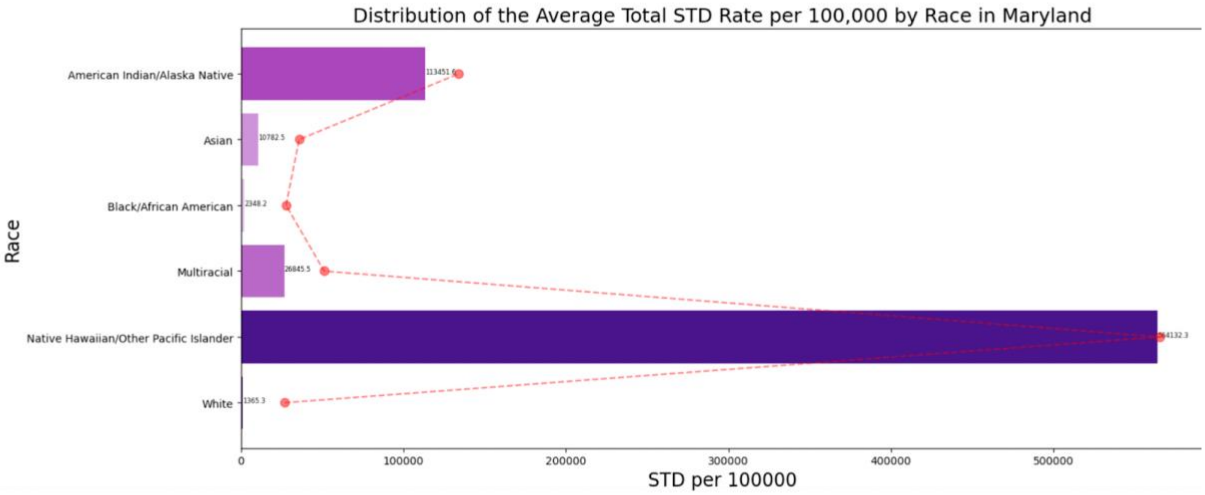


These results show a clear difference in STD rates between different racial groups in New York and its control states. The rankings show that STDs disproportionately affect Hawaii and other Pacific Island communities, consistently ranking first in both Georgia and control states. Also, the American Indian/Alaska Native population consistently ranks second in both cases, further highlighting disparities in public health outcomes between racial groups.

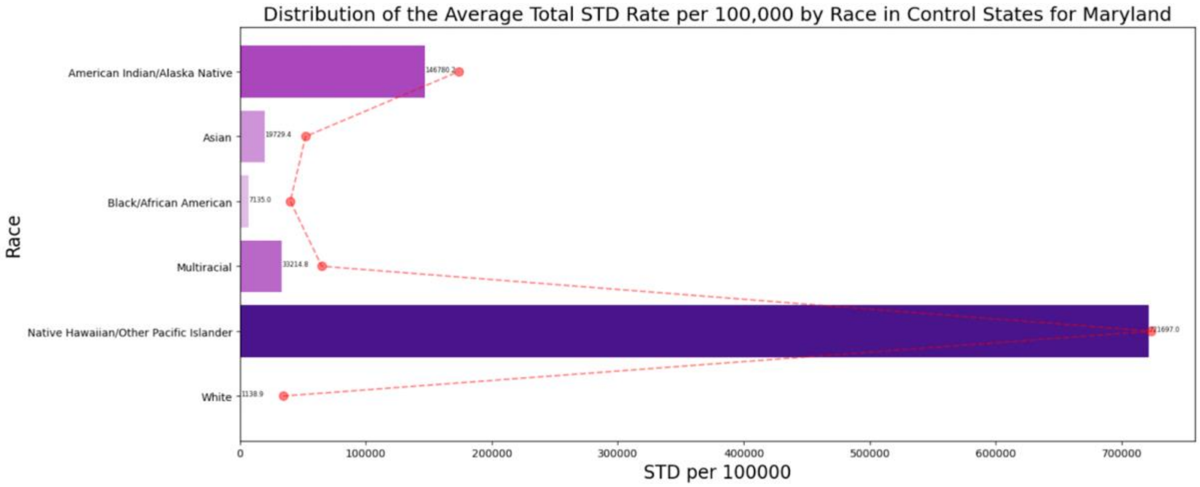
Maryland

In the histogram illustrating the average total STD rate per 100,000 by race in Maryland, it is evident that the Native Hawaiian/Other Pacific Islander community exhibits the highest average total STD rate with

564132 cases per capita. Following closely in second place, the American Indian/Alaska Native community experiences a rate of 113452 cases per capita.



Upon comparison with control states for Maryland, the findings reveal a similar pattern. The Native Hawaiian/Other Pacific Islander population retains the highest average total STD rate, amounting to 721697 cases per capita. Meanwhile, the American Indian/Alaska Native community holds the second-highest position with an average total STD rate of 146780 cases per capita.

