Drug Resistance and Susceptibility in Foodborne Illnesses

Data Exploration

Breanna Richards, Yanru Liao, Jina Yang

Link to GitHub Repository: https://github.com/brichards21/PHP2550-Final-Project

Data Exploration

The primary data source that we're using for this study is the National Center for Biotechnology Information's (NCBI) Pathogen database. This database includes information about sequenced bacterial pathogens that originate from a food source. It records isolates with source and genetic information. Recall that our study aims to assess if we can predict future isolates' performance against antibiotics. More specifically, can we learn from past isolates to develop a system to recommend antibiotics to new isolates? What kind of patterns can we extract from our current data about behaviors of drug susceptibility and resistance?

The focus of the study started with the Salmonella enterica species. This was partially due to the fact that Salmonella is one of the most common germs that causes foodborne illness in the United States. Due to the nature of the NCBI data, the column containing information about drug resistance and susceptibility is more often then not missing. Since our research question is contingent upon information about how these isolates interact with antibiotics, we only considered cases with non-missing information for this particular variable. Thus, the records that we had available to us to assess our study question were sparse. This led to the decision to extend our analysis outside of Salmonella to also include isolates from the E.coli and Shigella species as well as the Campylobacter jejuni species not only to increase our sample size but to also assess the similarities and differences between the three species.

Filtering records from all three of the species, our final dataset contained 8,672 records for Salmonella, 3,880 records for E.coli and Shigella, and 3,497 records for Campylobacter jejuni for a total of 16,049 records. We start by examining our variables of interest.

Variables

There are several variables in the NCBI dataset that we are considering in our study (National Center for Biotechnology Information).

Organism group

This variable refers to the taxonomy group that the isolate belongs to. For the purpose of our study this is going to take on one of three values: Salmonella enterica, E.coli Shigella, or Campylobacter jejuni.

Isolate

The isolate variable is unique for each record in our data and is how we identify individual organisms that have little genetic mixing.

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.

Collection Date

Collection date gives the date that the sample was collected.

Location

The geographical origin of the sample.

Source type

Source type gives the general category that an isolate originates from. Some examples of values of source type are 'Human', 'Animal feed', 'Environmental' and 'Food'.

Host and Host Disease

The 'Host' variable refers to the host species of the isolate. Examples include Homo sapiens, Bos taurus, pig, chicken, poultry, swine, and bovine. Host disease ties the isolate to a disease origin. Examples include gastroenteritis, salmonellosis, and diarrhea.

Serovar

The serovar gives the distinct variation within a species of bacteria. This column groups isolates based on their surface antigens, allowing the classification of isolates to the subspecies level.

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.

Isolation Source and Isolation Type

Isolation source describes the physical, environmental and/or local geographical source of the biological sample from which the sampled was derived. Examples include boneless beef, turkey, river, creek water, and water filtration membrane. Isolation type generalizes the isolation source into either clinical or environmental.

AST Phenotypes and AMR Genotypes

The AST Phenotype column gives the antibiotic resistant phenotype of the isolate. This variable is key to our study and lists antibiotics that the isolate has demonstrated resistance, susceptibility, intermediate responsiveness, and/or dose-dependent susceptibility to. AMR genotypes gives the antimicrobial resistant genes found in the isolate.

Computed Types

This formula gives the antigen formula and serotype results from an in-silico experiment. The antigen formula gives the detected presence of specific viral antigen which indicates viral infection. The serotype groups isolates by their distinctive surface structures/antigens.

SNP Cluster

A single nucleotide polymorphism (SNP) is a genomic variant at a single base position in DNA. It represents the difference in a single DNA building block. A SNP cluster is a group of isolates whose genome assemblies are closely related, depending on the clustering methodology used. This column gives each isolate's pathogen SNP cluster accession.

Min-Same and Min Diff

Min-same is the minimum SNP distance to another isolate of the same isolation type (clinical or environmental). Min-diff is the minimum SNP distance to another isolate of a different isolation type. For example, the minimum SNP difference from an environmental isolate to a clinical isolate.

