Statistical Machine Learning for Modeling Foodborne Disease Source Attribution

Amos Okutse Zexuan Yu Rophence Ojiambo

19 October, 2022

Abstract

Here is a line
Here is another line which can be longer.

Introduction

Some paper used the data set [1].

Related Literature

Methodology

Results

Exploratory data analysis

Variables Description

The listeria dataset as of last download on 10/18/2022 from the NCBI dataset had 53,725 observations with a total of 50 variables. Below is information on the variables of interest in our study as described in **Table 1.**

Organism group

This variable denotes the name of the taxonomy group that the isolate belongs to and is represented by the Genus species name and for our case we shall focus on Listeria monocytogenes.

Isolate

The isolate variable gives the unique Pathogen Detection accession of the isolate where each accession has a prefix "PDT," which stands for Pathogen Detection Target.

IFSAC category

This variable defines the categories of isolate sourcing information as developed by The Interagency Food Safety Analytics Collaboration (IFSAC).

Isolation Source and Isolation Type

Isolation source provides information on the physical, environmental and/or local geographical source of the biological sample from which the sampled was derived. Examples include cheese, blood,environmental swab, stool, tuna salad and smoked salmon. On the other hand, Isolation type categorizes the isolation sources into either clinical or environmental/other broad groups. These are the key variables in our study.

Strain

The strain variable gives the microbial strain name. It's used to distinguish a genetically distinct lineage separated from another strain by one or two mutations. Different strings of strains indicate different genetic variants or subtypes of microorganisms.

Host and Host Disease

Host refers to the host species of the isolate such as Animal, Homo sapiens, Sheep, Pigeon, Horse and Guinea pig whereas Host disease matches the identified isolate to a disease origin, for example Listeriosis, gastroenteritis, Meningitis and Septicaemia.

Collection Date and Create Date

Collection Date gives us the date the sample was collected which may differ from the Create date that gives the date on which the isolates were first seen by the Pathogen Detection system. Note: Decide on one of the two

Outbreak

The outbreak variable is a way to group isolates that originated due to the same breakout. It is a submitter given name for the occurrence of more cases of a disease than expected among a specific group of people or within a specific area over a period of time.

BioSample

BioSample variable describes the biological source materials used in experimental assays.

Lat/Lon and Location

Lat/Lon provides the geographical coordinates (latitude and longitude) of the location where the sample was collected while location provides the geographical origin of the sample (Country or Region)

Min-Same and Min Diff

Min-same variable represents the minimum single nucleotide polymorphism (SNP) distance to another isolate of the same isolation type for example, the minimum SNP distance from one clinical isolate to another clinical isolate, while the Min-diff is the minimum SNP distance to another isolate of a different isolation type. For example, the minimum SNP difference from a clinical isolate to an environmental isolate.

Missing Data

We shall start by assessing the missingness in our dataset. **Figure 1** shows us the overall missingness of our selected variables ordered from the least to the largest missing percentage. Key to note here is that there is over 86% missing observations in the variables Host Disease, Lat/Long, and Outbreak, with Outbreak having the most percentage (99.54%) of missing values. This amount of missingness will be a major limitation of our study as these variables may not be informative in our analysis and may limit the interpretation and generalizability of our study findings.

Variable Distributions

Next, we shall explore our data to get a sense of the variable distributions. Our data has 53,725 and 53694 unique isolates and BioSamples respectively. There are 42,7948 unique strains represented in our dataset, a reflection of the variation and heterogeneity of our data. Additionally, there are 3084 unique isolation sources which comprise of 17,344 clinical types and 30,356 environmental/other types. The data has 285 unique isolate sourcing categories as developed by IFSAC category scheme, 90 unique Hosts, and 67 unique Host Disease. To make meaningful comparisons in relation to our objective, we shall aggregate the IFASC category into 7 broad categories. Next, we examine the collection date and create date variables to see the time frame of our dataset. Since the collection date variable contains the date the sample was collected in the format the submitter supplied ranging from Month-Date-Year, Year-Month and Year only while the Create date is in the Year-Month-Date ISO format with time stamp the data was added into the Pathogen Detection Project, we shall reduce these variables to extract information on Year only to maintain consistency in terms of available year records.

As an initial exploration of the trends in number of Listeria monocytogenes over time, the line plot in Figure 2 shows a non-linear trend. There was a moderate increase in sample collected from the year 2010 to 2012, which was followed by a relatively sharp increase through the year 2016. From 2016 to 2020, there was variation in terms of steady increase and decrease that was later followed by a sustained decrease in the samples collected from the year 2020 through 2022.

Next, we aggregate the IFSAC isolate sourcing categories into 7 categories named; Dairy, Fruits, Leafy Greens, Meat, Poultry, Seafood and Vegetables, then further explore relationships with other variables in the dataset.

Statistical modeling

Discussion

Conclusion

References

[1] Sanaa M, Pouillot R, Vega FG, et al. GenomeGraphR: A user-friendly open-source web application for foodborne pathogen whole genome sequencing data integration, analysis, and visualization. *PloS one* 2019; 14: e0213039.