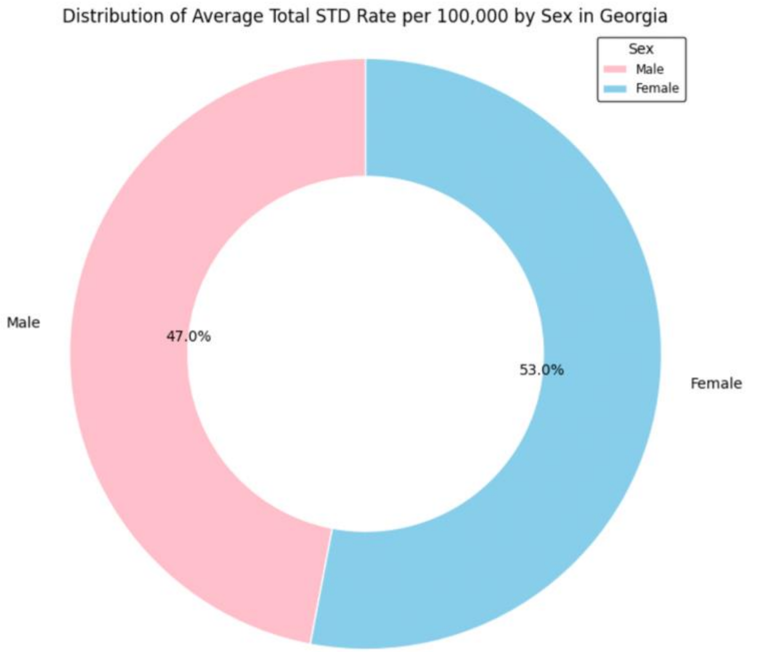


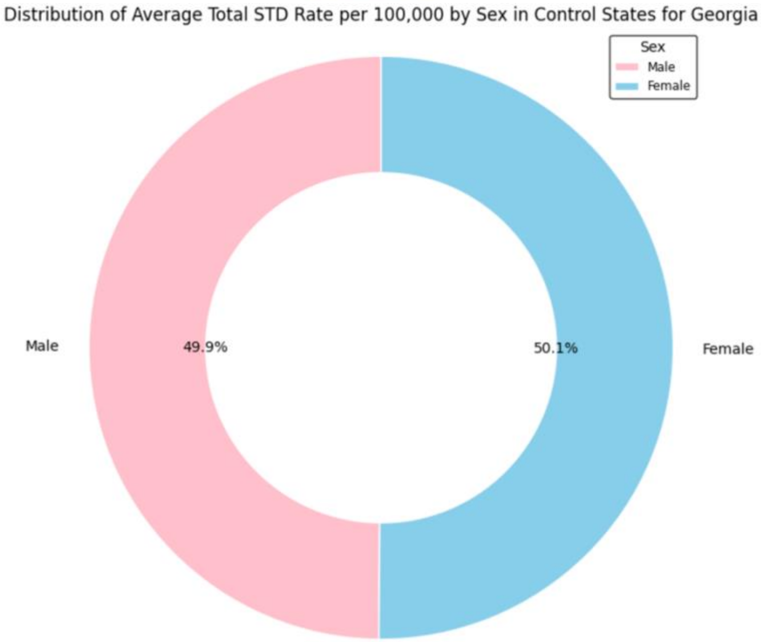
These results demonstrate marked differences in STD rates between different racial groups in Maryland and its control states. Rankings show that sexually transmitted diseases disproportionately affect communities in Hawaii and other Pacific islands, consistently ranking first in both Maryland and control states. Additionally, the Alaska Native/Native American population consistently ranked second in both cases, further highlighting disparities in public health outcomes between racial groups.

Differences Between Gender

Georgia

An examination of the distribution of the average total STD rate per 100,000 by sex in Georgia reveals that females account for a slightly higher proportion of the distribution at 53%, while males represent 47% of the distribution. This disparity between the sexes is also observed in the control states for Georgia, with females making up 50.1% of the distribution and males constituting 49.9%.

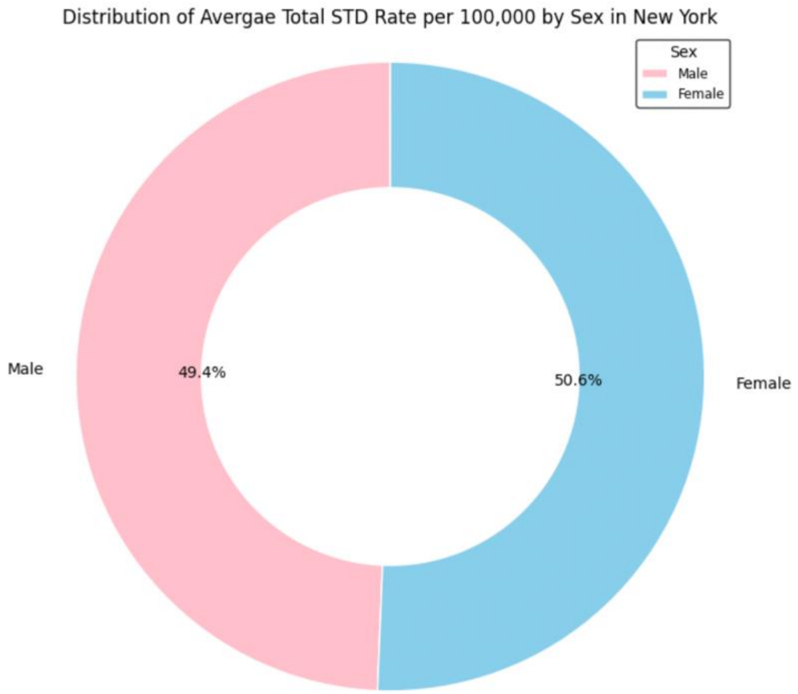

Distribution of Average Total STD Rate per 100,000 by Sex in Georgia

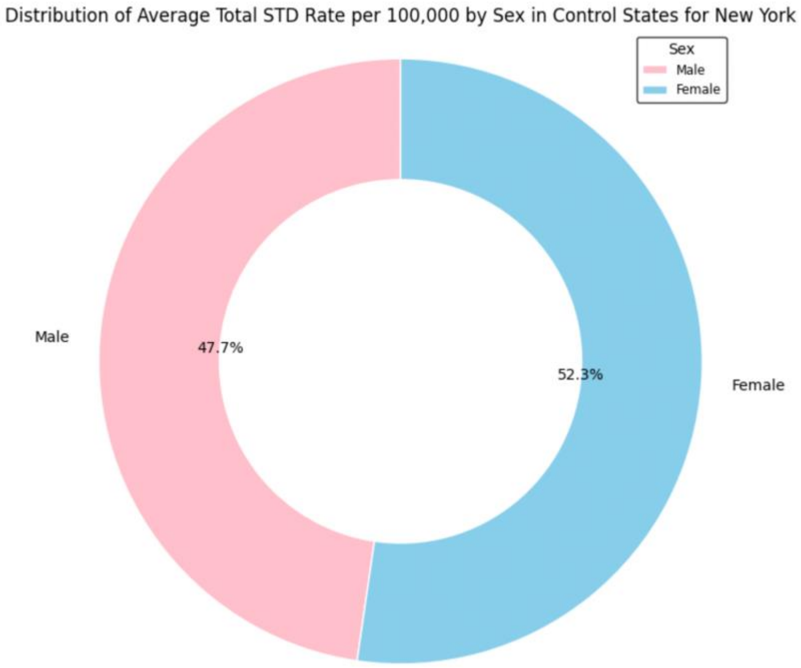

Distribution of Average Total STD Rate per 100,000 by Sex in Control States for Georgia

In both cases, the difference in the distribution of the average total STD rate between males and females is relatively small, indicating a fairly balanced distribution across the sexes. Despite females having a marginally higher share in both Georgia and its control states, the proportions remain closely aligned, suggesting that both males and females are similarly affected by STDs in these regions.