Table 1: Isolates Info from Most to Least Missing

	% Missing	Count Missing
outbreak	1.0000000	16049
host_disease	0.9768210	15677
host	0.9343261	14995
serovar	0.4806530	7714
computed_types	0.4596548	7377
min_diff	0.3587139	5757
min_same	0.2050595	3291
snp_cluster	0.1925354	3090
strain	0.0406256	652
amr_genotypes	0.0284130	456
collection_date	0.0216836	348
isolation_source	0.0195651	314
location	0.0070409	113

We start by assessing the missingness in our data.

Table 1 gives the variables with missingness and orders them from most missing to least missing. As a rule of thumb, we may want to omit the variables with over 90% missingness as they may not be informative in our analysis. The variables that fit this criteria in our data are `outbreak`, `host disease`, and `host`. We recognize that our study will not be informed by information about an isolate's host disease, host species, and origin from an outbreak.

After removing the high-missingness variables, we note that 54.7% of the isolate records have at least one missing value for the 14 variables that we have left in our data.

We advance by assessing the distributions of our variables, starting with **strain**. There are 15,348 unique strains represented in our dataset, meaning that 95.6% of the isolates in our data are unique. We may want to consider dropping this variable as a result as it will not be too informative to our research problem.

Next, we looked at the **collection date** variable to examine the time periods that our data span. Since the raw values of this variable differ in terms of the specificity of the dates, i.e., some values only record year while others record the exact month and day as well, we'll reduce this variable to only the year for consistency purposes.

We observe the proportion of years represented within each organism group.

Interestingly, Salmonella is the only bacteria that retains isolate data for 2002-2008. Data starts to be retained for E.coli and Shigella in 2009 but shows the most records in 2017 and thereafter. Data collection for isolates for Campylobacter started in 2016 and retained records up to and including the year of 2020. As we advance in this project, we'll use this variable to understand trends in our data overtime, more specifically as it pertains to changes in drug resistance and susceptibility.

Next, let's look at the distribution of **serovar** within each bacteria of interest. Upon further inspection, we see that we only have non-missing serovar information for Salmonella. There are 150 distinct serovars in our Salmonella data. To reduce the number of groupings and simplify, values will only be distinguished as their own group if they represent over 2% of the isolates in the Salmonella data. Remaining serovar categories will be grouped into an 'Other' classification.

With this new sub-classified serovar column, we are left with 13 distinct groupings. Aside from the 'Other' Classification group, Kentucky (~14%), Heidelberg (8%), and Typhimurium var. 5- (7%) are the most common serovar groupings. These groupings may be broad enough to be of interest in our overall research project as we will test their association with drug response, although we note that we can only be able to implement this information in the context of Salmonella due to the constraints of our data.

Table 2: Prop. of Sources Represented Per Bacteria

Organism	beef	chicken	other/unspecified	pork	stool	turkey	urine	water
Campylobacter jejuni	17.27%	67.44%	4.75%	4.15%	4.69%	1.7%	0%	0%
E.coli and Shigella	16.49%	16.73%	5.07%	12.55%	12.52%	26.76%	7.35%	2.52%
Salmonella enterica	5.44%	46.92%	0.75%	9.87%	1.18%	35.75%	0.09%	0%

The **AST phenotypes** variable, in its rawest form, comes to us as one long string with all of the recorded antibiotics that each isolate is resistant, susceptible, intermediate in response, and susceptible-dose dependent to. We separated and classified this variable into two new variables, collapsing resistant and intermediate drug responses into 'resistant', and collapsing susceptible and susceptible-dose dependent responses into 'susceptible', and listed the corresponding antibiotics for each category. This process was done for ease of use when we eventually go on to build our model and want to separate antibiotics by isolates' responses to them.

In Salmonella, tetracycline (22%), streptomycin (16%), sulfisoxazole (12%), ampicillin (11%), and amoxicillin-clavulanic acid (7%) are the top antibiotics that our observed isolates are resistant to.

