Introduction, Review, and Exploration

Nathan Provost and Antonella Basso

Introduction and Review of Literature

Listeria monocytogenes is a particularly problematic species of bacteria that causes numerous cases of food-borne illness each year both in the United States and across the world. The most vulnerable populations to its effects consist of infants, pregnant women, the elderly, and the immunocompromised, who frequently experience the common symptoms of infection that include flu-like symptoms (nausea, vomiting, fever, etc.) as well more severe symptomatic manifestations. (Rogalla and Bomar 2022) While it is often characterized medically as a food-borne illness, being found frequently in cheeses, cold meats, and unprocessed/improperly processed diary products, it exists environmentally in the soil and decaying organic matter, but does not surface as a point of infection nearly as frequently as it does in food-processing areas such as farms or improperly cleaned production facilities. (Rogalla and Bomar 2022) (Ward et al. 2004) Furthermore, the severity of strains extracted from differing environments, or more specifically, that have evolved to suit different environments, thus suggesting an increased presence, has been a subject of comparison and evaluation in recent literature (Ward et al. 2004), which is where the basis of our line of questioning lies.

It is essential to elaborate on the results of one recent article (Ward et al. 2004) that provides a rigorous overview of the differences in strains of listeria by serovars, small biological distinctions in the bacteria that separate one subgroup of the species from another. Developing robust methods of separating one strain of listeria from another is essential to providing comprehensive analysis of its pathology, since differences in serovars are often related to differing levels of disease manifestations in human beings as shown in this article. (Ward et al. 2004) Its results present a a phylogenetic tree with over 60 strains of listeria. These strains are grouped into three lineages (LI, LII, and LIII) which separate them by small genetic differences. These lineages are further organized by several different serovars, which distinguish the strains through ever more minute genetic and physiological differences. The sources from which each specimen was obtained is also provided, but the detail behind these sources is minimal, since each sources is listed as either human, animal, food, environmental, or missing. (Ward et al. 2004) While these source types provide some degree of clarity, the lack of specific details pertaining to these sources is unhelpful for further analysis of the relationship between sources and strains.

An important first step in sorting the impact of listeria on humanity is examining the variance of severity among different strains. It has been shown (Muchaamba et al. 2021) that strains possessing serovar 4B have led to longer virus survivability in organisms (zebrafish were used), which is instrumentally tied to worse clinical symptoms overall. (Muchaamba et al. 2021) This kind of distinction is of great importance to the medical examination of listeria, since any implication of differing severity across strains could better inform our decisions when faced with the isolation of one strain versus another. A specific result from this study showed that LI strains of listeria yielded an 85% mortality rate in the tested zebrafish population, whereas the LII strain only yielded a 17% mortality rate and the LIII strain only yielded a 2.5% mortality rate. These survival rates were fitted using a standard Kaplan-Meier estimation curve. (Muchaamba et al. 2021) From this report, it is clear that not all strains of listeria pose an equal threat to biotic organisms (the study uses a population of zebrafish but suggests that the gravity of the results can be generalized to human populations), but further trends have yet to be identified in this study. Specifically, a crucial point of interest is whether or not different sources of listeria are associated with more or less aggressive strains.

Studies have also focused exclusively on strains of listeria that have proven to be most severe compared to their counterparts. The study lists other subspecies of listeria that almost always do not cause human illness, but then proceeds to discuss the most potent strains of *Listeria monocytogenes* that were encountered. The

26 strains arose in Bucharest, Romania and were all clinically isolated from a total of 24 patients who were presenting with conventional symptoms over the years 2009 to 2013. (Borcan et al. 2014) Three clinical origins were specified: blood cultures, placenta swabs, and cerebrospinal fluid. Over half (16 in total) of the samples came from cerebrospinal fluid, while only a single isolate came from a placenta swab. The most common serovars among these isolates for the hospitalized patients were 1/2a, 1/2b, 3a, and 3b, but serovar 4b was also prominent among the selection of specimens. All of these serovars demonstrated resilience to traditional unifaceted antibiotic treatment, but there was no significant resilience shown against multidrug methods. Collectively, this study again demonstrates that the most severe strains of listeria present specific genetic characteristics that separate them from other strains, yet this approach is limited to solely clinical data, and does not offer any insight into where these people could have first encountered listeria. (Borcan et al. 2014)

