Exploring clinical and non-clinical(food) Listeria monocytogenes isolates

Alexis Gonzalez

The structure below is one possible setup for a manuscript, or a general data analysis project (including the course project). Adjust as needed. You don’t need to have exactly these sections, but the content covering those sections should be addressed.

This uses MS Word as output format. [See here](https://quarto.org/docs/output-formats/ms-word.html) for more information. You can switch to other formats, like html or pdf. See [the Quarto documentation](https://quarto.org/) for other formats.

**Authors**

* Alexis Gonzalez (ORCID: 0000-0000-1234-5678)
* Second Author

**Author affiliations**

1. College of Public Health, University of Georgia, Athens, GA, USA.
2. Another Affiliation, Athens, GA, USA.
3. Yet another Affiliation.

These authors contributed equally to this work.

Corresponding author: some@email.com

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

# 1. Part 1 Project Idea

For this project I will utilize data from the National Center for Biotechnology Information’s Pathogen Detection system. This is a database that collects genomic data from a various different sources such as state laboratories, private laboratories, and government agencies like CDC, FDA, USDA, and EPA. It is primarily used a surveillance tool to monitor isolates,clusters, antimicrobial resistance, and active outbreaks. As someone that keeps track of foodborne pathogen outbreaks, I have noticed the increase in Listeria monocytogenes outbreaks. With the NCBI data I would like to visualize the occurence of new listeria strains, compare trees for clinical vs nonclincal isolates, and maybe see the correlation between the number of clinical isolates to the number of non-clincal isolates, maybe investigate how often non-clinical isolates have a clinical in the same cluster. I could also look at distribution across the United States. There is some data for other countries, but far less than the U.S. . For this project I will filter the data to show isolates collected from January 1, 2015 to January 1, 2025. This gives me 22,980 isolates and 2,365 clusters.

# 2. Summary/Abstract

*Write a summary of your project.*

# 3. Introduction

## 3.1 General Background Information

*Provide enough background on your topic that others can understand the why and how of your analysis*

## 3.2 Description of data and data source

This project will utilize two data sources

1. National Center for Biotechnology Information’s Pathogen Detection system
2. CDC National Notifiable Disease Surveillance System

## 3.3 Questions/Hypotheses to be addressed

1. How has the presence of listeria in food products changes over time (NCBI data)
2. How has the rate of listeriosis cases changed over time
3. How does the positive rate of food cases compare to the rate of listeriosis cases
4. Comparing and contrasting the isolates found in food to the isolates found in clinical samples.
5. What are some commonalities between establishments and locations where listeria is found and where cases are diagnosed
6. Follow up to 4, are these large scale companies or small scale companies.
7. Maybe a prediction model based to predict listeriosis cases based on number of positive isolates found?

# 4. Methods

*Describe your methods. That should describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement.*

## 4.1 Schematic of workflow

Sometimes you might want to show a schematic diagram/figure that was not created with code (if you can do it with code, do it). [Figure 1](#fig-schematic) is an example of some - completely random/unrelated - schematic that was generated with Biorender. We store those figures in the assets folder.

|  |
| --- |
| Figure 1: A figure that is manually generated and shows some overview/schematic. This has nothing to do with the data, it’s just a random one from one of our projects I found and placed here. |

## 4.2 Data import and cleaning

Before important the NCBI data, I filtered the data to what I wanted using the sites filters. I filtered data down to organism = Listeria monocytogenes, location = USA , this left me with 33,264 isolates.

I will briefly use the NNDSS data because unfortunately, it is only availble from 2016 to 2022. The NCBI data containes clinical isolates that date farther back to we will consider those isolates to be “cases”. Those cases are what we will compare to the source isolates (environmental, food, etc)

The data cleaning code can be found in the processingfile under data->rawdata. The cleaning processed involved removing variables that I would not be using such as biosample numbers, assembly numbers, and variables that contained no data. Next I made some changes to date entries. They originally contained the time the isolate was uploaded so I removed that. I also created 2 new variables showing if the isolate was from a food sample or a clinical sample . The variable “Clinical” has a yes or no entry for if the sample is clinical, the same goes for the Food variable.

## 4.3 Statistical analysis

*Explain anything related to your statistical analyses.*

# 5. Results

## 5.1 Exploratory/Descriptive analysis

The full code for the EDA is under code->eda-code->eda.qmd

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

[Table 1](#tbl-summarytable) shows a summary of the data.