In E.coli, the top drugs that these isolates are resistant to are tetracycline (14%), ampicillin (11%), streptomycin (7%), sulfisoxazole (7%), and ceftriaxone (6%).

In Campylobacter, tetracycline (48%), ciprofloxacin (21%), nalidixic acid (20%), clindamycin (3%), and erythromycin (2%) are the top drugs that isolates show resistance to.

In general, these bacteria share some of antibiotics that they've grown most resistant to, but also exhibit a few differences.

Regarding the discussion on drug susceptibility, Salmonella isolates saw the most susceptibility to tetracycline (22%), streptomycin (16%), sulfisoxazole (12%), ampicillin (11%), and amoxicillin-clavulanic acid (7%).

In E.coli, isolates and strains were found to be most susceptible to meropenem (8%), ciprofloxacin (7%), nalidixic acid (7%), azithromycin (7%), and cefoxitin (7%).

Finally, in Campylobacter, the drugs those isolates were most susceptible to were gentamicin (13%), erythromycin (13%), florfenicol (12%), azithromycin (12%), and clindamycin (12%).

Interestingly, we see some overlap in certain bacteria being both susceptible and resistant to the same antibiotics. It'll be interesting to explore this more and see if we can examine potential shifts from susceptible to resistant status and vice versa for these illnesses.

Next, we look at the **locations** represented in our data. We'll reduce the location variable in our data to just be the countries. The vast majority of the records were collected in the United States with 99% of the Salmonella records hailing from the U.S., 85% of the E.coli and Shigella records from the U.S., and 95% of the Campylobacter jejuni records hailing from the U.S. The Netherlands and Germany are the next most popular countries represented. There are a total of 25 distinct countries represented in our dataset.

The **isolation source** variable originally hosted 180 unique values. Upon further inspection, we found that a lot of these distinctions were due to case sensitiveness as well as many variations of the same general source. For example, 'pork', 'pork chops', 'pork chop', and 'Pork'. To simplify, we grouped the isolates into broader categories based on the content we observed in the isolation source variable. We ultimately reduced this variable to 9 distinct categories: pork, chicken, turkey, beef, stool, urine, water, other/unspecified, and NA (for missingness).

The most common isolate source in our data is chicken for Campylobacter and Salmonella (**Table 2**). For E.coli and Shigella, it's turkey, with chicken being the second most common.

In terms of **isolation type**, there are 1,405 clinical isolates and 14,644 environmental/other isolates. This brings us to ~91% non-clinical (environmental/other) isolates. There's a difference in the distribution of isolation type in the E.coli bacterium group compared to the other two bacteria (**Table 3**). We're interested

Table 3: Prop. of Isolation Type Represented Per Bacteria

Organism	clinical	environmental/other
Campylobacter jejuni	4.83%	95.17%
E.coli and Shigella	24.85%	75.15%
Salmonella enterica	3.14%	96.86%

in assessing whether or not similarity in isolation source and isolation type is correlated with similarity and drug response and will explore this further.

There are 2,613 **SNP clusters** represented in our data, 811 clusters for Salmonella, 830 for E.coli and Shigella, and 974 for Campylobacter jejuni.

There are 3,085 unique sets of **AMR genotypes** represented in the data, which we are very interested in dissecting and using to supplement our AST phenotype information in our broader study. These genes will be parsed through during the process of clustering and evaluated for biological relevance in relation to drug response. Overall, mdsA (11%), mdsB (11%), blaEC (5%), and acrF (5%) are the most common antimicrobial resistant genes found in our isolates. Breaking it down by bacteria, within Salmonella, the most common genes are mdsA, mdsB, tet(A), aph(3'')-Ib, and aph(6)-Ib. In E.coli, the top 5 are blaEC, acrF, mdtM, tet(A), and aph(3'')-Ib. Finally, in Campylobacter, the most common genes are tet(O), blaOXA-193, 50S_L22_A103V, gyrA_T86I, and aph(3')-IIIa. We undoubtedly observe heterogeneity between the three species.

