Analysis of Tick-Bourne Disease Cases Reported to DPH Between 2019-2024

Authors

* Hope Grismer (ORCID: 0000-0000-1234-5678)

Author affiliations

1. College of Public Health, University of Georgia, Athens, GA, USA.
2. Southeastern Center of Excellence in Vector-Bourne Diseases (SECVBD), Gainesville, FL 32611.

Corresponding author: hope.grismer@uga.edu

Disclaimer: The opinions expressed in this article are the author’s own and don’t reflect their employer.

# 1. Abstract

This project explores patterns in tick-borne disease cases reported to the Georgia Department of Public Health (DPH) between 2019 and 2024. The analysis involved extensive data cleaning to retain only confirmed and positive cases, with additional processing to resolve inconsistencies in timelines and reporting. Exploratory data analysis (EDA) was used to examine disease distribution, demographic patterns, geographic hotspots, and data completeness over time. Building on these insights, predictive models were developed to assess whether patient and case-level factors—such as age, disease type, and county—could help identify cases at risk of delayed follow-up. This work theorectically supports improved understanding of disease surveillance data and highlights areas for strengthening tick-borne disease reporting and case management systems in Georgia.

# 2. Introduction

### 2.0.1 2.1 Background

In the past two decades, the study of tick-borne disease has gained increasing attention due to rising incidence rates and expanding geographic ranges of disease-carrying ticks. While early medical and veterinary research in the 20th century identified key tick-borne illnesses such as Rocky Mountain spotted fever and Lyme disease, recent decades have seen an acceleration in the detection of new pathogens, greater public health concern, and expanded surveillance efforts across the globe. Researchers have documented the ecology and behavior of ticks, explored the role of climate and land use in tick habitat expansion, and analyzed patterns of disease incidence in both human and animal populations. Despite the growing body of research, many knowledge gaps remain, particularly in understanding co-infections, long-term health effects, and effective prevention strategies. Tick-borne diseases (TBDs) are primarily transmitted through the bite of infected hard-bodied ticks (family Ixodidae), and include bacterial, viral, and protozoan pathogens. Today, TBDs are recognized as a major public health concern in both rural and suburban settings, with transmission risk influenced by ecological, climatic, and sociocultural factors.

### 2.0.2 2.1.1 General Background Information

Tick-borne illnesses are diseases that humans can contract from the bite of an infected tick. Ticks are small, blood-sucking arachnids that are often found in wooded, grassy, or suburban areas. When a tick feeds on an infected animal, it can acquire pathogens—such as bacteria, viruses, or parasites—that can then be transmitted to humans. Examples of tick-borne diseases include, but are not limited to, Lyme disease, Anaplasmosis, and Babesiosis. These diseases are of increasing public health concern, especially as the geographic range and seasonality of ticks expand due to environmental and ecological changes.

### 2.0.3 2.1.2 Transmission and Reservoirs

Ticks act as both vectors and reservoirs for a variety of pathogens, transmitting disease to humans and animals during blood meals. Most tick-borne diseases are zoonotic, with small mammals, deer, and birds acting as key reservoirs in natural cycles. Transmission to humans typically occurs when a nymph or adult tick feeds for an extended period, allowing the pathogen to transfer from the tick’s salivary glands into the host’s bloodstream. Common tick-borne pathogens include Borrelia burgdorferi (the agent of Lyme disease), Rickettsia rickettsii (Rocky Mountain spotted fever), Anaplasma phagocytophilum, Ehrlichia chaffeensis, and viruses such as Powassan and Heartland virus. Ticks themselves are influenced by multiple environmental factors, including temperature, humidity, vegetation, and host availability. Human exposure risk is increased in areas with high host density, edge habitats between forest and residential zones, and during seasons with peak tick activity. Additionally, land use changes—such as suburban development, deforestation, and agricultural expansion—can modify tick habitats and affect pathogen dynamics.

### 2.0.4 2.1.3 Ecological and Health Impacts

The ecological consequences of tick-borne disease extend beyond human health. Tick populations and their associated pathogens can alter wildlife health and community composition. Infected animals may suffer sublethal effects that reduce fitness, while high burdens of tick parasitism can influence population dynamics, particularly in rodents and ungulates. From a public health perspective, TBDs can cause a range of acute and chronic symptoms in humans, from fever and fatigue to long-term neurological and musculoskeletal complications. Co-infections—when an individual is infected with more than one pathogen—are increasingly recognized and may complicate diagnosis and treatment. The public health burden is also shaped by disparities in access to health care, awareness of TBDs, and availability of diagnostic testing, which may result in underreporting or misdiagnosis. The costs associated with tick-borne disease—including medical care, productivity loss, and prevention efforts—represent a growing concern for health systems, particularly in endemic areas.

### 2.0.5 2.1.4 Relevance

Tick-borne diseases have been documented in all regions of the United States, with the southeastern, northeastern, and upper midwestern states experiencing particularly high incidence rates. Climate models suggest that warming temperatures and shifting precipitation patterns may expand the range of ticks and lengthen their active season, potentially increasing the geographic footprint and seasonality of tick-borne disease transmission. In the southeastern U.S., for example, Amblyomma americanum (the lone star tick) has expanded in abundance and geographic range over the past decades, contributing to increased cases of ehrlichiosis and the emergence of Alpha-Gal Syndrome, a red meat allergy triggered by tick bites. Understanding the ecology, distribution, and drivers of tick-borne disease is therefore crucial to predicting future trends, identifying at-risk populations, and developing effective prevention and response strategies.

# 3. 3. The Data

# 4. 3.1 Data Acquisition & Description

The data used in this analysis was obtained through my internship with the Georgia Department of Health, in collaboration with the Centers for Disease Control and Prevention (CDC). Specifically, it is derived from the Nationally Notifiable Diseases Surveillance System (NNDSS) and focuses on tick-borne illnesses. This data set provides key insights into the trends, distribution, and characteristics of reported cases, supporting public health efforts in disease monitoring and prevention.2.1.5 Description of Data and Data Source

This data set includes all human cases reported to the Georgia Department of Public Health (GA-DPH) between 2019 and 2024. It was graciously provided by DPH professionals in the epidemiology department and includes all available data for each case reported across the state over this five-year period. Proper de-identification practices—such as removal of first name, last name, and date of birth (DOB)—were followed. The dataset includes variables such as disease type, patient demographics (gender, age, district, state, county, and ZIP code), medical details (incident ID, report ID, lab test dates, results, serotype, and titers), and disease outcomes (death status, fever, migraines, and administrative status). Reported diagnoses include Toxoplasmosis, Dengue, Rocky Mountain Spotted Fever, Anaplasma phagocytophilum, Malaria, Lyme Disease, West Nile Virus, Chikungunya, LaCrosse Virus, Zika, Ehrlichia chaffeensis, and various forms of Typhus. Some cases, based on adherence to specific case definitions and subsequent lab results, are categorized as suspect, probable, confirmed, or unconfirmed. By examining case reports, public health professionals can better understand the prevalence, demographics, and regional variations of these diseases. This analysis aims to further inform the public health understanding of tick-borne disease in Georgia.

## 4.1 3.1.2 Questions and Hypotheses to Be Addressed

This project seeks to explore temporal, spatial, and epidemiologic patterns of vector-borne disease (VBD) in Georgia using reported case data from 2019–2024. Specifically, it aims to address the following questions:

* How complete is the data for vector-borne diseases, and how has data completeness evolved over time?
* What is the timeline between key dates (onset, lab test, and date of entry), excluding same-day onset and lab results, and how does this timeline vary over time?
* Can we predict whether a confirmed tick-borne disease case will result in a longer-than-typical diagnostic delay based on demographic and geographic features (county of origin of the case and Gini Index)?
* We hypothesized that tick-borne diseases would exhibit seasonal trends corresponding with tick life cycles, that higher incidence would be observed in areas with more natural and edge habitat, and that case counts would correlate with both environmental conditions and patterns of human interaction with tick habitats.