Note the loading of the data providing a **relative** path using the ../../ notation. (Two dots means a folder up). You never want to specify an **absolute** path like C:\ahandel\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path. You can also use the here R package to create paths. See examples of that below. I generally recommend the here package.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Data summary table.   | skim\_type | skim\_variable | n\_missing | complete\_rate | Date.min | Date.max | Date.median | Date.n\_unique | character.min | character.max | character.empty | character.n\_unique | character.whitespace | numeric.mean | numeric.sd | numeric.p0 | numeric.p25 | numeric.p50 | numeric.p75 | numeric.p100 | numeric.hist | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Date | Create date | 0 | 1.0000000 | 2011-05-10 | 2024-12-31 | 2019-05-14 | 2502 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | | character | #Organism group | 0 | 1.0000000 | NA | NA | NA | NA | 22 | 22 | 0 | 1 | 0 | NA | NA | NA | NA | NA | NA | NA | NA | | character | Strain | 1092 | 0.9667883 | NA | NA | NA | NA | 2 | 28 | 0 | 31714 | 0 | NA | NA | NA | NA | NA | NA | NA | NA | | character | Isolate identifiers | 0 | 1.0000000 | NA | NA | NA | NA | 5 | 110 | 0 | 32854 | 0 | NA | NA | NA | NA | NA | NA | NA | NA | | character | Isolate | 0 | 1.0000000 | NA | NA | NA | NA | 14 | 15 | 0 | 32880 | 0 | NA | NA | NA | NA | NA | NA | NA | NA | | character | Location | 0 | 1.0000000 | NA | NA | NA | NA | 3 | 24 | 0 | 110 | 0 | NA | NA | NA | NA | NA | NA | NA | NA | | character | Isolation source | 7041 | 0.7858577 | NA | NA | NA | NA | 3 | 157 | 0 | 2743 | 0 | NA | NA | NA | NA | NA | NA | NA | NA | | character | Isolation type | 787 | 0.9760645 | NA | NA | NA | NA | 5 | 58 | 0 | 55 | 0 | NA | NA | NA | NA | NA | NA | NA | NA | | character | Food origin | 15861 | 0.5176095 | NA | NA | NA | NA | 3 | 26 | 0 | 110 | 0 | NA | NA | NA | NA | NA | NA | NA | NA | | character | SNP cluster | 5603 | 0.8295925 | NA | NA | NA | NA | 6 | 19 | 0 | 3098 | 0 | NA | NA | NA | NA | NA | NA | NA | NA | | character | AMR genotypes | 30 | 0.9990876 | NA | NA | NA | NA | 12 | 216 | 0 | 252 | 0 | NA | NA | NA | NA | NA | NA | NA | NA | | character | clinical | 787 | 0.9760645 | NA | NA | NA | NA | 2 | 3 | 0 | 2 | 0 | NA | NA | NA | NA | NA | NA | NA | NA | | character | food | 0 | 1.0000000 | NA | NA | NA | NA | 2 | 3 | 0 | 2 | 0 | NA | NA | NA | NA | NA | NA | NA | NA | | numeric | Min-same | 6725 | 0.7954684 | NA | NA | NA | NA | NA | NA | NA | NA | NA | 5.272567 | 7.858745 | 0 | 0 | 2 | 7 | 57 | ▇▁▁▁▁ | | numeric | Min-diff | 15215 | 0.5372567 | NA | NA | NA | NA | NA | NA | NA | NA | NA | 16.506255 | 12.026130 | 0 | 6 | 16 | 26 | 67 | ▇▆▃▁▁ | |

## 5.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

[Figure 2](#fig-result) shows a scatterplot figure produced by one of the R scripts.

|  |
| --- |
| Figure 2: Height and weight stratified by gender. |

## 5.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example [Table 2](#tbl-resulttable2) shows a summary of a linear model fit.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Linear model fit table.   | term | estimate | std.error | statistic | p.value | | --- | --- | --- | --- | --- | | (Intercept) | 149.2726967 | 23.3823360 | 6.3839942 | 0.0013962 | | Weight | 0.2623972 | 0.3512436 | 0.7470519 | 0.4886517 | | GenderM | -2.1244913 | 15.5488953 | -0.1366329 | 0.8966520 | | GenderO | -4.7644739 | 19.0114155 | -0.2506112 | 0.8120871 | |

# 6. Discussion

## 6.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 6.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 6.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (1) discusses types of analyses.

These papers (2,3) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like.

# 7. References

1. Leek JT, Peng RD. [Statistics. What is the question?](https://doi.org/10.1126/science.aaa6146) *Science (New York, N.Y.)*. 2015;347(6228):1314–1315.

2. McKay B, Ebell M, Billings WZ, et al. [Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery.](https://doi.org/10.1093/ofid/ofaa494) *Open forum infectious diseases*. 2020;7(11):ofaa494.

3. McKay B, Ebell M, Dale AP, et al. [Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients.](https://doi.org/10.1098/rspb.2020.0496) *Proceedings. Biological sciences*. 2020;287(1927):20200496.