Next, we looked at the **min-same** and **min-diff** variables. We observed that the distributions of minimum SNP distance to another isolate whether that isolate be of the same isolation type or different isolation type, are generally right-skewed (**Figure 1**). This may speak to the presence of outbreaks, isolates closely related to one another that may even have occurred near the same time. Our E.coli and Shigella records seem to be slightly distinct from the other two bacteria in its distributions as if presents a bit more variance in its min-same and min-diff values.

We created a new count variable to indicate the number of isolates that are "close" to another isolate. We'll define a closely related isolate as an isolate within 7 SNPs of another. If an isolate is within 7 SNPs of another clinical isolate and another environmental isolate, that counts as 2. If an isolate is only 'close' to an isolate of the same type but not close to an isolate of a different type and vice versa, that counts as a value of 1. Thus, only going off of the data that is available to us, this variable will only take on discrete values from 0-2.

Briefly looking at the distribution of this new count variable, we observed that ~66% of isolates from E.coli and Shigella are not 'close' to any other isolates while this proportion is around 25-26% for Campylobacter and Salmonella. Our Campylobacter jejuni isolates are the most close with ~44% having a min-diff *and* min-same value less than 7. This is followed by Salmonella with 35% of records having a min-diff and min-same value less than 7, and then E.coli with only 9% of records having a min-diff and min-same value less than 7.

The last variable that we'll look at broadly is **computed types**. We start by separating computed types into the two pieces of information that the variable provides: antigen formula and serotype. Perhaps the most important piece of information that we want to extract from this variable is the antigen formula because we already have serotype information from the serovar variable. We checked to see if this information serotype and serovar from the two different variables, was comparable or not and found that it was. So, from the computed types variable, we only preserved the antigen formula information.

This left us with 126 unique antigen formulas. We reduced this further by looking at the distribution of antigen formulas represented. Upon further investigation, we found that the only isolates that don't have missingness for antigen formula are isolates that belong to Salmonella. Categories representing at least 2% of the Salmonella isolates remained their own distinct columns while categories representing less than 2% were grouped into an 'Other' category.