# 5. 4. Methods

## 5.1 4.1 Workflow Summarized

* Cleaned and prepared surveillance data on tick-borne disease cases reported in Georgia from 2019–2024, including filtering for confirmed positive cases and addressing issues with missing values and inconsistent timelines.
* Conducted exploratory data analysis (EDA) to examine temporal trends, geographic distribution, age patterns, diagnostic test outcomes, and data completeness, using visualizations to inform hypotheses and modeling decisions.
* Developed and evaluated multiple predictive models to assess whether demographic and case-specific factors could explain delays in case follow-up, using cross-validation and test/train splitting to assess model quality and performance.

## 5.2 4.2 Uncleaned Data

The data contains 6255 entries representing discrete cases of tick-bourne diseases. For each case, some or all of the following data variables were collected: disease type, patient demographics (gender, age, district, state, county, and ZIP code), medical details (incident ID, report ID, lab test dates, results, serotype, and titers), and disease outcomes (death status, fever, migranes, and administrative status). For some cases, specific variables were not reported and, therefore, are reported as blank spaces in the data set.

## 5.3 4.3 Initial Data Cleaning

The data cleaning and preprocessing workflow begins with loading essential R packages for data handling (dplyr, tidyr, readxl), summary visualization (skimr), and file path management (here). The raw dataset, stored in an Excel file, is imported into R for structured processing. To assess data quality and structure, we utilize functions such as glimpse(), head(), and summary(), which provide an overview of column names, data types, and potential inconsistencies like missing or unexpected values. Additionally, skimr::skim() offers a detailed summary, highlighting numerical distributions and missing data patterns.

The initial cleaning step removes records where the STATUS field is either missing or not classified as “CONFIRMED,” ensuring that only officially verified cases are included. The dataset is then refined further by retaining only cases with a confirmed “positive” test result, excluding records with missing, blank, or ambiguous values (e.g., “inconclusive”). This process initially removed 15 cases that did not meet the case definition in Georgia, leaving 95 valid records. Upon further review, filtering criteria were adjusted to ensure all relevant positive cases were retained while systematically excluding non-positive values such as “Negative,” “Equivocal,” “Not detected,” and other irrelevant categories. This refinement resulted in a more accurate dataset and a clearer summary table of key case information. The cleaned dataset is saved at each stage to maintain version control and reproducibility.

#### 5.3.0.1 4.3.1 Looked for Missing Data & General Trends

* Excluded cases with a missing or non-“CONFIRMED” STATUS to retain only confirmed cases.
* Refined the dataset by retaining only confirmed “positive” test results, excluding missing or ambiguous results, leaving 95 valid cases.
* Additional filtering ensured correct identification of positive cases while excluding non-positive values.

#### 5.3.0.2 Addressed Data Quality Issues

* Identified negative time values between lab test dates, indicating data entry errors.
* Implemented a filtering step to remove records with invalid date sequences.

## 5.4 4.4 Exploratory/Descriptive Analysis

#### 5.4.0.1 4.4.1 Identified Disease Burden Trends & Hotspot Counties

* Generated a histogram to visualize the distribution of tick-borne diseases in Georgia between 2019 and 2024.
* Created a bar chart highlighting the five most common diseases.
* Specifically analyzed Lyme Disease cases, focusing on test results.

#### 5.4.0.2 4.4.2 Checked Age Distribution

* Verified that no cases had missing age values in the cleaned dataset.
* Conducted an exploratory analysis of disease counts by age group.

#### 5.4.0.3 4.4.3 Evaluated Data Completeness Over Time

* Extracted onset date information to track missing data trends.
* Created a plot distinguishing between blank and non-blank STATUS entries over time to identify periods of high missingness, indicating possible data entry errors or reporting gaps.

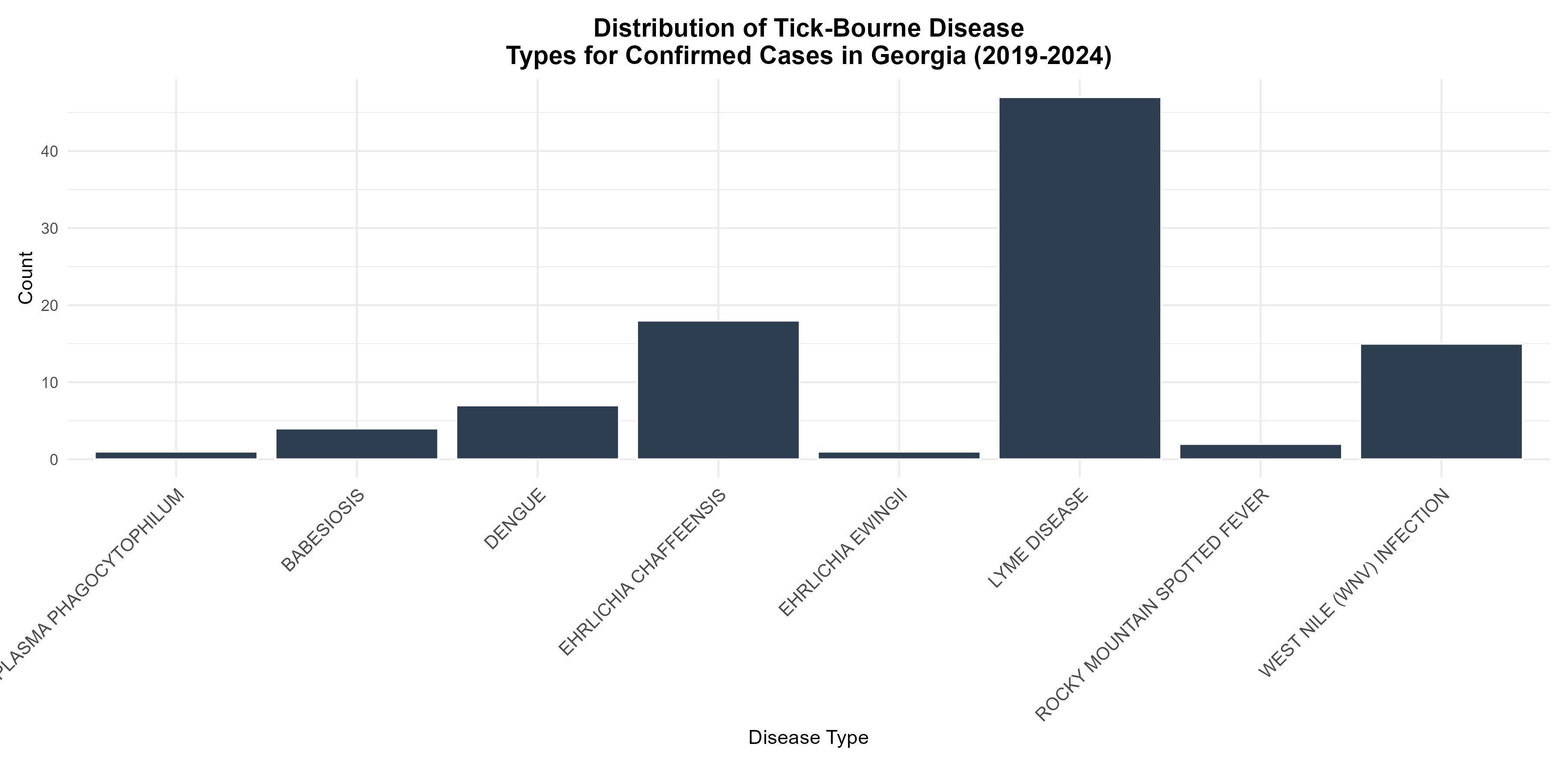
#### 5.4.0.4 4.4.4 Analyzed Timeline Between Key Case Events

* Calculated time differences for:
  + Onset date (DOO) to first lab test date (LABDATE1)
  + First lab test (LABDATE1) to second lab test (LABDATE2)
  + Second lab test (LABDATE2) to case update date (UPDATEDATE)
* Summed these values to determine the total case duration.
* Saved the processed dataset as a CSV file for further analysis.
* Created a formatted HTML table to present the extracted time intervals.