A more expansive approach to investigating the source-severity dynamic of listeria in the human population examines the movement of listeria through several *different* food products across eastern Europe during two distinct time periods (from 2001 to 2005 and 2019 to 2020). (Psareva et al. 2021) These strains were only of two distinct lineages (either LI or LII) and they all came from food products broadly consisting of dairy, meat and poultry, and fish. Strains from the first outbreak period (from 2001 to 2005) were of greater serovar diversity and contained a larger proportion of the LI strains of listeria, which as previously mentioned (Muchaamba et al. 2021) have been associated with worse symptomatic manifestations overall. Furthermore, LI strains were shown to be associated with dairy products in greater proportion than other products included in the study, whereas LII strains were shown to have originated from a more balanced proportion of dairy products versus fish and meats products (combined). It was shown that dairy products have a statistically significant association with greater specimen diversity when compared to the other two groups. (Psareva et al. 2021) Furthermore, the essential conclusion of this study was that dairy products seem to serve as the main origin of LI strains in this outbreak, which is instrumentally tied to the severity of infections from dairy products since LI strains have been shown to yield greater clinical severity in the past. (Muchaamba et al. 2021) This study therefore points us in an important direction when it comes to matching sources to strains, since it indirectly proves an association between more severe strains and dairy products.

To better understand which specific kinds of listeria are found where, several methods have been used to group (or more accurately cluster) them together. A study conducted less than three years ago made use of single linkage clustering in the process of backtracking and forward checking the propagation of listeria through meat distribution in the case of a particular provider. (Luth et al. 2020) This case's methodology was a promising point of inspiration for us, since we wondered whether or not we could employ similar methods from a more direct, intrinsic source. The study employed single-linkage clustering to group different listeria outbreaks and isolates by genetic makeup through a process called core genome multi locus sequence types (cgMLST). (Luth et al. 2020) The empirical rule for assigning clusters was that two isolates would be placed together as long as they had less than eleven genetic allele differences in accordance with the dictation of cgMLST. While this means that the strains were not directly clustered by isolation source (likely due to the fact that the potential sources were limited to either food, food processing, or clinical detection), the sources were readily comparable with the clusters themselves through several visuals provided in the study. Most notably, two clusters comprise all of the clinical sources in listed in the data, while the remaining 15 clusters have strictly isolates from food or food processing environments.

These two clusters are both part of a single outbreak that occurred in Germany over the years 2013 to 2018, with all of the cluster 2 cases falling between 2015 and 2017 and the cluster 1 cases spanning the entire period. Cluster 1 had 72 cases in total and cluster 2 had 11 cases in total, resulting in the outbreak consisting of 83 total cases. The remaining clusters are not discussed at length, which is not empirically helpful, but it is mentioned that clusters 9, 10, and 12 through 16 all had an identical medical feature (they all had the same virulence factor composition, which consists of genes associated with the outbreak clusters) that tied them to the two outbreak clusters. (Luth et al. 2020) In a general sense, a Mann-Whitney U test done by the study found that gene counts pertaining to virulence in clusters 1 and 2 differed extremely significantly (with a p-value less than 0.0001) from the gene counts pertaining to virulence of all the other clusters. While this realization is important, it is somewhat obvious given the background the study provides, and does not critically examine the potentially significant association between source and strain severity. (Luth et

al. 2020) This study made promising progress in sorting the strains of listeria by genetic differences, going even further than sorting by serovar, and its use of traditional clustering methods was relatively successful and highlighting some trends pertaining to the disease, but it is important to explore other, more general methods when studying a disease that had been shown to be multifaceted and intrinsically complicated.