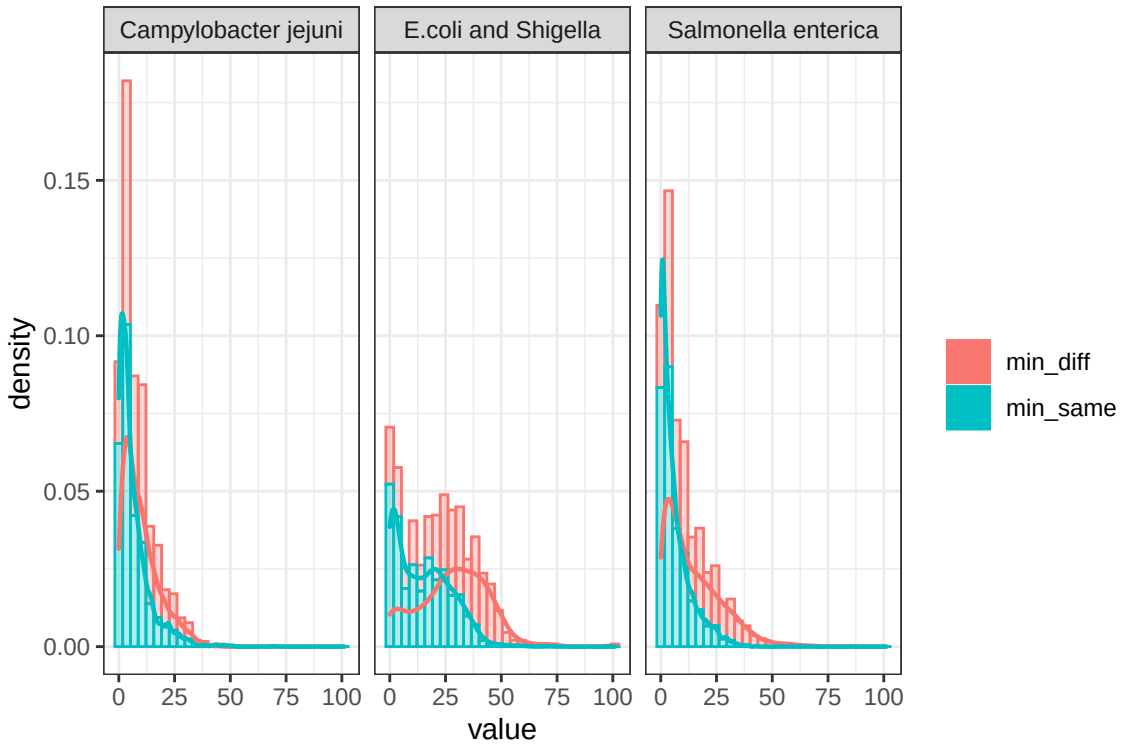

Figure 1: Distributions of min-same and min-diff of isolates belonging to each bacteria

Aside from the "other" category, '8:i:z6' (15%), '4:i:1,2' (13%), and '4:r:1,2' (8%), are the top three most common antigen formulas represented in the Salmonella isolates that we have records on.

To start to motivate and inform our broader study, we explored two initial questions.

Initial Subquestions

The first question that we considered is:

Are isolates within the same SNP cluster susceptible/resistant to the same antibiotics?

In order to assess this, we will compute the similarity scores between the drugs that isolates within the same clusters have. To do this, we'll use the `stringsim` function in the `stringdist` package in R (van der Loo M, 2014). The `stringsim` function computes pairwise string similarities between elements of a vector. It returns values between 0 and 1 (inclusive), corresponding to perfect dissimilarity at a value of 0, and perfect similarity at a value of 1. We used the default method for distance calculation, optimal string alignment (restricted Damerau-Levenshtein distance). We assessed the average similarity values for isolates within the same clusters. These values are broken down by drug resistance and drug susceptibility.

Table 4 shows that on average, the drug behaviors of isolates within the same cluster tend to be fairly similar with mean and median values hovering between 0.74-0.86. Despite high averages, there are some clusters that seem to have low similarity based on our data (min cluster average of 0.09 for drug susceptibility and 0.05 for drug resistance).

We decided to group clusters into two categories, 'close' and 'distant'. Close clusters are defined as clusters with average min-diff or average min-same less than or equal to 10 SNPS. All other clusters are categorized

Table 4: Drug Behavior Similarity between Clusters

	Susceptible Sim.	Resistant Sim.
Min.	0.0903446606874123	0.0528846153846154
1st Qu.	0.738099732302261	0.541858136668346
Median	0.851528384279476	0.783018867924528
Mean	0.826901569660862	0.74259930994572
3rd Qu.	0.942149292149292	1
Max.	1	1

Table 5: Average drug responses for close and distant clusters with closeness being defined as an average min-diff or min-same SNP distance of < 10 SNPS

Cluster Type	Average Susceptibility	Average Resistance	Num Clusters
close	0.9276393	0.8860877	1502
distant	0.8590357	0.7449361	314

as distant. We looked at the average similarity measures for close and distant clusters and found that in general, clusters that we considered 'close' had high similarity when it came to their isolates' drug responses (**Table 5**).

The second question that we considered is:

Does drug resistance and susceptibility change over time?

Since we've established that clusters tend to be very similar in terms of their drug responses, we approached this problem by looking at how the mean drug responses per cluster have changed over time. Collection date/Year is the variable that we used for time. In terms of drug response, an isolate within a cluster was given a value of '1' in either a variable representing drug resistance or susceptibility, if it was susceptible or resistant to that drug. Otherwise, an isolate was given a value of 0. When we group these isolates into their clusters and take each cluster's mean response variable by year, these means will be between 0 and 1 (inclusive) because we're taking means of binary responses.

For illustration purposes, we randomly sampled 5 clusters to show potential changes over time. Let's start with drug resistance to the most popular drug, tetracycline.