## 5.5 4.5 EDA Revelant Takeaways

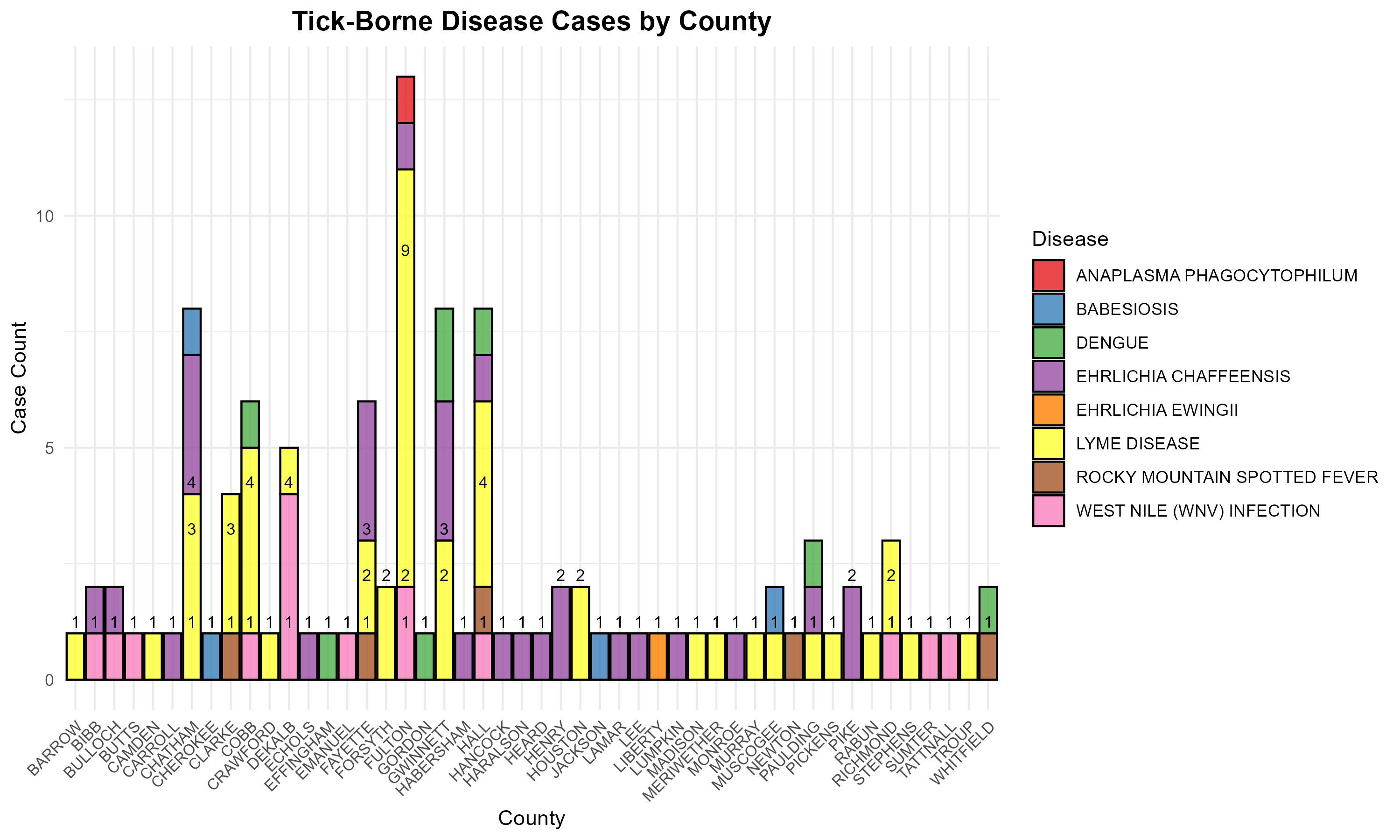
#### 5.5.0.1 4.5.1 Most Common Tick-Borne Diseases

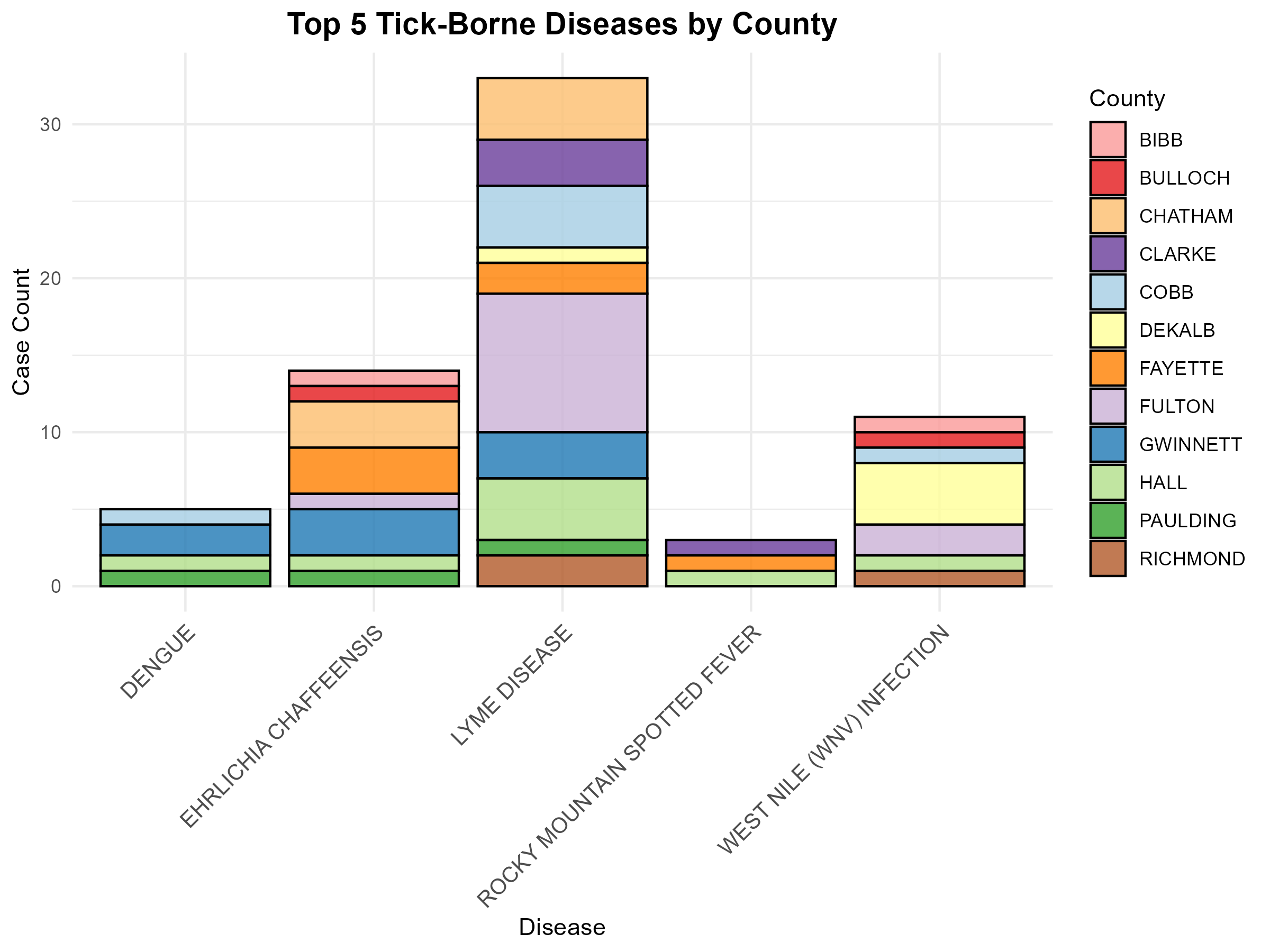
The dataset is filtered to identify the five most frequently reported tick-borne diseases. A bar chart visualizes their case counts, offering an overview of significant public health concerns.

