ECOL 8910

Lyme Cases Redefinition Project

James, Josiah, Kane, and Nicholas

2024-11-13

Background:

To make the most of epidemiological data, consistency in the pattern of data collection is needed. This is important for a good understanding of any disease and to inform intervention. Lyme disease has increased in incidence, becoming one of the most common and widely distributed pathogens in the United States (Mead et al., 2024). In the United States, it is estimated that there are 476,000 Lyme disease cases per year. While mortality associated with Lyme disease is rare, its impacts on public health make it important to understand how this disease spreads using existing data on incidence rates.

Since Lyme was first described by Steere et al. (1977), public health organizations like the CDC have developed and revised their surveillance conditions based on clinical and laboratory criteria. Case definitions or surveillance case definitions are established at the state and local levels. This ultimately ensures that cases are being counted consistently. Currently, the CDC has done case definitions six times—in the years 1995, 1996, 2008, 2011, 2017, and 2022, with the most radical change being in the year 2022. Under the latest revision, reporting criteria for low-incidence areas changed minimally, requiring the collection of both clinical and laboratory data to identify and classify cases. In high-incidence areas, however, cases are reportable as probable cases based on positive laboratory results alone, absent any clinical information, begging the question of how much influence the lack of complementary clinical information has on incidence rate reports, since high-incidence jurisdictions report cases based on laboratory evidence alone.

Although these changes improve the standardization of surveillance across jurisdictions, they preclude detailed comparisons with historical data, therefore making effective deployment difficult. We observed that after the implementation of a revised surveillance case definition in 2022, the number of reported Lyme disease cases in the United States increased 68.5% over the average reported during 2017–2019; in high-incidence jurisdictions, the number of cases increased 72.9%, whereas, in low-incidence jurisdictions, the number of cases increased 10.0% (Kugela et al., 2024). One outstanding question is whether these changes in reported cases are due to the case redefinition or a true reflection of the increasing disease burden. Therefore, reconciling these datasets to a scale that makes them comparable is critical for their effective usage in public health analysis.

Questions:

1. How does the pre-2022 case definition relate to the 2022 redefinition?

2. How could the absence of clinical information influence the reporting rate in high-incidence areas?

Note: We are using Lyme incidence data from the year 2008 to 2022 across 51 states. Sixteen of these states are categorized as belonging to the high-incidence jurisdiction and the remaining 35 as low-incidence jurisdiction.

```
library(deSolve)
library(ggplot2)
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ lubridate 1.9.3     ✔ tibble    3.2.1
## ✔ purrr     1.0.2     ✔ tidyr     1.3.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(stringr)
library(magrittr)

##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract

library(dplyr)
library(ggplot2)
library(lmtest)   # For statistical tests

## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

library(sandwich) # For robust standard errors
library(zoo)      # For time series manipulation
library(segmented) # For interrupted time series

## Loading required package: MASS
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
##
## Loading required package: nlme
##
## Attaching package: 'nlme'
##
## The following object is masked from 'package:dplyr':
##
##     collapse

Lyme_data <- read.csv ("Lyme_jurisdiction_data.csv")
head(Lyme_data); tail(Lyme_data)

##                states pre2022_cases year2022_cases percent_change
## 1         Connecticut         1,714          2,022             18
## 2            Delaware           590            298          ?49.5
## 3 District of Columbia           88             77          ?12.5
## 4               Maine         1,807          2,653           46.8
## 5            Maryland         1,563          2,035           30.2
## 6       Massachusetts           144          5,052       3,408.30
##   pre2022_incidence year2022_incidence incidence_change jurisdiction
## 1              47.5               56.1              8.5         high
## 2           59.6            30.1          ?29.5       high
## 3           12.7            11.2           ?1.5       high
## 4          132.7           194.7           62.1       high
## 5           25.3            32.9            7.6       high
## 6            2.1            71.9           69.8       high
```

```
##        states pre2022_cases year2022_cases percent_change pre2022_incidence
## 46 South Dakota          10             12             20               1.1
## 47    Tennessee          40             32          ?20.0               0.6
## 48        Texas          49             23          ?53.1               0.2
## 49         Utah          24             16          ?33.3               0.7
## 50   Washington          33             23          ?30.3               0.4
## 51      Wyoming           3              4           33.3               0.5
##    year2022_incidence incidence_change jurisdiction
## 46                1.4              0.3          low
## 47                0.5             ?0.1          low
## 48                0.1             ?0.1          low
## 49                0.5             ?0.2          low
## 50                0.3             ?0.1          low
## 51                0.7              0.2          low
```

```r
str(Lyme_data)
```

```
## 'data.frame':    51 obs. of  8 variables:
##  $ states           : chr  "Connecticut" "Delaware" "District of Columbia" "Maine" ...
##  $ pre2022_cases    : chr  "1,714" "590" "88" "1,807" ...
##  $ year2022_cases   : chr  "2,022" "298" "77" "2,653" ...
##  $ percent_change   : chr  "18" "?49.5" "?12.5" "46.8" ...
##  $ pre2022_incidence : chr  "47.5" "59.6" "12.7" "132.7" ...
##  $ year2022_incidence: chr  "56.1" "30.1" "11.2" "194.7" ...
##  $ incidence_change  : chr  "8.5" "?29.5" "?1.5" "62.1" ...
##  $ jurisdiction      : chr  "high" "high" "high" "high" ...
```

```r
summary(Lyme_data)
```

```
##     states          pre2022_cases      year2022_cases     percent_change    
##  Length:51          Length:51          Length:51          Length:51         
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##  pre2022_incidence  year2022_incidence incidence_change   jurisdiction      
##  Length:51          Length:51          Length:51          Length:51         
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
```

```r
#Imprt the second data
data2 <- read.csv ("Lyme_Incidence_by_States.csv")
head(data2)
```

```
##       states X2008 X2009 X2010 X2011 X2012 X2013 X2014 X2015 X2016 X2017 X2018
## 1    Alabama   0.2   0.1   0.0   0.5   0.5   0.5   1.3   0.5   0.8   0.8   0.7
## 2     Alaska   0.9   1.0   1.0   1.5   1.4   1.9   1.1   1.2   2.0   1.3   1.5
## 3    Arizona   0.1   0.1   0.0   0.2   0.2   0.5   0.3   0.2   0.2   0.4   0.1
## 4   Arkansas   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.1   0.2   0.1
## 5 California   0.2   0.3   0.3   0.2   0.2   0.3   0.2   0.3   0.3   0.4   0.3
## 6   Colorado   0.1   0.0   0.1   0.0   0.0   0.0   0.0   0.0   0.0   0.1   0.1
##   X2019 X2020 X2021 X2022
## 1   1.3   0.3   1.0   0.6
## 2   0.4   0.0   0.0   1.0
## 3   0.1   0.0   0.0   0.1
## 4   0.6   0.4   0.4   0.1
## 5   0.4   0.1   0.3   0.2
## 6   0.1   0.0   0.0   0.2

#I just want to join these two datasets, by states

Combined_data <- inner_join(Lyme_data, data2, by="states")
head(Combined_data)

##                 states pre2022_cases year2022_cases percent_change
## 1          Connecticut         1,714          2,022             18
## 2             Delaware           590            298          ?49.5
## 3 District of Columbia            88             77          ?12.5
## 4                Maine         1,807          2,653           46.8
## 5             Maryland         1,563          2,035           30.2
## 6        Massachusetts           144          5,052       3,408.30
##   pre2022_incidence year2022_incidence incidence_change jurisdiction X2008
## 1              47.5               56.1              8.5         high 111.2
## 2              59.6               30.1            ?29.5         high  88.1
## 3              12.7               11.2             ?1.5         high  12.5
## 4             132.7              194.7             62.1         high  68.8
## 5              25.3               32.9              7.6         high  39.2
## 6               2.1               71.9             69.8         high  70.0
##   X2009 X2010 X2011 X2012 X2013 X2014 X2015 X2016 X2017 X2018 X2019 X2020 X2021
## 1 118.1  85.7  84.7  73.9  81.3  65.6  70.8  48.8  57.4  52.0  34.6  17.3  15.0
## 2 111.2  72.9  96.2  73.1  55.1  44.7  46.2  53.3  63.5  53.8  65.6  36.0  14.0
## 3  10.2   6.9   0.0   0.0   5.4   6.0  17.9  15.0  12.1  11.2  14.1  16.8  13.6
## 4  73.6  56.6  75.7  83.7 103.3 105.2  90.4 111.6 138.5 104.8 161.0  83.5 110.0
## 5  35.5  27.9  23.1  28.0  20.2  23.0  28.9  31.1  31.4  22.9  23.4  13.9  14.9
## 6  79.7  49.7  37.4  77.1  78.8  78.4  62.1   2.9   6.0   0.2   0.1   1.6   0.4
##   X2022
## 1  55.8
## 2  29.7
## 3  11.5
## 4 192.6
## 5  33.0
## 6  72.3

#This is not even what I need. I just need a simple dataset with columns for all the years and another for all the jurisdictions.
fin_data <- Combined_data[, -2:-7] #excludes column 2 to 7
head(fin_data)

##                 states jurisdiction X2008 X2009 X2010 X2011 X2012 X2013 X2014
## 1          Connecticut         high 111.2 118.1  85.7  84.7  73.9  81.3  65.6
## 2             Delaware         high  88.1 111.2  72.9  96.2  73.1  55.1  44.7
## 3 District of Columbia         high  12.5  10.2   6.9   0.0   0.0   5.4   6.0
## 4                Maine         high  68.8  73.6  56.6  75.7  83.7 103.3 105.2
## 5             Maryland         high  39.2  35.5  27.9  23.1  28.0  20.2  23.0
## 6        Massachusetts         high  70.0  79.7  49.7  37.4  77.1  78.8  78.4
##   X2015 X2016 X2017 X2018 X2019 X2020 X2021 X2022
## 1  70.8  48.8  57.4  52.0  34.6  17.3  15.0  55.8
## 2  46.2  53.3  63.5  53.8  65.6  36.0  14.0  29.7
## 3  17.9  15.0  12.1  11.2  14.1  16.8  13.6  11.5
## 4  90.4 111.6 138.5 104.8 161.0  83.5 110.0 192.6
## 5  28.9  31.1  31.4  22.9  23.4  13.9  14.9  33.0
## 6  62.1   2.9   6.0   0.2   0.1   1.6   0.4  72.3
```

```
#view(fin_data)

#write.csv(fin_data, file="Merged_Lyme_data.csv")

# Pivot to get desired columns
Final_Lyme_data<-fin_data%>%
 pivot_longer(cols = 3:17, names_to = "year", values_to = "incidence_rate")
str(Final_Lyme_data)
```

```
## tibble [765 × 4] (S3: tbl_df/tbl/data.frame)
##  $ states        : chr [1:765] "Connecticut" "Connecticut" "Connecticut" "Connecticut" ...
##  $ jurisdiction  : chr [1:765] "high" "high" "high" "high" ...
##  $ year          : chr [1:765] "X2008" "X2009" "X2010" "X2011" ...
##  $ incidence_rate: num [1:765] 111.2 118.1 85.7 84.7 73.9 ...
```

```
#I need to remove x attached to the years
Final_Lyme_data <- Final_Lyme_data %>%
 mutate(year = str_remove(year, "^X")) #drop X attached to the years
```