New York

An examination of the distribution of the average total STD rate per 100,000 by sex in New York reveals that females account for a slightly higher proportion of the distribution at 50.6%, while males represent 49.4% of the distribution. This disparity between the sexes is also observed in the control states for New York, with females making up 52.3% of the distribution and males constituting 47.7%.

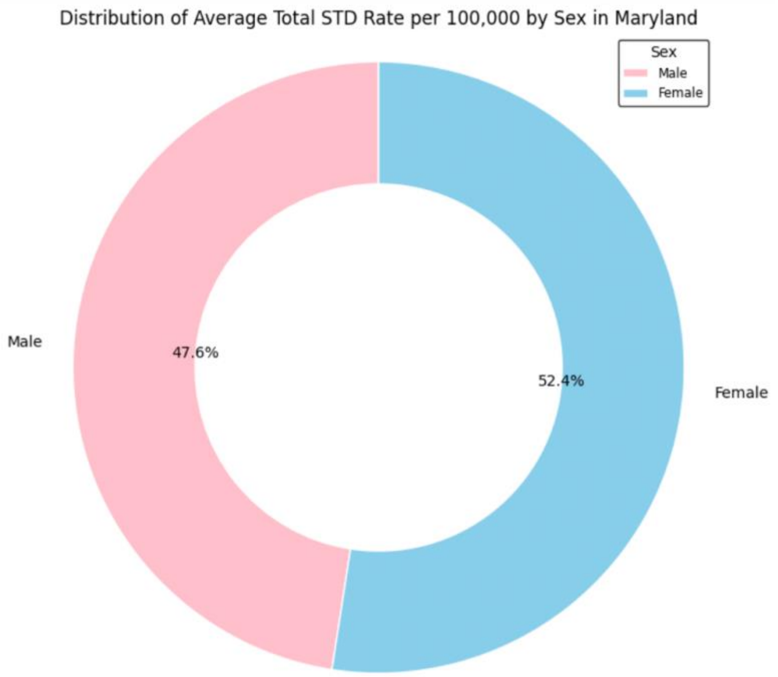




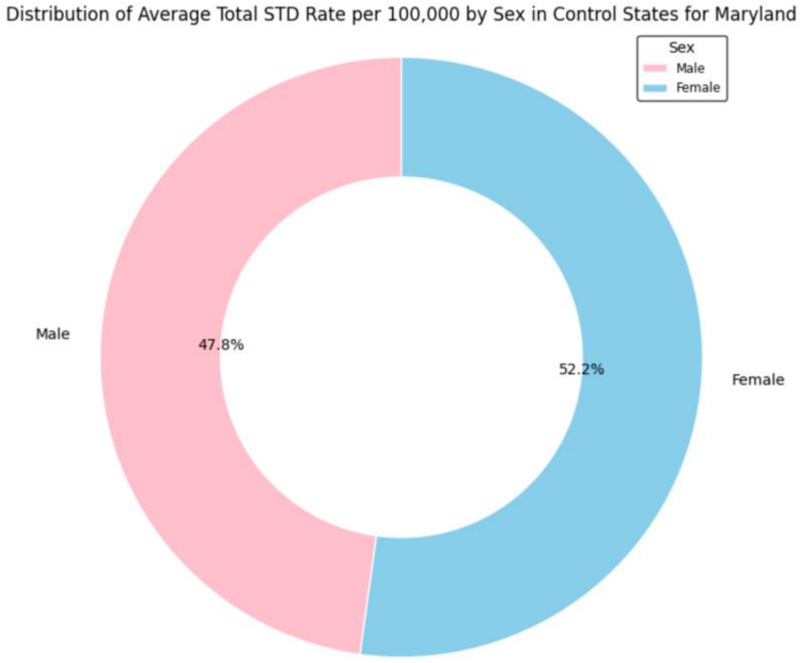
In both cases, the difference between the distribution of the average total number of sexually transmitted diseases between men and women is relatively small, which indicates a fairly balanced distribution between the sexes. Despite slightly higher proportions of women in both New York and its control states, the proportions are closely aligned, suggesting that both men and women have similar STD rates in those areas.

Maryland

An examination of the distribution of the average total STD rate per 100,000 by sex in Maryland reveals that females account for a slightly higher proportion of the distribution at 52.4%, while males represent 47.8% of the distribution. This disparity between the sexes is also observed in the control states for Maryland, with females making up 52.2% of the distribution and males constituting 47.2%.



Distribution of Average Total STD Rate per 100,000 by Sex in Maryland

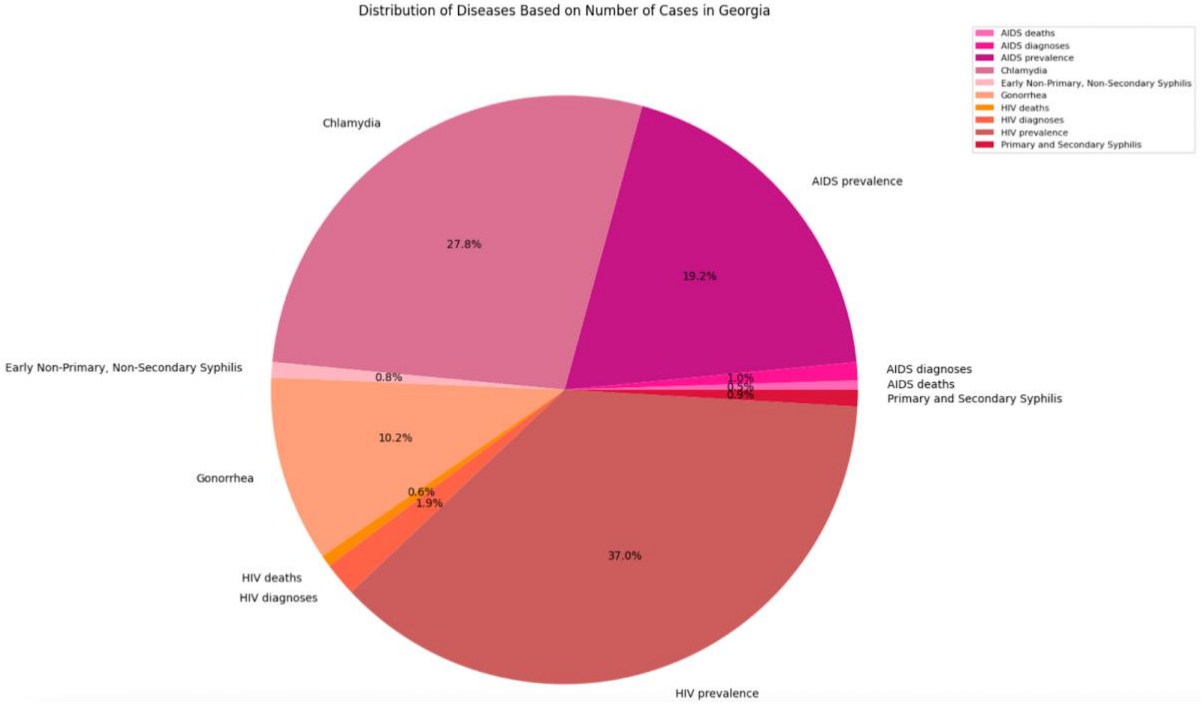


Distribution of Average Total STD Rate per 100,000 by Sex in Control States for Maryland