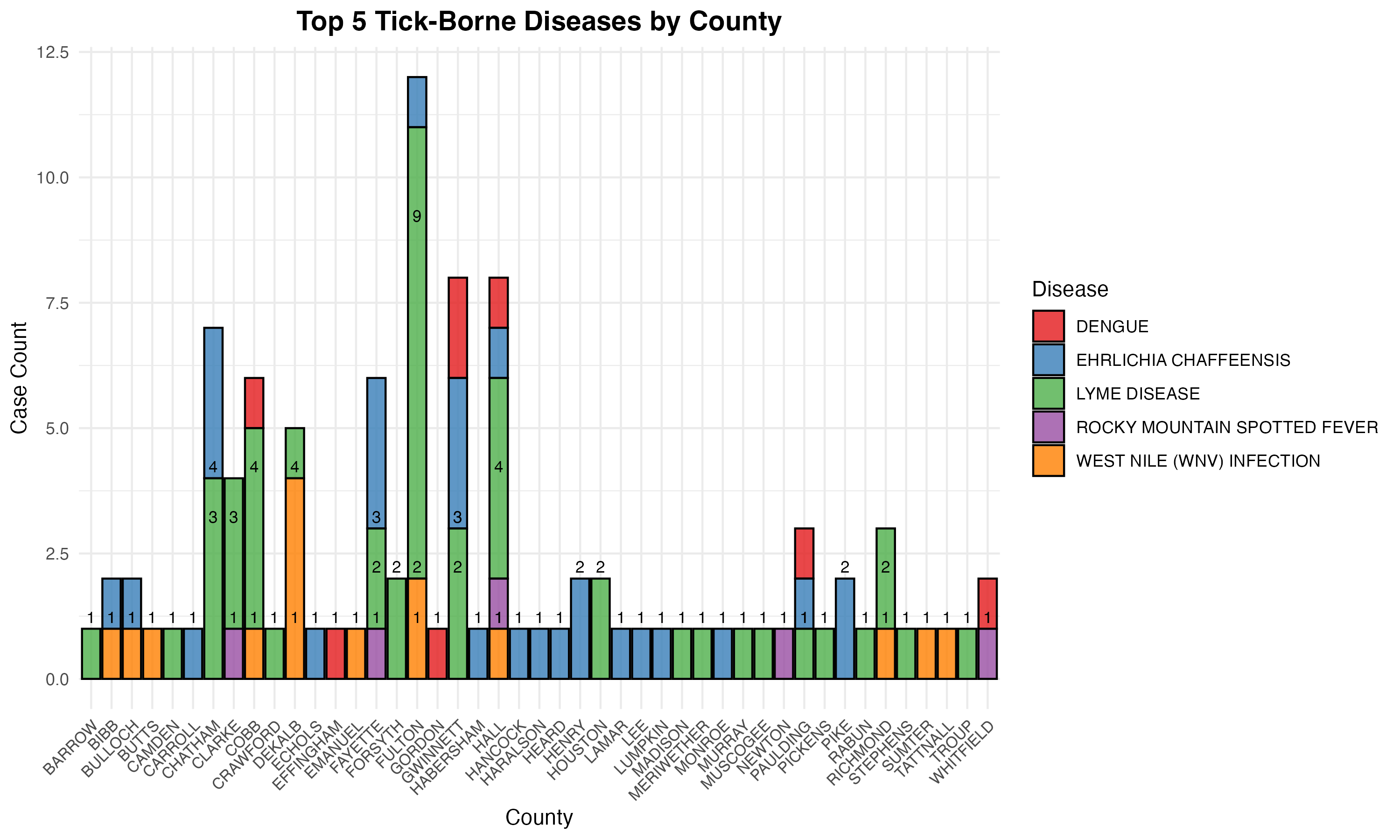


#### 5.5.0.2 4.5.2 Geographic Distribution

A grouped bar chart maps case counts across Georgia counties, categorized by disease type to analyze disease burden trends and identify hotspot counties. Due to the density of information, this visualization is refined to focus on the top five tick-borne diseases and their distribution across the 12 counties with the highest confirmed case counts.