Evidently, we observed that in general, isolates that seem to be very similar in terms of their drug responses (i.e. isolates within the same cluster), demonstrate variability in their drug resistance over time (**Figure 2**). If we draw a line at 0.5 and classify average resistance at or above that line as being drug resistant and responses below that line as not being drug resistant, then we can see that for similar isolates, over time, mean responses have crossed the line between being drug resistant and not.

We did the same thing to observe changes over time for drug susceptibility.

We observed similar behaviors in drug susceptibility. Even in just a random subset of 5 clusters, we observed that over time, some clusters' mean responses crossed that 0.5 threshold of being susceptible or not being susceptible (**Figure 3**). Thus, we may infer that drug susceptibility and resistance are dynamic over time and is a complex behavior that we may want to understand and adjust for as we move forward in our study.

Limitations

We recognize that there are several limitations that our data pose to fully exploring our research question. As aforementioned, by only retaining records from the NCBI database that have non-missingness in drug response, we've limited the isolates and cases that we're able to observe as the drug response variable

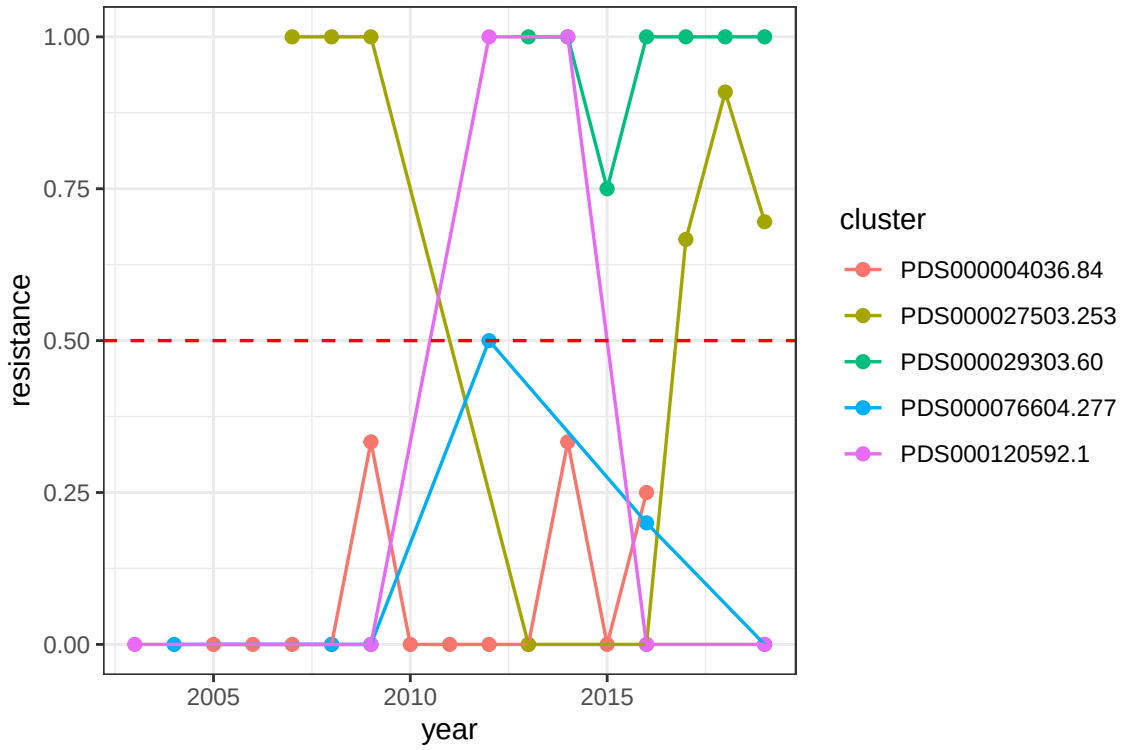

Figure 2: 5 clusters' mean resistance to Tetracycline over time

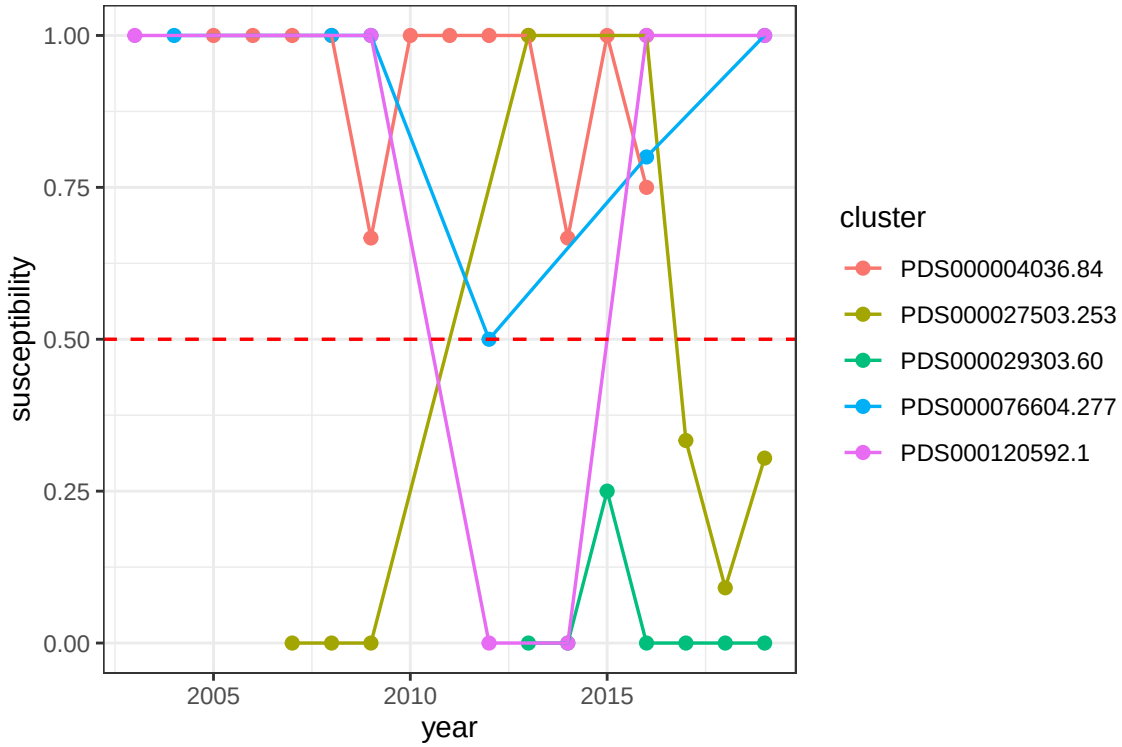

Figure 3: 5 clusters' mean susceptibility to Tetracycline over time

(`AST_phenotypes`) is more often than not missing. Additionally, this variable is defined by the data submitter, so there could be some error or objectivity involved in that process. The NCBI Pathogen Detection Help Documentation site notes that these data are typically submitted using CLSI or EUCAST guidelines, which change over time. Alternatively, drug response classification is decided by an automated instrument which may infer cutoffs. This