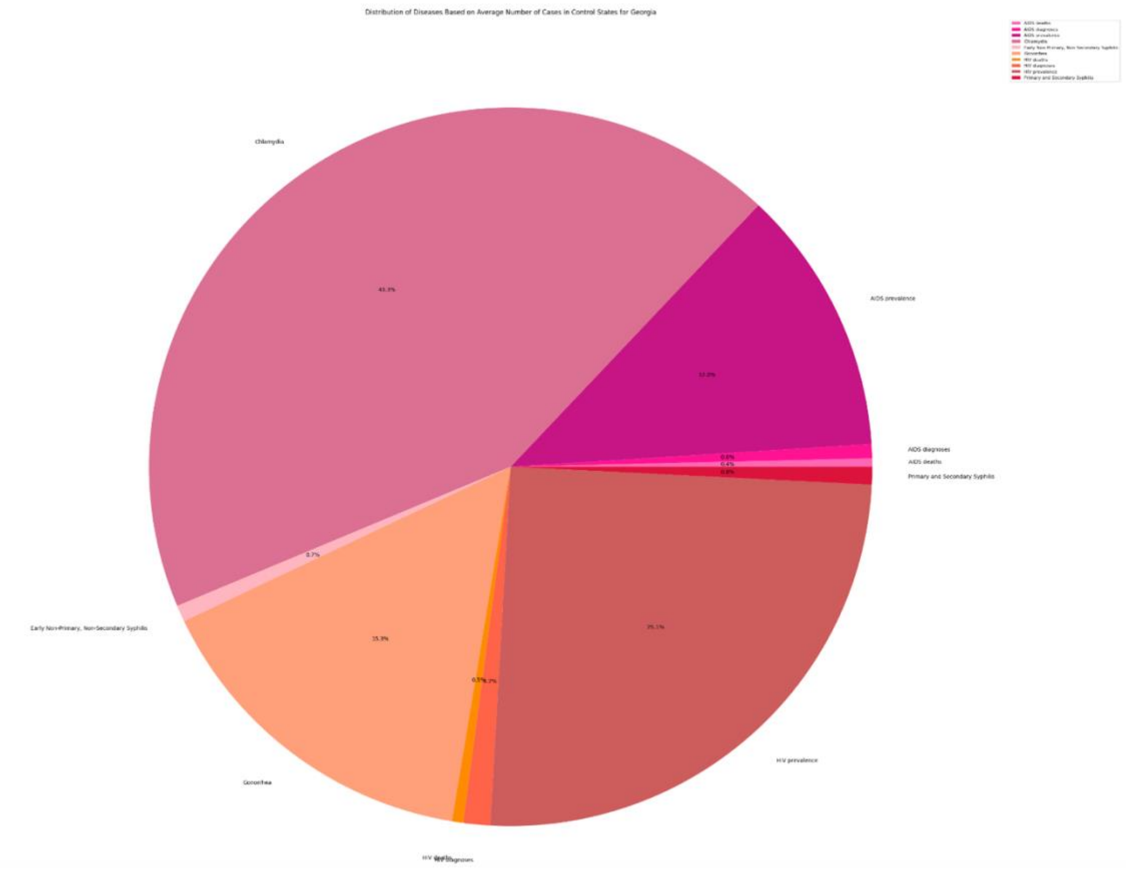
In both cases, the difference in the distribution of the mean total number of sexually transmitted infections between men and women is relatively small, indicating a fairly balanced distribution between men and women. Despite slightly higher proportions of women in New York and its governing states, the proportions are nearly identical, suggesting that both men and women in these areas have similar rates of STDs.

Differences Between STD Cases

Georgia

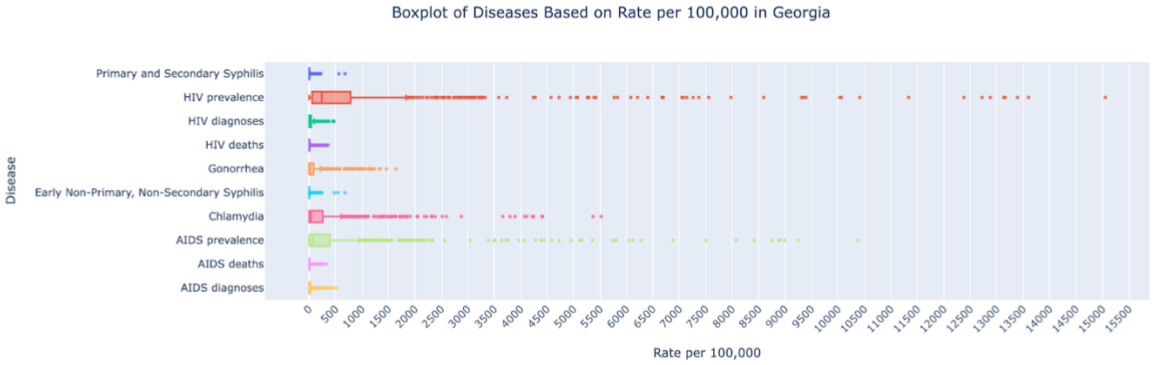


According to the distribution of diseases based on the number of cases in Georgia, the four most prevalent sexually transmitted diseases (STDs) are HIV prevalence, Chlamydia, AIDS prevalence, and Gonorrhea. In Georgia, HIV prevalence accounts for 37% of cases, Chlamydia contributes 27.8%, AIDS prevalence makes up 19.2%, and Gonorrhea represents 10.2% of the cases. These four diseases appear to be the most common STDs in the state.