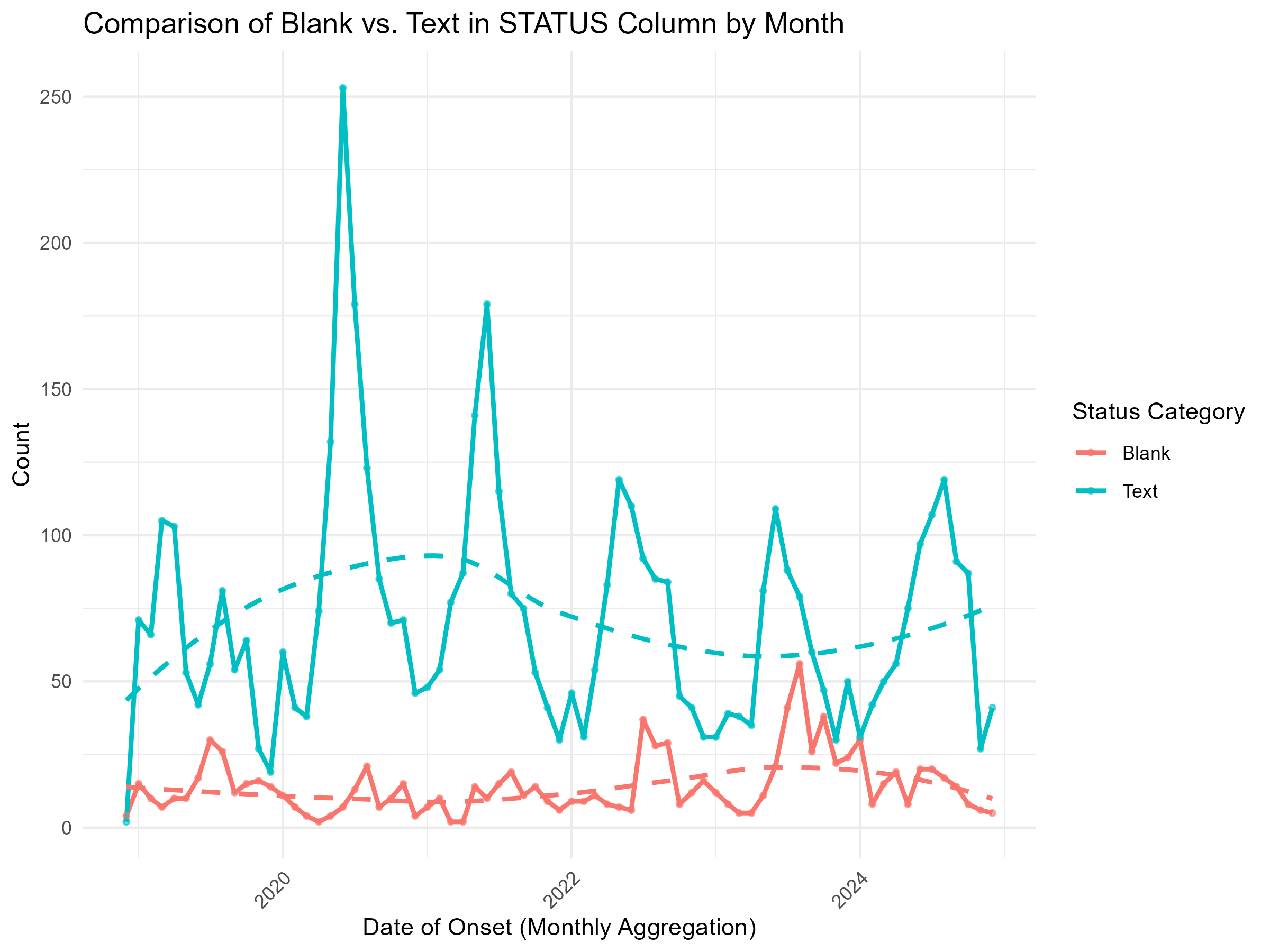






#### 5.5.0.3 4.5.3 Data Completeness Over Time

To evaluate how data completeness has evolved, the dataset is processed to extract onset date information. A visualization distinguishes between blank and non-blank entries in the STATUS column over time, helping to identify periods with high missingness. Spikes in blank values may indicate data entry errors, system issues, or reporting gaps that should be addressed before conducting further analysis.



For more of my initial exploratory data analysis, please see the Supplemental Figures section at the bottom of the report.

#### 5.5.0.4 4.5.4 Timeline Analysis of Key Case Events

Given the frequent loss-to-follow-up observed in tick-borne illness reporting, an analysis is conducted to examine time gaps between critical dates, including:

* Onset date (DOO) to first lab test date (LABDATE1)
* First lab test (LABDATE1) to second lab test (LABDATE2)
* Second lab test (LABDATE2) to case update date (UPDATEDATE)

These durations are summed to determine the total case timeline for each record. The processed dataset is saved as a CSV file for further analysis, and a formatted HTML table is generated to present the extracted time intervals in a structured manner. Data Quality Adjustments Upon reviewing the time calculations, inconsistencies were identified where negative time values appeared between initial lab results and follow-up tests. These errors, likely due to data entry mistakes, made it impossible to determine the correct dates. To address this, a filtering step was added to remove erroneous records where follow-up test dates occurred before initial lab dates.

## 5.6 4.6 Statistical Analysis

To evaluate whether diagnostic delays could be predicted from case-level characteristics, a binary outcome variable (delay\_over\_threshold) was created to indicate whether the total time from symptom onset to final update exceeded 30 days. Predictor variables included patient age, sex, county of residence, and disease type. Age was normalized, and categorical variables were one-hot encoded.

Three classification models—logistic regression, decision tree, and random forest—were trained using the tidymodelsframework. Model performance was assessed using 5-fold cross-validation repeated three times. Metrics included accuracy and area under the receiver operating characteristic curve (AUC).

Both basic bivariate analysis and statistical modeling was conducted. For the statistical modeling, three classification models were trained and evaluated using 5-fold cross-validation repeated 3 times. Performance was assessed using accuracy and area under the ROC curve (AUC).

