Linda Deckerli

**Lyme Disease Data Warehouse Design**

Contents

[Introduction 2](#_Toc111555979)

[Key Questions 2](#_Toc111555980)

[Constellation Schema 2](#_Toc111555981)

[Measures: 3](#_Toc111555982)

[ETL PROCESSES 5](#_Toc111555983)

[Data Extraction: 5](#_Toc111555984)

[Data Transformation and Loading 5](#_Toc111555985)

[Data Visualization 5](#_Toc111555986)

[Summary 7](#_Toc111555987)

[References: 7](#_Toc111555988)

# Introduction

According to the CDC (2022), Lyme disease is the most common vector-borne disease in America. Humans become infected with Lyme disease through the bite of infected ticks. Typical symptoms include fever, headache, joint pain, skin rash and fatigue. Lyme disease can also cause severe damage to the body if left untreated, including potential spread to the heart or nervous system (CDC, n.d.). It is therefore a significant public health issue, and this project thus utilizes data analysis to provide a more specific understanding of Lyme disease.

In particular, the goal of this project is to design a data warehouse that can be utilized by researchers to better understand the geographical and environmental factors of Lyme disease in the US. Multiple datasets were utilized in this effort, mainly from CDC and John Hopkins websites. The main steps for this project included designing dimensional schemas for a data warehouse, applying ETL processes to create staging tables to store the raw data from different sources into each table, cleaning, transforming and loading the data into a data warehouse based on the constellation schema design, and utilizing Microsoft PowerBI to create visualization of the data and address the below key questions.

# Key Questions

As detailed above, the goal of this project is to create a data warehouse to aid in better understanding the geographical and environmental factors of Lyme disease. The use of a data warehouse could aid researchers in answering certain key questions in this regard:

* In what parts of the US (states, counties, divisions, and regions) is Lyme disease most prevalent?
* What is the trend of the incidence of Lyme disease in comparison to its historical data in each county, state, or region over the past ten years?
* Where are blacklegged ticks (Lyme disease is transmitted by blacklegged (Ixodes) ticks) mostly located in the US?
* Are blacklegged ticks spreading over the years?
* Does Lyme disease have any correlation with environmental factors such as temperature and precipitation?

# Constellation Schema

Based on the above key questions, the constellation schema ERD design is below:

Diagram, timeline

Description automatically generated

There are three fact tables: Incidence, IxodesTransmission, and EnviromentalFactors. The dimension tables are State, County, Year, Season, and Month. The Incidence and IxodesTransmission fact tables and their related dimension tables were utilized to address the geographical questions about Lyme disease: understanding the geographical distribution and expansion of Lyme disease and the ticks that mainly cause Lyme disease in the US. Additionally, the EnvironmentalFactors table was utilized to record the average temperatures and total precipitations of each county and state in the US for each month and season from 2010-2020 so as to analyze whether such environmental factors contribute to the transmission of Lyme disease.

## Measures:

|  |  |
| --- | --- |
| Incidence | |
| Measure | Stored information |
| NumberOfCases | Stores the number of Lyme disease cases in each county from 2010-2020 in the US. |
| IncidenceRate | Stores the number of confirmed cases per 100,000 persons in each county from 2010-2020. |
| IncidenceCategory | stores two categories (high incidence and low incidence).  \*High Incidence=average incidence of at least 10 confirmed cases per 100,000 persons for three reporting years  \*Low Incidence=average incidence has not reached 10 confirmed cases per 100,000 persons for three reporting years. (CDC, 2022) |

|  |  |
| --- | --- |
| IxodesTransmission | |
| Measure | Stored information |
| Ixodes\_scapularis\_County\_Status | Stores the ‘spread status’ of Ixodes scapularis in each county. The status will be recorded as ‘Established’, ‘Reported’, and ‘No Records’. |
| Ixodes\_pacificus\_County\_Status | Stores the ‘spread status’ of *Ixodes pacificus* in each county. |
| Note:  **Established**: “Six or more *I. scapularis* or *I. pacificus* of a single life stage, or more than one life stage of the tick collected in the county within a 12-month period.” (CDC, April)  **Reported**: “Less than six I. scapularis or I. pacificus of a single life stage collected in the county within  a 12-month period.” (CDC, April)  **No Records**: “Counties classified as ‘no records’ should not be interpreted as ticks being absent. No records could arise either from a lack of: sampling efforts, tick collection, reporting or publishing the results of sampling efforts.” (CDC, April) | |

|  |  |
| --- | --- |
| EnviromentalFactors | |
| Measure | Stored information |
| AvgTemperature | Stores the average temperature of each county and state in the US from 2010-2020. |
| TotalPrecipitation | Stores the monthly total precipitation of each county and state in the US from 2010-2020. |

## ETL PROCESSES

## Data Extraction:

In order to apply ETL processes to project data, flat csv files from multiple sources were first extracted. Then these were loaded onto the SQL server and a staging table for each file was created.