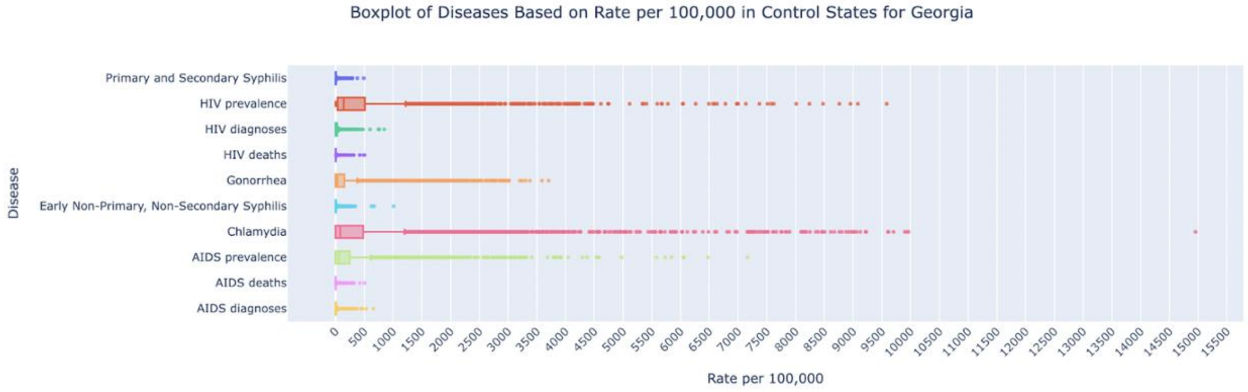


In comparison, the average number of cases in the control states shows a different distribution. Chlamydia has the highest proportion, accounting for 43.3% of cases, followed by HIV prevalence with 25.1%, Gonorrhea at 15.3%, and AIDS prevalence with 12% of cases. Despite the variation in proportions, HIV, Chlamydia, AIDS, and Gonorrhea remain the most prevalent STDs in both Georgia and the control states.

Based on the boxplot analysis of diseases in Georgia, three key metrics were used to compare the prevalence of HIV, AIDS, and chlamydia: median, minimum, and maximum rates per 100,000. The results showed that HIV had the highest median rate per 100,000, with a value of 238.3, followed by AIDS with a median rate of 63.4 per 100,000, and chlamydia with a median rate of 36.75 per 100,000. When it comes to outliers, HIV was the disease with the most outliers, ranging from 0 to 15,000 per 100,000. AIDS had the second most outliers, ranging from 0 to 10,000 per 100,000, followed by chlamydia, with a range of 0 to 5,522.6 per 100,000. These results indicate that HIV is the disease with the most significant variability in rates of prevalence in Georgia.

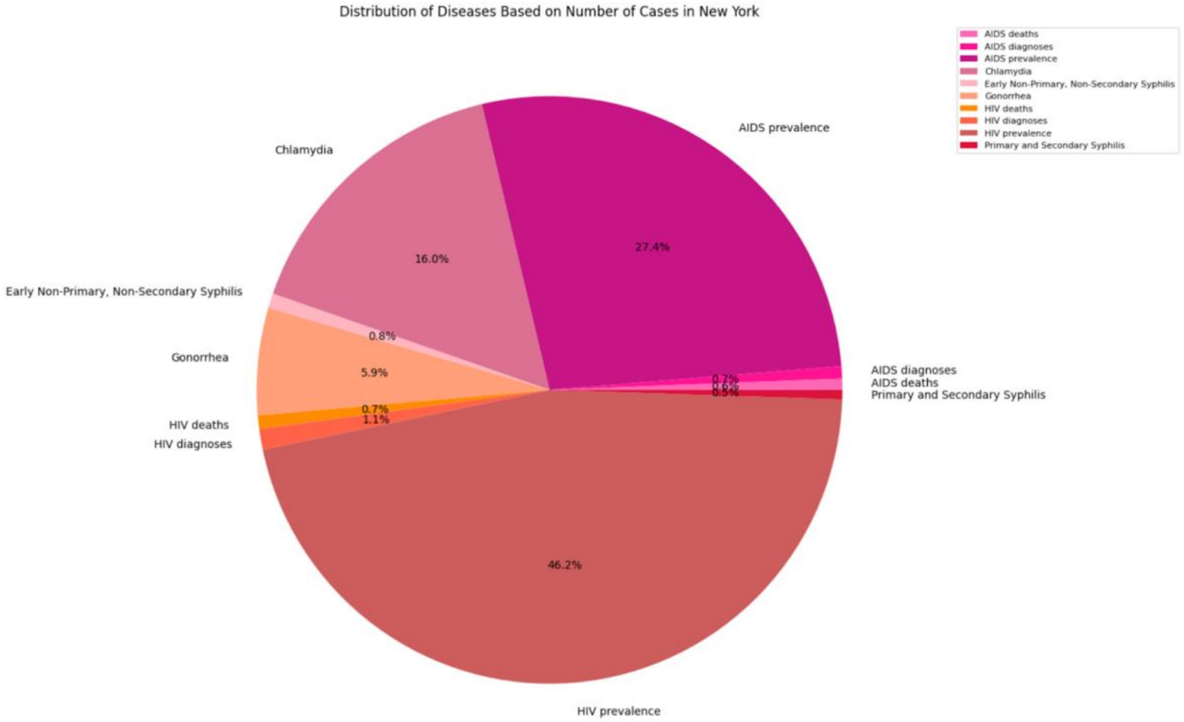


In comparison to control states, HIV prevalence had the highest median rate per 100,000, with a value of 142.35, followed by chlamydia with a median rate of 80.65 per 100,000, and AIDS prevalence with a median rate of 56.35 per 100,000. However, when it comes to outliers, chlamydia had the most, ranging from 0 to 14953 per 100,000, followed by HIV prevalence with a range of 0 to 9583 per 100,000, and AIDS prevalence with a range of 0 to 7164 per 100,000. These results suggest that chlamydia is the disease with the most significant variability in rates of prevalence when compared to control states.