could mean that records submitted using an earlier standard/guideline may have different resistance and susceptibility criteria for the same antibiotic compound than it would if using a later standard. Even for the same organism and same isolate, different tests may yield different results. This may bring some nuance into the problem space of what we observe in our data as criteria don't seem to be uniform overtime and vary depending on the different standards used to define drug response.

Additionally, some of the variables that we observe only have information for Salmonella isolates. These variables are antigen formula and serovar/serotype. And so, these variables may only be helpful to refine our future analysis for Salmonella but won't be able to contribute to our other two bacteria of interest. It's also a bit of a nuanced process to assess changes in drug response over time. Strains in our dataset are more unique than not, and it's rare that we observe the same strain twice, especially in different years. To try and remedy this, because of the similarity that we saw between isolates within the same cluster, we approached the problem by assessing each cluster's mean response and imposing a 0.5 threshold on those mean responses to categorize those isolates generally being susceptible or not, and resistant or not. This approach is also taking a strong assumption that if an isolate is susceptible or resistant to an antibiotic, then that would recorded in the data, which reasonably, doesn't always hold true. Additionally, we note that genomic responses to antibiotics are more than likely to change over time as bacteria start to learn more about the antibiotics that are used to treat them, and in response may eventually build a resistance to them. This problem space surrounding treatment is ongoing and constantly in flux due to this. Thus, we realize that what is represented in our data now may or may not be a good scope for what could be represented in data like this for future years.