The main files include:

* 'LD Case Counts by County’: this dataset provides the number of confirmed cases by US county from 2010-2019.
* ‘Incidence Rate’: this file provides the number of confirmed cases per 100,000 persons in the US.
* ‘Lyme disease incidence\_ categories’: provides the incidence category (high incidence or low incidence) of each county in the US).
* ‘Public\_Use\_Ixodes\_County\_Table (2018-2021)’: records the status (established, reported, no records) of two types of ticks (*Ixodes scapularis* and *Ixodes paci*ficus) in each county in the US from 2018 to 2021).
* ‘counties\_pcpn’: provides total precipitation for each month and season for each county in the US from 2010-2020).
* ‘temperature’: provides average temperature for each month and season for the counties in the US from 2010-2020.

## Data Transformation and Loading

After creating these staging tables, data transformation was conducted using SQL to create the fact and dimension tables based on the constellation schema. This included the use of large datasets (some with over 400,000 rows). SQL functions such as join, union, select, update, where, and group by etc. were utilized to clean, reorganize and integrate the data from different staging tables to predesigned fact and dimension tables. Additionally, consideration was made of performance tuning when applying data transformation (such as for creating indexes on some columns). Finally, all the data was successfully loaded to the data waremart/house. More details can be found in the attached code.

# Data Visualization

Clean data was then loaded to Microsoft Power BI to help provide a better visualized understanding of the data and therefore to better address the key questions.

To take Franklin County in Maine as an example of the use of Microsoft Power BI, the following screenshot shows visualization of the number of cases and the incidence rates for this county from 2010-2019. The histograms on the right- hand side clearly indicate that the number of cases and the incidence rate almost doubled from 2010 to 2019. Additionally, from the incidence category map, we can see the state of Maine is considered as a high-incidence state.

Graphical user interface, application

Description automatically generated

Likewise, the following screenshot shows visualization of the national ‘spread statuses’ of the two types of ticks (*Ixodes pacificus* and *Ixodes scapularis*) which cause Lyme disease. From the maps, we can clearly see that *Ixodes pacificus* ticks are primarily located on the West Coast, and *Ixodes scapularis* ticks are primarily located in the Northeast. Additionally, from the histograms on the left side of the below graph, we can see that *Ixodes pacificus* ticks have not spread (the number of counties that had ‘established’ reports of ticks remained 94 from 2018 to 2021). However, the number of counties that had ‘established’ reports of *Ixodes scapularis* ticks increased almost 10% from 2018 to 2021.

Chart

Description automatically generated

Finally, the following screenshot demonstrates the manner in which the project enables searches for average monthly and seasonal temperature and precipitation from 2010-2020 for each county in the US. Future work on the project envisions enabling the use of such functionality to conduct comparison of average temperature and precipitation for different states/divisions with tick-related criteria such as habitat to provide quantitative and qualitative assessment of the links between environmental factors and the incidence of Lyme disease in humans.

Graphical user interface, chart, application, line chart

Description automatically generated

# Summary

This project involved construction of a data warehouse to aid in analysis of the geographical and environmental factors involved in the spread of Lyme disease. Star schemas were designed based on the predefined key questions, ETL was applied to extract and combine data from multiple data sources and to transform this data into a structured and consistent datastore, and data was then loaded into the data warehouse. This approach allowed for Microsoft Power BI utilization so that visualizations can be created to present research results. The result is a database that allows researchers and the public to dynamically analyze trends in the spread of Lyme disease, an important and dangerous public health threat.

# References:

CDC. (2022 January). [Lyme Disease Data Tables: Historical Data | Lyme Disease | CDC](https://www.cdc.gov/lyme/stats/tables.html)

CDC. (2022 July). [Tick surveillance data sets | Ticks | CDC](https://www.cdc.gov/ticks/surveillance/TickSurveillanceData.html)

CDC (2022, April). [Blacklegged tick (Ixodes scapularis) surveillance | Ticks | CDC](https://www.cdc.gov/ticks/surveillance/BlackleggedTick.html)

CDC. (n.d.). [LymeDisease\_9211\_county | Data | Centers for Disease Control and Prevention (cdc.gov)](https://data.cdc.gov/dataset/LymeDisease_9211_county/smai-7mz9)

John Hopkins Bloomberg School of Public Health. (n.d.). [Data Explorer | Johns Hopkins Lyme and Tickborne Disease Dashboard (hopkinslymetracker.org)](https://www.hopkinslymetracker.org/explorer/)

CDC. (n.d.) [TABLE 2j. Annual reported cases of notifiable diseases, by region and reporting area, United States and U.S. Territories, excluding Non-U.S. Residents\*, 2019<sup class= 'nndss-sup-title'>†</sup> (cdc.gov)](https://wonder.cdc.gov/nndss/static/2019/annual/2019-table2j.html)

CDC. (n.d.). [Lyme Disease Data Tables: Historical Data | Lyme Disease | CDC](https://www.cdc.gov/lyme/stats/tables.html)

CDC. (n.d.).[Lyme Disease Data Tables: Most Recent Year | Lyme Disease | CDC](https://www.cdc.gov/lyme/datasurveillance/tables-recent.html)

CDC. (n.d.). [Lyme Disease Maps: Most Recent Year | Lyme Disease | CDC](https://www.cdc.gov/lyme/datasurveillance/maps-recent.html)

World Population Review. (n.d.) [List of State Abbreviations (Download CSV, JSON) (worldpopulationreview.com)](https://worldpopulationreview.com/states/state-abbreviations)