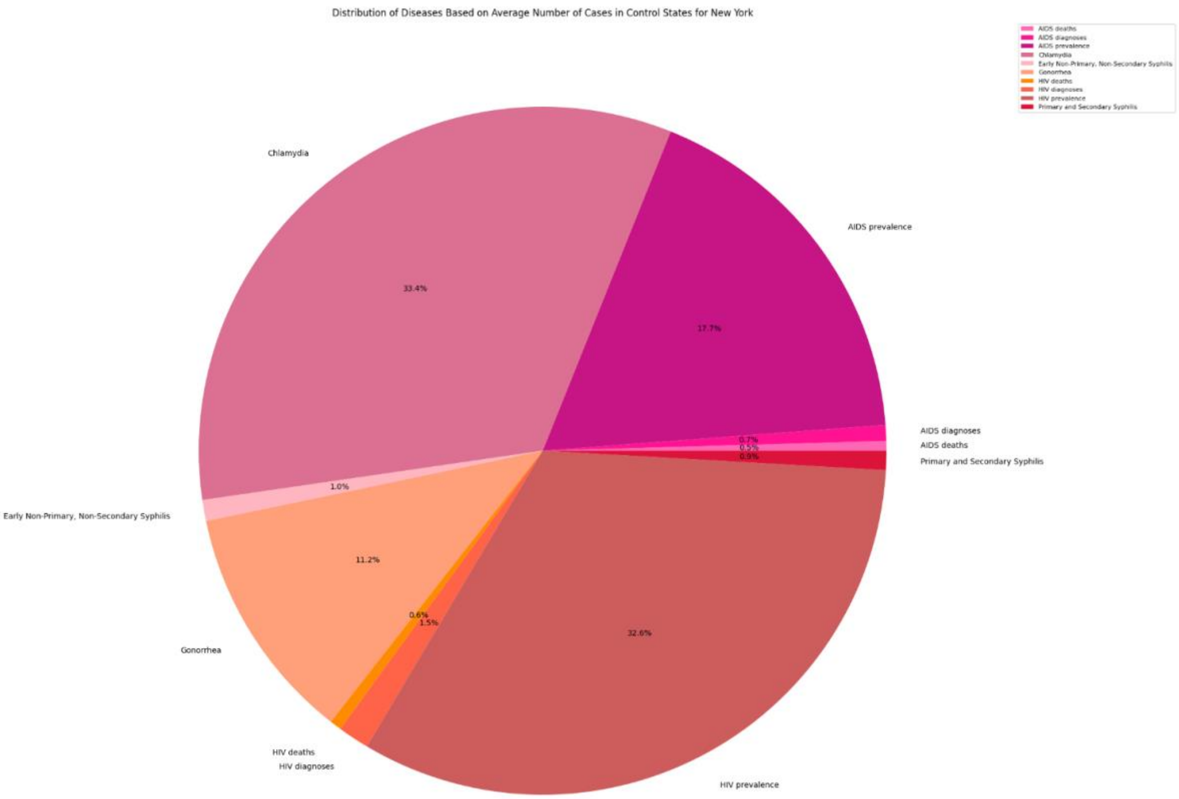


In conclusion, the boxplot indicates that HIV prevalence has the highest median rate of prevalence in Georgia and the most significant variability in rates, while chlamydia has the highest median rate in control states and the most significant variability in rates when compared to control states. The plot also revealed that HIV, AIDS, and chlamydia are the diseases with the highest prevalence rates in Georgia.

New York

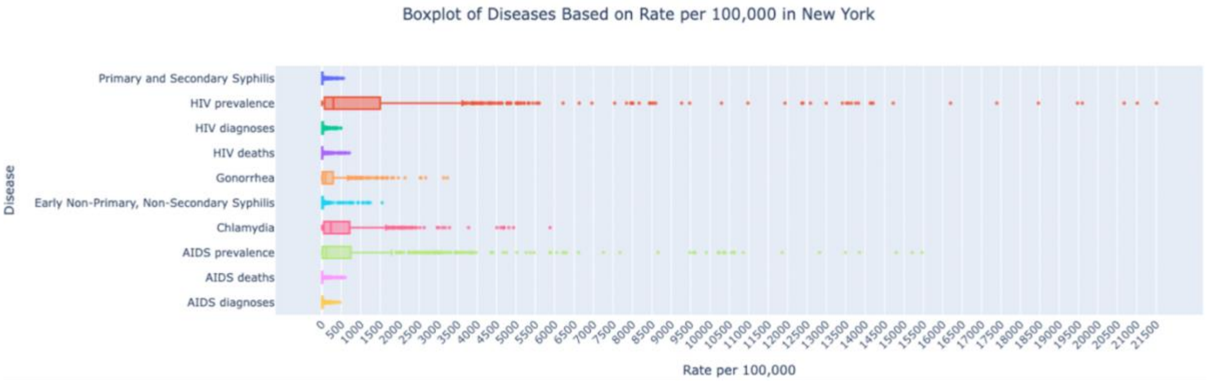


According to the distribution of diseases based on the number of cases in New York, the four most prevalent sexually transmitted diseases (STDs) are HIV prevalence, Chlamydia, AIDS prevalence, and Gonorrhea. In New York, HIV prevalence accounts for 46.2% of cases, AIDS prevalence makes up 27.4%, Chlamydia contributes 16%, and Gonorrhea represents 5.9% of the cases. These four diseases appear to be the most common STDs in the state.

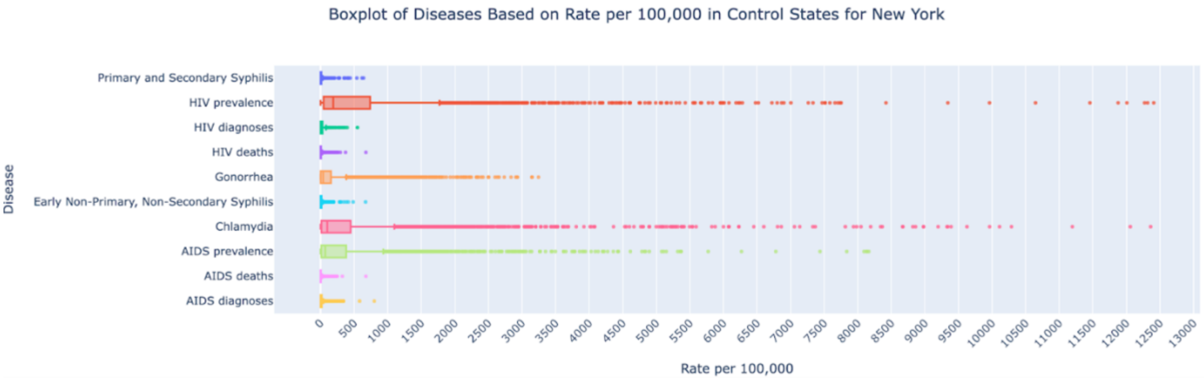


In comparison, the average number of cases in the control states shows a different distribution. Chlamydia has the highest proportion, accounting for 33.4% of cases, followed by HIV prevalence with 32.6%, AIDS prevalence with 17.7% of cases, and Gonorrhea at 11.2%, and Despite the variation in proportions, HIV, Chlamydia, AIDS, and Gonorrhea remain the most prevalent STDs in both New York and the control states.