Conclusion

The use of antibiotics for the treatment of foodborne illnesses is of growing interest in the field of human health. We have grasped a solid understanding of our data and the distributions of the variables represented in our data. We've learned that variables within the same cluster and variables that are considered 'close' in terms of their SNP distance, tend to have similar responses to certain drugs. Similarity in genomic, clinical, and environmental characteristics is going to define the conversation surrounding this study. By assessing the similarity of our isolates, we may be able to pull significant information about patterns of drug response and how they relate to these values/points of similarity.

Moving forward, we aim to go more in depth with our data and extract patterns and relationships between our covariates of interest as they relate to antibiotic resistance and susceptibility. We will specifically focus more on the genes found in our isolates and the connections that we can draw with antibiotic response in that aspect. The broader goal of this study is to investigate the efficacy of and build a classification model to predict a pathogen's response to an antibiotic (resistant or susceptible). Essentially, we want to approach this as a binary classification problem. This model will be trained to consider the optimal hyper-parameters and covariates for a well-considered bias-variance trade-off, and tested to assess accuracy in prediction. The specifications of the model(s) that we will want to try are not solidified as of now but we are considering logistic regression and multiple machine learning methods for binary classification like random forests. We will also assess how effectively we can cluster antibiotic resistant genes and assess how well those clusters may match and define our observed antibiotic resistance and susceptibility behaviors to treatment.

Overall our models will be assessed using classification and clustering performance metrics like area under the ROC curve (AUC), and silhouette scores.

References

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England), 26(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Hashempour-Baltork, F., Hosseini, H., Shojaee-Aliabadi, S., Torbati, M., Alizadeh, A. M., & Alizadeh, M. (2019). Drug Resistance and the Prevention Strategies in Food Borne Bacteria: An Update Review. Advanced Pharmaceutical Bulletin, 9(3), 335–347. https://doi.org/10.15171/apb.2019.041

Hassani, S., Moosavy, M.-H., Gharajalar, S. N., Khatibi, S. A., Hajibemani, A., & Barabadi, Z. (2022). High prevalence of antibiotic resistance in pathogenic foodborne bacteria isolated from bovine milk. Scientific Reports, 12(1), Article 1. https://doi.org/10.1038/s41598-022-07845-6

Rajaei, M., Moosavy, M.-H., Gharajalar, S. N., & Khatibi, S. A. (2021). Antibiotic resistance in the pathogenic foodborne bacteria isolated from raw kebab and hamburger: Phenotypic and genotypic study. BMC Microbiology, 21(1), 272. https://doi.org/10.1186/s12866-021-02326-8

Ren, Y., Chakraborty, T., Doijad, S., Falgenhauer, L., Falgenhauer, J., Goesmann, A., Hauschild, A.-C., Schwengers, O., & Heider, D. (2022). Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. Bioinformatics, 38(2), 325–334. https://doi.org/10.1093/bioinformatics/btab681

Van Camp, P.-J., Haslam, D. B., & Porollo, A. (2020). Prediction of Antimicrobial Resistance in Gram-Negative Bacteria From Whole-Genome Sequencing Data. Frontiers in Microbiology, 11, 1013. https://doi.org/10.3389/fmicb.2020.01013

National Center for Biotechnology Information. (n.d.). Home - Pathogen Detection - NCBI. National Center for Biotechnology Information. Retrieved October 23, 2022, from https://www.ncbi.nlm.nih.gov/pathogens/

van der Loo M (2014). "The stringdist package for approximate string matching." *The R Journal*, *6*, 111-122. https://CRAN.R-project.org/package=stringdist.