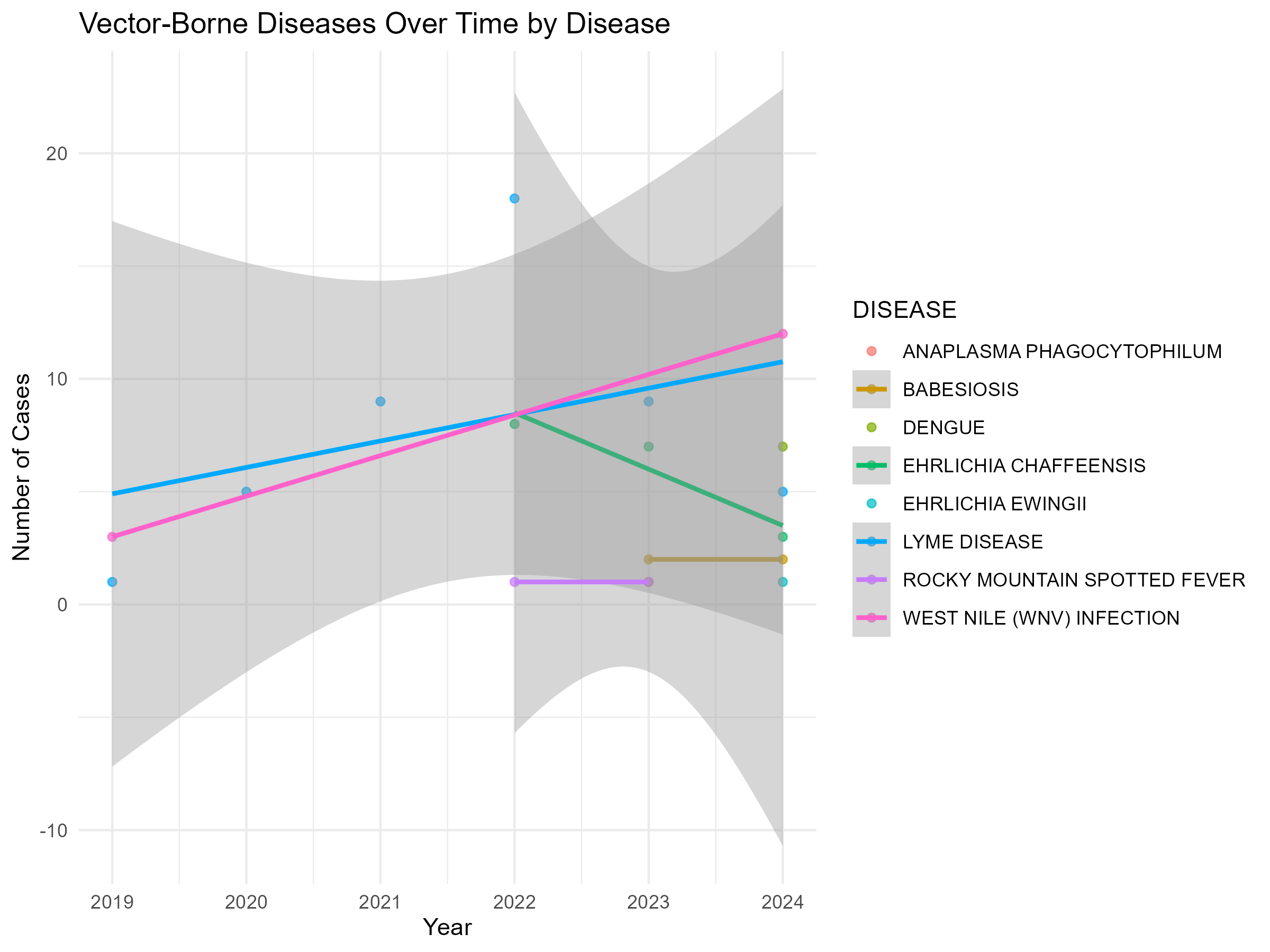
# 6. 5. Results

## 6.1 5.1 Basic statistical analysis

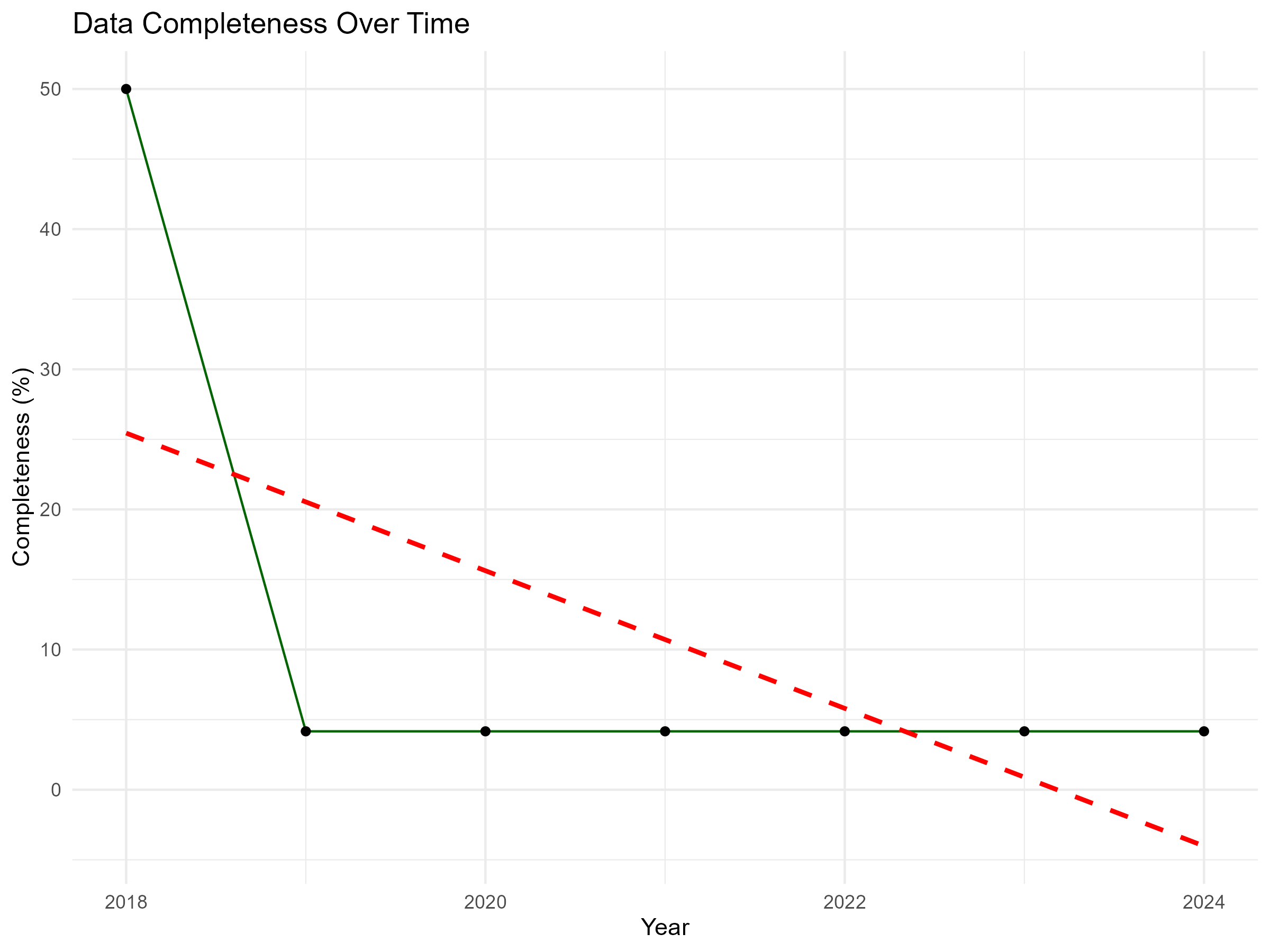
For the first key question related to the disease trends over time (in years). In looking at the plot, it appears that there have been increases over time were observed in most of the diseases. However, the rate of increase was relatively slow as indicated by the lack of statistical significance of these trends (p>0.5) which may indicate an attribution to the increase in reporting rather than an increase in overall disease burden. In the full analysis and second iteration of this, these trends will be compared year to year, comparing each year to the 2018 standard.

| term  <chr> | estimate  <dbl> | s td.error  <dbl> | s tatistic  <dbl> | p.value  <dbl> |
| --- | --- | --- | --- | --- |
| ( I n tercept) | -2 3 2 8 .5151515 | 1 57 3.052665 | 1 . 48025251 | 0 .1729390 |
| YEAR | 1 .1515152 | 0.777581 | 1 . 48089418 | 0 .1727716 |
| B A BESIOSIS | 0 .4242424 | 5.484565 | 0 . 07735207 | 0 .9400358 |
| DENGUE | 4 .8484848 | 6.364775 | 0 . 76176850 | 0 .4656907 |
| E HRLICHIA C H A FFEENSIS | 5 .0000000 | 5.157889 | 0 . 96938889 | 0 .3576794 |
| E HRLICHIA EWINGII | * 1 .1515152 | 6.364775 | 0 . 18092002 | 0 .8604390 |
| LYME DISEASE | 8 .5606061 | 4.963745 | 1 . 72462659 | 0 .1186814 |
| ROCKY MOUNTAIN SPOTTED FEVER | 0 .5757576 | 5.484565 | 0 . 10497781 | 0 .9186958 |
| WEST NILE (WNV) I NFECTION | 8 .2272727 | 5.593721 | 1 . 47080497 | 0 .1754208 |

9 rows



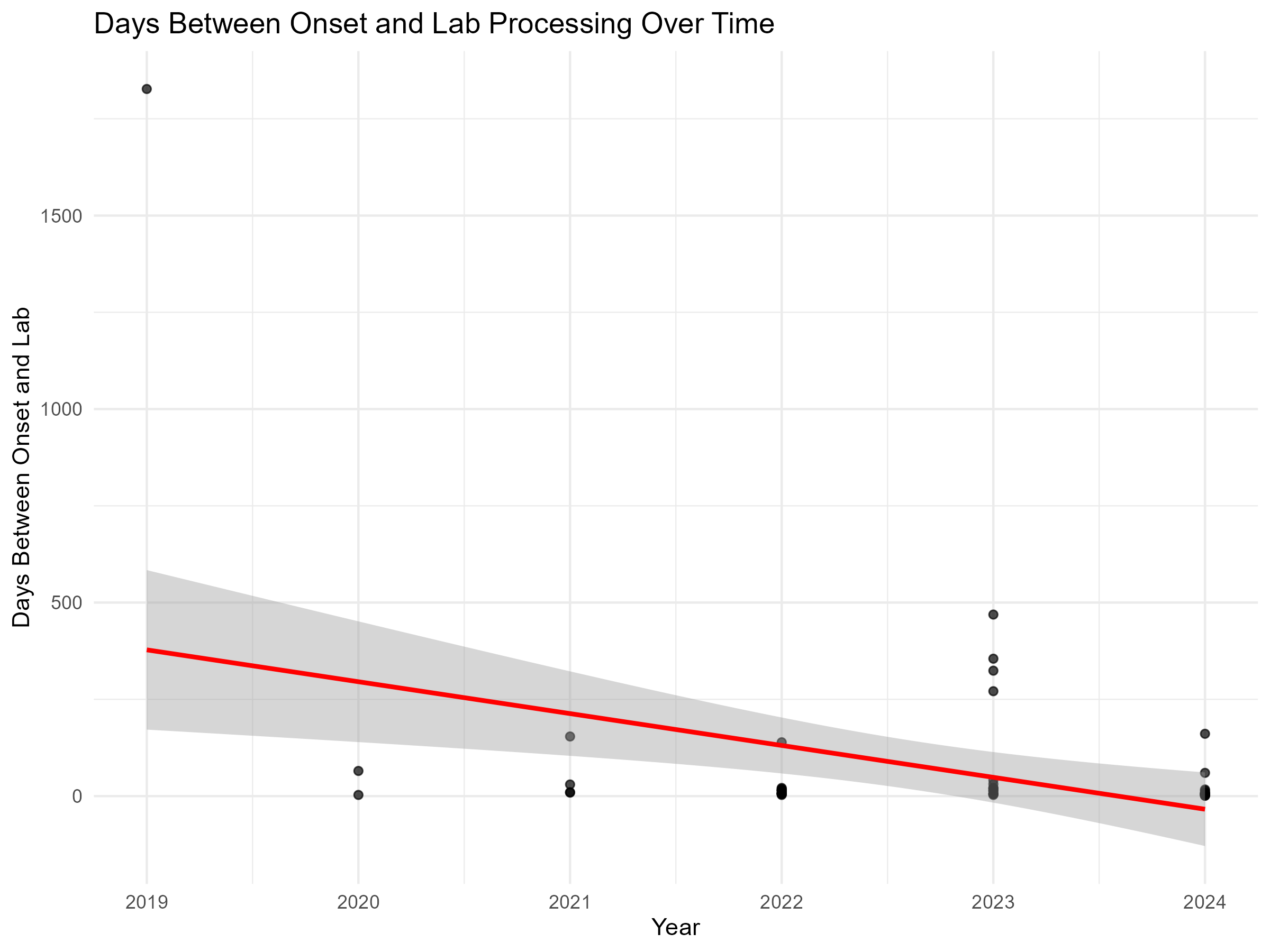
The most interesting finding from the initial modeling was in relation to the data completeness research question. In modeling the differences across time in whether a final STATUS was inputted rather than left blank, a general decrease was seen across the full 2018-2024 time scale. Specifically, completion rates of the reporting form were approximately 50% in 2018 but dropped rapidly to 20% in 2019– around which the rate of completeness has hovered since 2019. However, in linear modeling, the changes across time in the aggregate, p > 0.05, meaning observed differences from year to year were not statistically significant.