Going back to the environmental observations we previously referenced (Ward et al. 2004), applying a cluster-style methodology to group listeria specimens by source could be an important preventative step in modeling the propagation of listeria through human systems. Furthermore, the dynamics of bacterial growth as estimated by frequentist, parametric models have been known to exhibit error as a result of confounding experimental variability when estimating necessary parameters (not specifically in listeria growth studies, but in studies examining the growth of similar organisms). (Pouillot et al. 2003) One recent study has taken steps to overcome one side of this dilemma by employing hierarchical Bayesian inference to model the microbial growth of listeria originating strictly from milk samples (that consist of either whole, skim, or partial-skim milk). This model was able to convey a greater deal of nuance by listing specific mean and standard deviation estimates for several strain subsets based on uninformed prior distributions and a likelihood distribution that is semi-uninformed, with slight influence from experimental data concerning microbial growth. (Tonner et al. 2020) Ultimately, the models used to simulate microbial growth yielded convergent Markov chain Monte Carlo estimates and produced results that were consistent with the researchers' observed results (though no further comments are details on this are provided). A table of the generated means and standard deviations for each of the model parameters, of which there are numerous, was also provided. These parameters corresponded to various physical attributes of listeria, such as the minimum and maximum temperatures at which it can grow, and the maximum temperature the milk could have reached. The lack of elaboration and source diversity in this study is one that seems to arise in many nuanced modeling approaches. A similar paper that investigated the growth of listeria in smoked salmon also took a Bayesian modeling approach that was quite similar to the one used in the other study, but also lacked source diversity as it only considered listeria that grew on smoked salmon and immediately related flora. (Deligentte-Muller et al. 2006).

More intense, theoretical approaches have recently been introduced when it comes to modeling the behavior of food-borne illnesses like listeria. A recent study (Wang et al. 2021) explored a variety of machine learning methods when trying to classify a wide variety of food-borne illnesses, one of which was listeria. While the focus of the study was on other food-borne illnesses, the researchers emphasize that the approach is easily generalizable to similar microbial food-borne illnesses and include listeria in their aggregate dataset of potential illnesses to which the methods can be applied. Four different machine learning methods were applied to the analysis of these food-borne illnesses: basic decision trees, random forests, gradient boosted decision trees, and adaptive boosting methods. Each of these methods can be collectively viewed as nonparametric, which is good for the purposes of escaping the limitations of parametric models that arise from the experimental errors and imprecisions that hinder an analysis of microbial behavior. (Pouillot et al. 2003) It is good to specify the benefits and drawbacks of each approach, not only in terms of their theoretical structure, but also in terms of their functionality in view of the dataset.In the context of this study, basic decision trees use source information, as well as other biological and scientific post-isolation information (such as serovar type, location, etc.) to classify the given strains by lineage and serovar. This way the model could be used in the future to quickly predict the kind of strain taken from a given source without extensive laboratory testing. Unfortunately, the basic decision tree method performed the worst of the four methods with only 63% accuracy, likely due to its simplistic structure. (Wang et al. 2021) Random forest methods fared a little better than decision trees, since their primary structure consists of using multiple decision trees. This improvement was rather small however, since the accuracy of this method for the food-borne illness data was only 64%, a 1% improvement over the previous model. (Wang et al. 2021) The methods start to improve substantially when the other two approaches are used to model the data. The gradient boosting decision tree (GBDT) method follows a similar initial setup to that of a random forest, but instead of using bagging methods to sample from the potential outcomes, it uses a boosting method. This distinction proved to be much better than the previous approaches with 69% accuracy overall, marking a 5% increase over the random forests method. Adaptive boosting was also tested, whereby poor classification methods are tried but then assigned lower weights when compared to better fitting classification methods. This did better than basic decision trees and random forests, but not as well as the GBDT approach, since adaptive boosting was only 67% accurate. (Wang et al. 2021) Ultimately, we see that nonparametric machine learning methods possess a

degree of power when it comes to predicting microbial strains of food-borne illness. Specifically, upwards of almost 70% accuracy was achieved in the previous study (Wang et al. 2021), which suggests that we may be able to recreate a nonparametric machine learning approach that better models predicts that strain of listeria based on several relevant biological covariates.