Test

We tried the interrupted time series analysis, but our desire to find something more basic and relatable combined with the complexity of its interpretability made us drop it. We, therefore, resorted to a more generic approach.

Generic Approach

Our thought process: We tried to consider our data in this way: The Lyme incidence records pre-2022 in both high and low jurisdictions are probably short of the true incidence record. Let the records before 2022, using both lab and clinical tests be labelled P for low jurisdiction and P2 for high jurisdiction. Q will represent incidence rates from high jurisdiction for the year 2022 which do not require clinical examination (based on the 2022 case redefinition for high-incidence jurisdiction).

```
# Label data as P or Q based on the criteria
Final_Lyme_data <- Final_Lyme_data %>%
 mutate(Label = ifelse((jurisdiction == "low"), "P",
               ifelse((year < 2022 & jurisdiction == "high"), "P2", "Q")))
```

Now, we decided to approach this problem this way: We try to find patterns in incidence rate across years (maybe the rate of increase from one year to another), separately for low- and high-jurisdiction. This should be done for all years in the low-incidence jurisdiction and for all years, except, 2022 for the high-incidence jurisdiction.

We hope that it tells us something about the rate of increase in incidence rate from one year to the other. We can then use our knowledge of this rate to see how much our estimate for 2022 in the high-incidence jurisdiction differs from the actual reported value.

This approach focuses on identifying trends in year-over-year increases in incidence rates, and then comparing the 2022 high-incidence value to the expected rate based on previous years. We can follow these steps to implement it:

Calculate Yearly Rate of Increase: For each jurisdiction (low- and high-incidence), calculate the rate of increase in incidence rate between consecutive years, excluding 2022 for high-incidence jurisdictions.

```
# Step 1: Calculate Yearly Rate of Increase
high_pre2022_data <- Final_Lyme_data %>%
 filter(Label == "P2") %>%
 group_by(states) %>%
 arrange(year) %>%
 mutate(rate_increase = (incidence_rate - lag(incidence_rate)) / lag(incidence_rate)) %>%
 mutate(rate_increase = ifelse(is.nan(rate_increase), 0, rate_increase))%>%
 filter(!is.infinite(rate_increase) & !is.na(rate_increase))  # Remove rows with NA or Inf in rate_increase
```

```
low_data <- Final_Lyme_data %>%
 filter(Label == "P") %>%
 group_by(states) %>%
 arrange(year) %>%
 mutate(rate_increase = (incidence_rate - lag(incidence_rate)) / lag(incidence_rate)) %>%
  mutate(rate_increase = ifelse(is.nan(rate_increase), 0, rate_increase))%>%
 filter(!is.infinite(rate_increase) & !is.na(rate_increase))  # Remove rows with NA or Inf in rate_increase


high_2022_data <- Final_Lyme_data %>%
 filter(Label == "Q" & jurisdiction=="high") %>%
 group_by(states) %>%
 arrange(year) %>%
  mutate(rate_increase = (incidence_rate - lag(incidence_rate)) / lag(incidence_rate)) %>%
   mutate(rate_increase = ifelse(is.nan(rate_increase), 0, rate_increase))%>%
 filter(!is.infinite(rate_increase) & !is.na(rate_increase))  # Remove rows with NA or Inf in rate_increase

### Note: Our values are bounded between -1 and +inf; 0=nothing change; 1=double! But the incidence rate can triple, etc (value>1)
#Note that NAs could be true zero values. For example, if year1=o and year2=0, R calculates it as NaN, but in actual sense, we can say
there is no change = 0, therefore NaNs==0
#Again, For example: if year2=1 and year 1=0, r will be calculated as (year2-year1)/year1 = ((1-0)/0) = inf
```

```
# Step 2: Calculate the Average Rate of Increase for Each Jurisdiction
avg_increase_high <- high_pre2022_data %>% summarize(avg_rate_increase = mean(rate_increase, na.rm = TRUE))
%>% pull(avg_rate_increase)
avg_increase_low <- low_data %>% summarize(avg_rate_increase = mean(rate_increase, na.rm = TRUE)) %>%
pull(avg_rate_increase)
```

We will now have to predict the 2022 high-incidence rate based on average increase

Here we will use the incidence rate from 2021 in high-incidence jurisdictions, along with the calculated average yearly rate of increase to predict what the 2022 incidence rate would have been if clinical data were included. This approach provides a predicted 2022 incidence rate based on historical trends, allowing for comparison with the actual 2022 rate that relied solely on lab data. This comparison can reveal how the exclusion of clinical data in high-incidence jurisdictions may have impacted the reported rate in 2022.

```
# Step 1: Get the 2021 incidence rate for high-incidence jurisdiction
last_incidence_2021_high <- high_pre2022_data %>%
 filter(year == 2021) %>%
 pull(incidence_rate)

# Step 2: Predict the 2022 incidence rate
predicted_2022_high <- last_incidence_2021_high * (1 + avg_increase_high)

## Warning in last_incidence_2021_high * (1 + avg_increase_high): longer object
## length is not a multiple of shorter object length

# Output the predicted value
mean_predicted_2022_high <- mean(predicted_2022_high)
```

Let us compare with the Actual incidence rate reported in the data

We will compare the predicted value with the actual 2022 data to estimate the effect of the case definition change.

```
# Step 3: Retrieve the actual 2022 incidence rate
actual_2022_high <- Final_Lyme_data %>%
 filter(Label == "Q") %>%
 pull(incidence_rate)

mean_actual_2022_high <- mean(actual_2022_high)
# Step 4: Calculate the difference
difference_2022 <- mean_actual_2022_high - mean_predicted_2022_high

# Output the difference
difference_2022

## [1] 51.61204

# Print results
cat("The Mean Predicted 2022 Incidence Rate for High-Incidence Jurisdiction is:", mean_predicted_2022_high, "\n")

## The Mean Predicted 2022 Incidence Rate for High-Incidence Jurisdiction is: 35.31921

cat("The Mean Actual 2022 Incidence Rate for High-Incidence Jurisdiction is:", mean_actual_2022_high, "\n")

## The Mean Actual 2022 Incidence Rate for High-Incidence Jurisdiction is: 86.93125

cat("The difference due to lack of clinical data in 2022 is:", difference_2022, "\n")

## The difference due to the lack of clinical data in 2022 is: 51.61204
```

Can we come up with a single parameter k by which we can tell how impact the lack of clinical information has on incidence rate in 2022, given this result? To derive a single parameter (let's say k) that quantifies the impact of the lack of clinical information on the incidence rate in 2022, we can express k as the relative difference between the actual and predicted 2022 incidence rates, aggregated across all states in the high-incidence jurisdiction.

We can calculate k as:

$$k = \frac{\text{mean of actual 2022 incidence rate} - \text{mean of predicted 2022 incidence rate}}{\text{mean of predicted 2022 incidence rate}}$$

```
# Calculate k
k <- (mean_actual_2022_high - mean_predicted_2022_high) / mean_predicted_2022_high

# Output k
k

## [1] 1.461302
```

Result:

Since k>0: Indicates that the actual incidence rates were higher than predicted, suggesting underestimation due to the lack of clinical information. k = 1.461302 is equivalent to a 146% overestimation of the incidence rate in the high-incidence jurisdiction in the year 2022.

Next Step - if time permits:

Can we try to reconcile the Lyme incodence data sets and bring all the years to the same comparable scale? We will try to rescale the whole data!