| term  < chr> | e s t imate  < dbl> | s t d . error  < dbl> | s t a t istic  < dbl> | p . value  < dbl> |
| --- | --- | --- | --- | --- |
| ( I n t e r cept) | 9 9 3 5 . 2 67857 | 5 7 2 9 . 9 46482 | 1 . 7 33920 | 0 . 1 4 34648 |
| YEAR | * 4 . 9 10714 | 2 . 8 35202 | * 1 . 7 32051 | 0 . 1 4 38108 |

2 rows

On a positive note for the utilization of the reporting system and trends towards better tick-borne disease response, there has been statistically significant decreases in the time between onset and an initial lab visit/interaction. In 2018, the average time between these two occurrences ranged from approximately 150-520 days. With marked decreases, the 2024 data indicates that patients with tick-borne disease occurrences may be interacting with providers or lab testing in the same month, same week or even same day.



## 6.2 5.2 Further Statistical Analysis

To assess whether delays in disease follow-up could be predicted based on demographic or case-specific factors, a binary classification variable (delay\_over\_threshold) was created. This variable indicates whether the total case duration—from symptom onset to final update—exceeded 30 days. The outcome was treated as a binary factor (1 = delay > 30 days, 0 = delay ≤ 30 days) and used as the target variable for modeling.

Predictors included:

* AGE: Patient’s age at time of case entry
* SEX: Reported gender
* COUNTY: County of residence
* DISEASE: Tick-borne disease type (e.g., Ehrlichiosis, Lyme disease, RMSF)

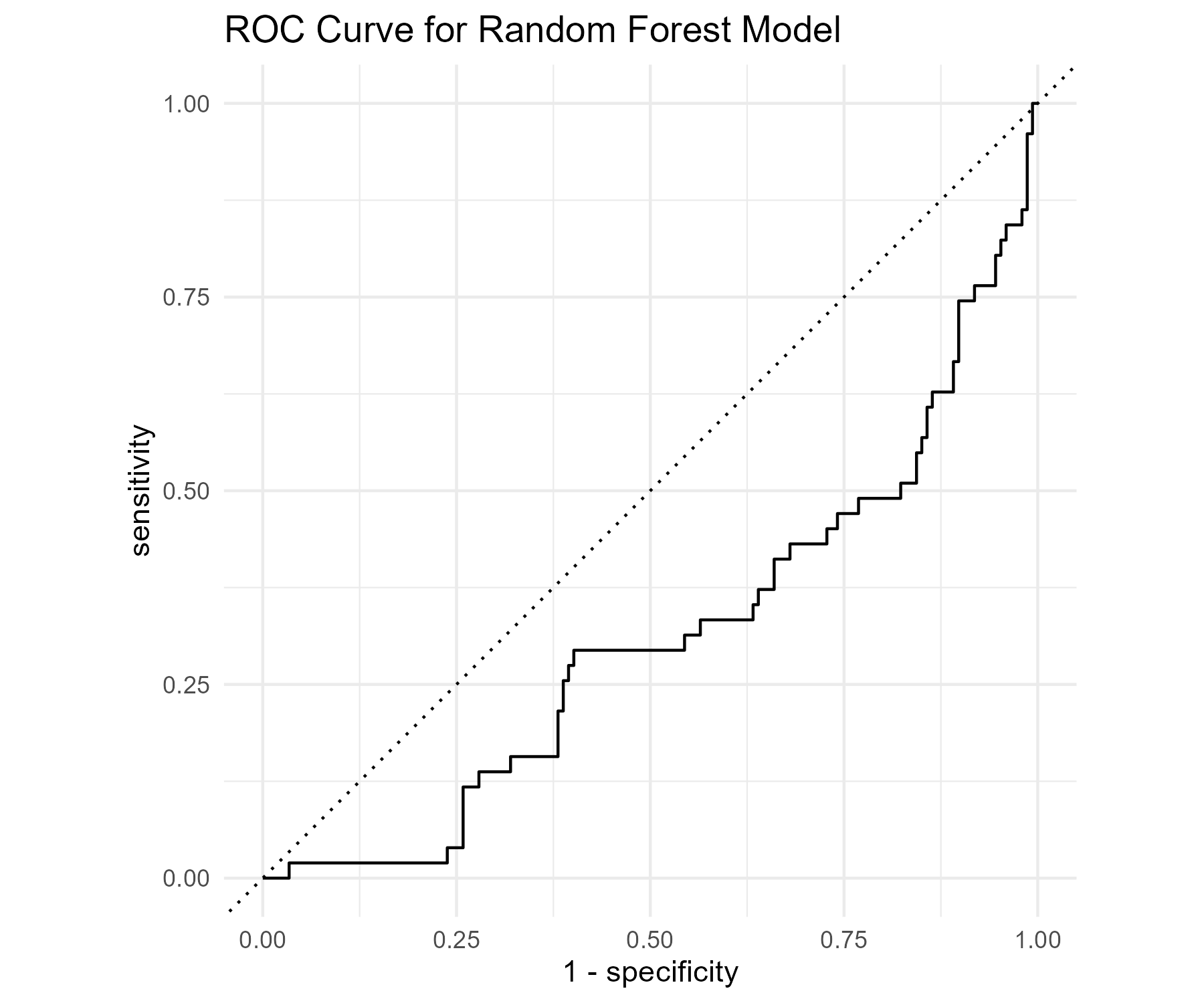
These predictors were selected based on their relevance to health access and geographic exposure patterns. All categorical variables were converted into dummy variables for modeling, and the age variable was normalized. Three classification models were trained and evaluated using the tidymodelframework:

1. Logistic Regression
2. Decision Tree
3. Random Forest

To evaluate model performance, 5-fold cross-validation was repeated 3 times on the training dataset. Model quality was assessed using both accuracy (overall correct classification rate) and area under the ROC curve (AUC), which measures the model’s ability to discriminate between delayed and non-delayed cases.

| Model | Accuracy | ROC AUC |
| --- | --- | --- |
| Logistic Regression | 72.8% | 0.65 |
| Decision Tree | 77.8% | 0.61 |
| Random Forest | 78.1% | 0.68 |

The logistic regression model provided an interpretable baseline with modest discriminatory power (AUC = 0.65). The decision tree achieved higher accuracy but lower AUC, indicating possible overfitting or limited generalizability. The random forest model outperformed the other two in both metrics, achieving 78.1% accuracy and an AUC of 0.68, and was selected as the best-performing model. Given these results, the random forest model was selected for final evaluation on the test data set.