Based on the boxplot analysis of diseases in New York, three key metrics were used to compare the prevalence of HIV, AIDS, and chlamydia: median, minimum, and maximum rates per 100,000. The results showed that HIV prevalence had the highest median rate per 100,000, with a value of 294.3, followed by chlamydia with a median rate of 229.3 per 100,000, and AIDS prevalence with a median rate of 111.75 per 100,000. When it comes to outliers, HIV was the disease with the most outliers, ranging from 0 to 21,000 per 100,000. AIDS had the second most outliers, ranging from 0 to 15,000 per 100,000, followed by chlamydia, with a range of 0 to 5,881 per 100,000. These results indicate that HIV is the disease with the most significant variability in rates of prevalence in New York.



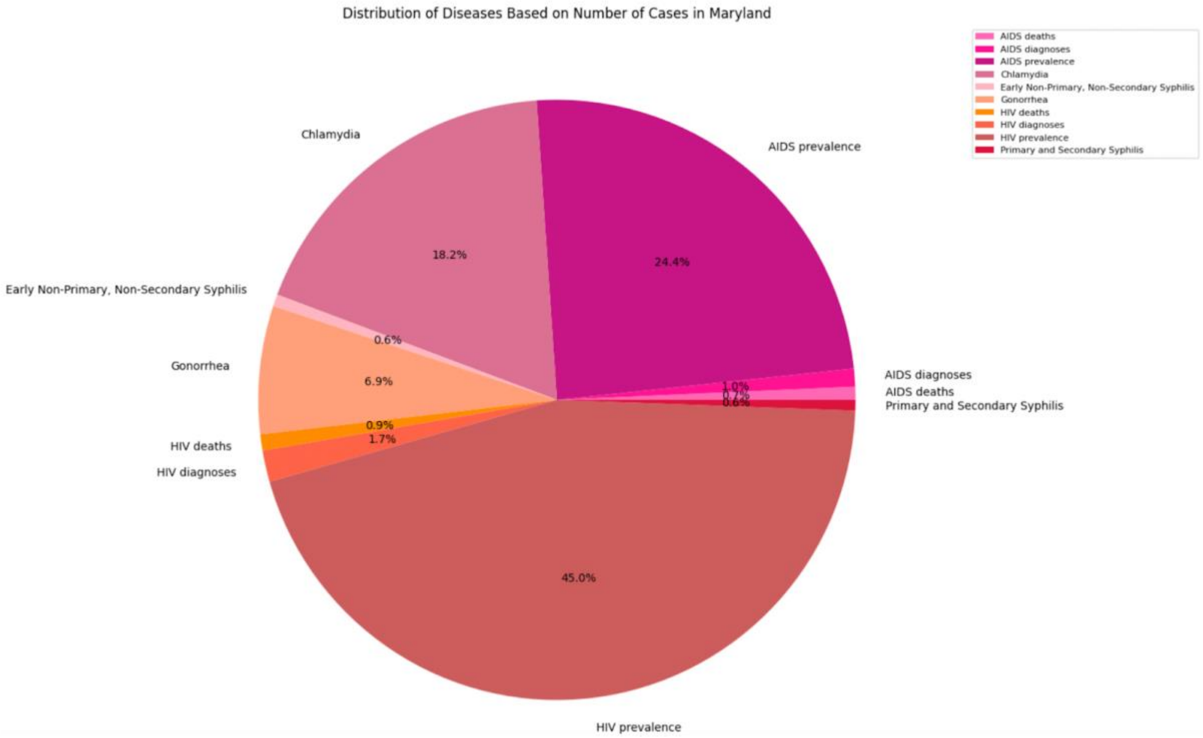
In comparison to control states, HIV prevalence had the highest median rate per 100,000, with a value of 188.9, followed by chlamydia with a median rate of 98.95 per 100,000, and AIDS prevalence with a median rate of 70.8 per 100,000. However, when it comes to outliers, HIV had the most, ranging from 0 to 12406 per 100,000, followed by chlamydia with a range of 0 to 12,359.6 per 100,000, and AIDS with a range of 0 to 8,164.7 per 100,000. These results suggest that chlamydia is the disease with the most significant variability in rates of prevalence when compared to control states.



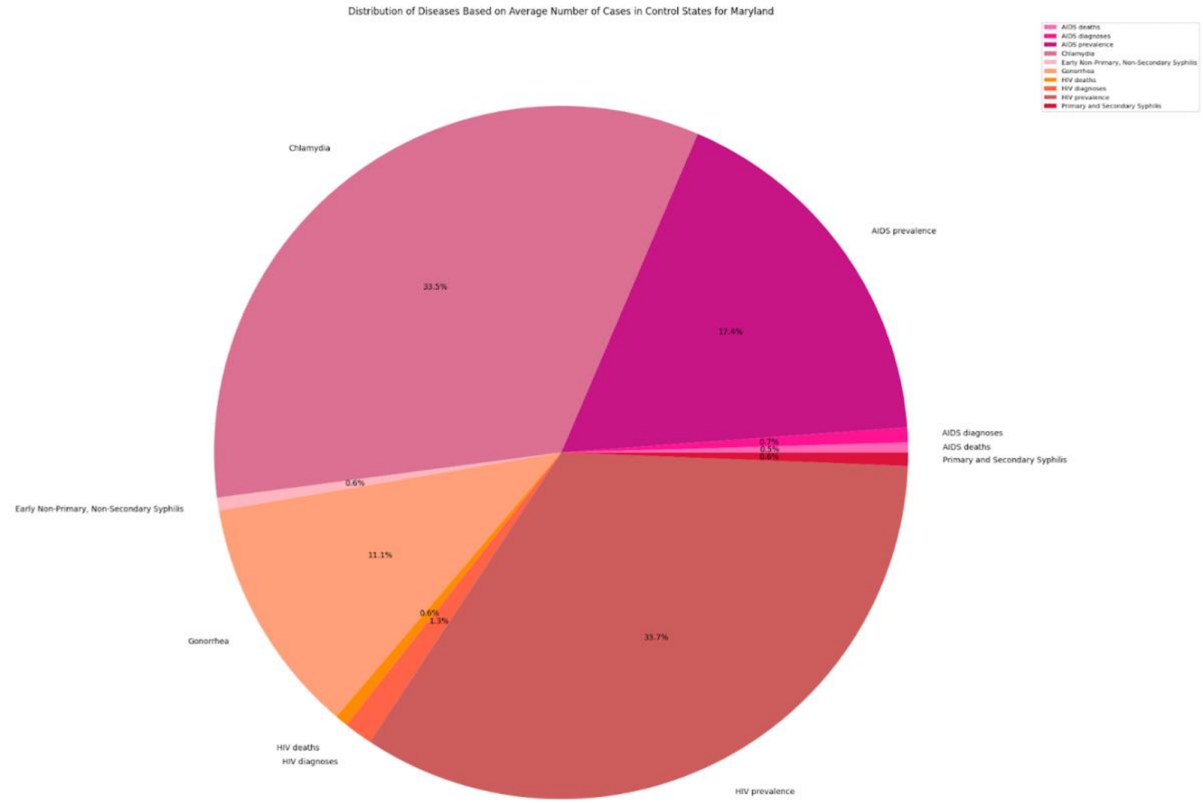
To sum up, the boxplot indicates that HIV prevalence has the highest median rate of prevalence in New York and the most significant variability in rates, while HIV prevalence has the highest median rate in control states and chlamydia has the most significant variability in rates when compared to control states. The plot also revealed that HIV, AIDS, and chlamydia are the diseases with the highest prevalence rates in New York.