## 6.3 5.3 Adding in Poverty Rate Data by County

Now, I was curious. Reflecting back on the EDA analysis and the stark differences in county and cases, if the modeling of delay would better predicted if we merged county-level poverty rate data. This poverty rate data is reasonable, potential structural determinant of diagnostic delay due to persistent healthcare inequities. This variable was included in model training alongside demographic and disease characteristics. Preprocessing steps were updated accordingly, and models were re-evaluated using the same cross-validation and test evaluation procedures to assess any added predictive value.

County-level poverty rate was evaluated as a predictor of whether a county had a high proportion of cases with diagnostic delays exceeding 30 days. A random forest classification model using only poverty rate as a predictor achieved 88.4% accuracy, but its ability to discriminate between delayed and non-delayed counties was poor (ROC AUC = 0.46). This suggests that although the model often guessed the majority class correctly, it lacked the ability to meaningfully differentiate high-risk counties.

Similarly, a linear regression model found no statistically significant association between poverty rate and delay rate (p = 0.87, β = 0.0049). The random forest regression model explained less than 4% of the variance in delay rates (R² = 0.038) with a relatively high prediction error (RMSE = 0.37).

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |

# 7. 6. Discussion

## 7.1 6.1 Summary and Interpretation

The random forest model for the modeling of delay as an outcome all predictors selected from cross-validation was refitted to the full training data and evaluated on the reserved 20% test dataset to assess generalizability. On this unseen data, the model achieved an accuracy of 74.7%, meaning nearly three-quarters of tick-borne disease cases were correctly classified as having experienced delayed (>30 day) or non-delayed follow-up. The model’s ability to discriminate between delayed and non-delayed cases was moderate, with an area under the receiver operating characteristic (ROC) curve of 0.688, closely mirroring performance observed during cross-validation.

The ROC curve showed consistent separation between sensitivity and 1-specificity, further supporting model validity. These results suggest that case characteristics such as age, sex, county, and disease type contain meaningful predictive information regarding follow-up delays in surveillance data.Such insights can inform improvements in case monitoring, triaging, and outreach in counties or patient groups with higher risk for delays.

County-level poverty rate was evaluated as a predictor of the proportion of cases with diagnostic delays exceeding 30 days. A linear regression model indicated no statistically significant relationship between poverty rate and delay rate (p = 0.87). The estimated effect size was minimal (β = 0.0049), and the 95% confidence interval included zero.

A random forest regression model using the same predictor resulted in low predictive performance. The R² was 0.038, indicating that poverty rate explained less than 4% of the variance in delay rate across counties. The RMSE was 0.37, reflecting high residual error in predicting delay proportions.

In comparison, the earlier classification models using individual-level predictors (e.g., age, sex, disease type, county) performed considerably better in both accuracy and AUC, suggesting that county-level poverty alone does not substantially contribute to explaining or predicting diagnostic delays in this dataset.

### 7.1.1 What this means:

* The poverty-only model had higher accuracy, but a lower ROC AUC.
* Accuracy tells us how often the model got it right — so poverty alone seems to classify “delay” correctly most of the time.
* However, the low ROC AUC suggests that the model struggles to rank cases by risk — in other words, it doesn’t distinguish well between high and low risk of delay.
* The all-predictors model performed more balanced, with a decent accuracy and a stronger ROC AUC, meaning it did a better job discriminating between delayed and non-delayed cases.

| Model Type | Predictors Used | Accuracy | ROC AUC |
| --- | --- | --- | --- |
| Random Forest | Age, Sex, County, Disease | 78.1% | 0.68 |
| Random Forest - Poverty Rate Only | Poverty Rate By County | 88.4% | 0.46 |

## 7.2 6.2 Strengths and Limitations

This analysis leveraged a comprehensive dataset spanning six years and multiple counties across Georgia, allowing for exploration of both individual- and community-level patterns in diagnostic delay for tick-borne disease cases. The geographic breadth of the data is a key strength, enabling comparisons across rural and urban areas and supporting more generalizable insights. Additionally, the use of several modeling strategies—including logistic regression, decision trees, and random forests—provided a well-rounded evaluation of predictor importance and model performance. The inclusion of both cross-validation and a holdout test set added rigor and reduced the risk of overfitting.

Another limitation is the use of county-level poverty rate as a proxy for individual socioeconomic status (SES). While poverty data at the county level offers a broad socioeconomic context, it can mask significant within-county income disparities. In counties with wide economic variation, aggregate poverty rates may not accurately represent the lived experience or access to healthcare of individuals affected by tick-borne disease.

Furthermore, individual exposure risk is often tightly linked to SES. For instance, people in outdoor labor roles—such as landscaping, farming, or construction—may face greater tick exposure than those working indoors, regardless of the county’s average income level. These more granular, individual-level occupational or behavioral factors were not captured in this dataset, which limits the precision of predictive modeling based solely on geography and aggregated socioeconomic indicators.

However, several limitations should be acknowledged. Inconsistencies in reporting across cases and counties required extensive data cleaning, which ultimately reduced the dataset to a much smaller number of usable observations. This constrained sample size may have limited model performance and generalizability. The dataset also lacked clinical and healthcare access variables—such as symptom severity, insurance status, or provider responsiveness—which likely influence diagnostic delays. Without these contextual details, the models may have missed key drivers of delay. Finally, the poverty rate was measured at the county level, which may not reflect intra-county disparities or the lived experience of patients.

# 8. 7. Conclusions

* Individual-level characteristics such as age, sex, disease type, and county were moderately effective in predicting whether a tick-borne disease case experienced a diagnostic delay of over 30 days. The best-performing model (random forest) achieved 78.1% accuracy and an AUC of 0.68, indicating meaningful—though not perfect—predictive value.
* County-level poverty rate alone was not a strong predictor of diagnostic delay. Both linear regression and random forest regression showed weak associations and low explanatory power, suggesting that broader socioeconomic indicators may not capture the nuanced drivers of individual-level follow-up delays.
* Data quality and completeness were major challenges: extensive cleaning was required, and missing or inconsistent entries significantly reduced the number of usable cases. This limitation underscores the need for more standardized reporting practices in public health surveillance systems.

# 9. Supplemental Figures

A full set of exploratory and supplemental visualizations—used to guide data cleaning, quality checks, and hypothesis generation—are available in a separate folder.

Please refer to the following file for all supplemental figures:

File: supplemental-figures.zip

This archive includes: - Disease distribution histograms - County-level maps - Test result breakdowns - Titer plots - Age distribution and data completeness trends that I did in the exploratory data analysis but opted to not include in my analysis per my reviewers’ recommendations that I narrow my scope!

# 10. References