Maryland

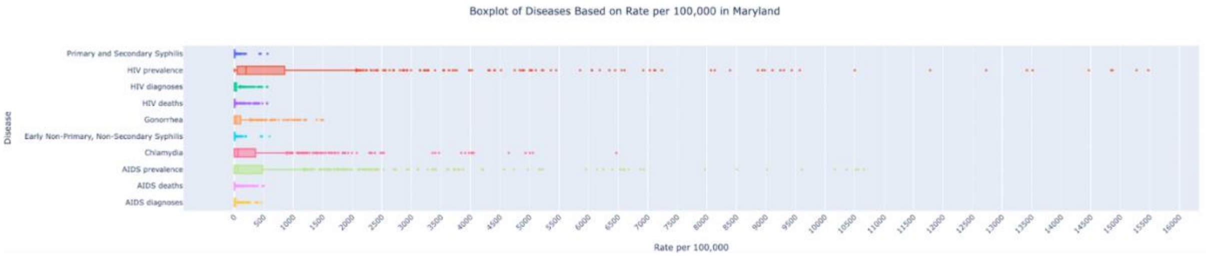
According to the distribution of diseases based on the number of cases in Maryland, the four most prevalent sexually transmitted diseases (STDs) are HIV prevalence, Chlamydia, AIDS prevalence, and Gonorrhea. In Georgia, HIV prevalence accounts for 45% of cases, AIDS prevalence makes up 24.4%, Chlamydia contributes 18.2%, and Gonorrhea represents 6.9% of the cases. These four diseases appear to be the most common STDs in the state.



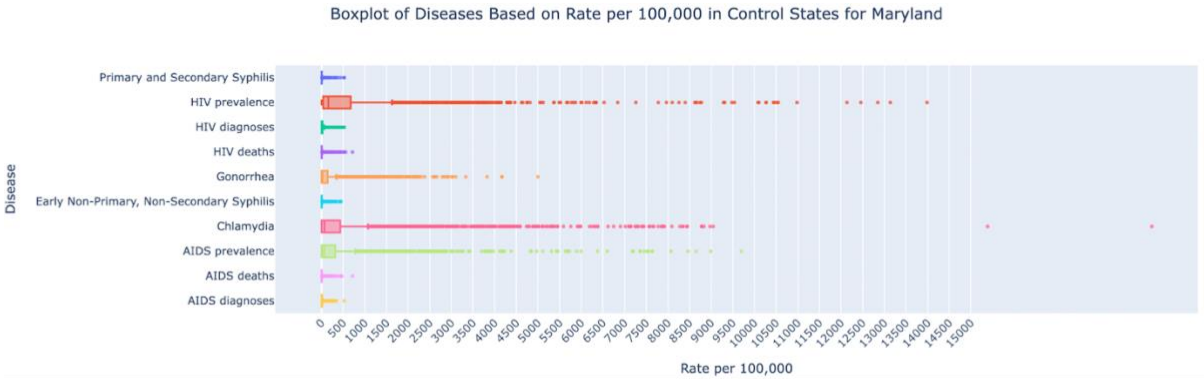
In comparison, the average number of cases in the control states shows a different distribution.
HIV prevalence has the highest proportion, accounting for 33.7% of cases, followed by Chlamydia prevalence with 33.5%, AIDS prevalence with 17.4% of cases, and Gonorrhea at 11.1%. Despite the variation in proportions, HIV, Chlamydia, AIDS, and Gonorrhea remain the most prevalent STDs in both New York and the control states.



Based on the boxplot analysis of diseases in Maryland, three key metrics were used to compare the prevalence of HIV, AIDS, and chlamydia: median, minimum, and maximum rates per 100,000. The results showed that HIV prevalence had the highest median rate per 100,000, with a value of 196.6, followed by chlamydia with a median rate of 60.75 per 100,000, and AIDS prevalence with a median rate of 59.05 per 100,000. When it comes to outliers, HIV was the disease with the most outliers, ranging from 0 to 15474.4 per 100,000. AIDS had the second most outliers, ranging from 0 to 10657.2 per 100,000, followed by chlamydia, with a range of 0 to 6467.7 per 100,000. These results indicate that HIV is the disease with the most significant variability in rates of prevalence in Maryland.



In comparison to control states, HIV prevalence had the highest median rate per 100,000, with a value of 159.1, followed by chlamydia with a median rate of 69.05 per 100,000, and AIDS prevalence with a median rate of 59.55 per 100,000. However, when it comes to outliers, chlamydia had the most, ranging from 0 to 19178.1 per 100,000, followed by HIV prevalence with a range of 0 to 13986.4 per 100,000, and AIDS prevalence with a range of 0 to 9696.7 per 100,000. These results suggest that chlamydia is the disease with the most significant variability in rates of prevalence when compared to control states.



In summary, the boxplot shows that Maryland has the highest median prevalence of HIV prevalence and the highest rate variability. Control states, on the other hand, have the highest median HIV prevalence and the most variable chlamydia rates compared to control states. The plot also showed that HIV, AIDS and chlamydia are the most prevalent diseases in New York.