This body of previous work has fallen into two heavily divided categories that we believed should be reconciled to gain a better understanding of the origin and propagation of listeria. Research that uses clustering methods usually makes use of a traditional clustering method (like single-linkage clustering in the previously discussed study (Luth et al. 2020) or k-means clustering), which is effective, but is subject to parameter misspecification that arises when trying to experimentally determined which parameters are best to use (Pouillot et al. 2003). As a result, using prior distributions in a Bayesian framework can alleviate some of these misspecification issues, since parameters are allowed to vary stochastically. However, most Bayesian methods aim to model growth rates of listeria (as in (Tonner et al. 2020) and (Deligentte-Muller et al. 2006)) as opposed to any form of relationship between strains and sources. It would be better to combine Bayesian framework with a clustering-style approach to investigate the frequency of certain strains, lineages, and serovars in conjunction with their source of extraction. Furthermore, nonparametric machine learning methods like those previously discussed (Wang et al. 2021) have shown potential in predicting strains of microbial food-borne illness using pertinent biological data like source information, location, and/or time.

It may therefore be fruitful to combine the freedom of a nonparametric machine learning method with Bayesian modeling structures and clustering methods to best predict the strain of a given listeria isolate using several biological covariates. The idea is to let there be a great variety of flexibility within the model, allowing for an examination of the source of a listeria isolate and how this affects the probability of such an isolate having a particular serovar, being from a specific lineage, and being a particular strain. While some relationships between certain food products and worse strains of listeria have already been established as previously discussed (Psareva et al. 2021), a broader unifying model for all potential sources had not been thoroughly explored using such a model. Many qualitative clinical studies have investigated the relationship between strain, serovar, and severity of symptoms, but these studies looked to establish non-quantitative relationships through the use fo structures like phylogenetic trees and crudely estimated survival rates. (Ward et al. 2004) (Muchaamba et al. 2021) A Bayesian nonparametric clustering method which combines all three aspects of these approached might therefore be of interest to both practitioners and analysts, as it would incorporate positive aspects from each of these models and yield quantitative evidence either in favor or against the implications made by any previously constructed phylogeny of listeria. (Ward et al. 2004) (Muchaamba et al. 2021)

Collectively, the body of work surrounding listeria is vast, but points to the same general conclusion. Listeria has been widely separated by serovar and lineage, implying that some strains are more potent and destructive than others. Specifically, LI lineage strains and 4b serovar strains of listeria are tied to severe cases of listeria in human beings, as proven clinically over several studies.((Ward et al. 2004) (Muchaamba et al. 2021) (Borcan et al. 2014) (Luth et al. 2020)) The growth of listeria and similar organisms has been successfully modeled with Bayesian hierarchical models, which provided a bigger picture into the modeling possibilities pertaining to listeria in the human population. ((Tonner et al. 2020) (Deligentte-Muller et al. 2006)) Standard clustering has also proven useful, though the extent to which this clustering has been performed is limited and could be expanded to more general datasets and more general techniques. (Luth et al. 2020) Nonparametric machine learning methods have also experienced a relative degree of success in clustering microbial outbreaks, but these methods are limited in by the parametric impositions of the frequentist paradigm and have not explicitly been applied to the analysis of listeria. Synthesizing a Bayesian nonparametric clustering approach and comparing it to a traditional k-means clustering approach may therefore be of use when furthering the body of research concerning the possible ties between listeria and its source. In the exploration that follows, we will explore varies relationships between the sources of listeria isolates and the strains (as evidenced by serovar, since lineage data is spurious and specific isolate methods would be too convoluted). We will try also to investigate other possible avenues of severity and see how they compare to serovars alone. By building distributions for our variables of interest, we can get a sense of the behavior of listeria in our dataset, and compare this to the behavior recorded in the literature.

Data Exploration

The data we will be working with is taken from the National Library of Medicine which is operated and organized under the auspices of the National Institutes of Health. (NLM:NCBI 2022) The central contributors to this database are the CDC, the FDA, the USDA, and PHE (Public Health England). Collected by numerous different agencies this database encompasses a wide variety of often profoundly unreliably labeled variables, some of which are missing in great quantities as discussed below. This dataset contains an excessive array of variables, many of which are either missing in great quantities or not relevant to our study, so we will begin by discussing the most important variables first. In terms of the collection of this data, researchers submit their results to the NIH for the NLM and report variables that they have for each isolate. As we will see, this leads to many different ways of denoting the same feature for a given observation, which makes comprehensive analysis extremely difficult. We discuss examples of this below. Several variables (for example, serovars) are recorded clinically uses standard biological procedures, whereas other variables are automatically observed by the researchers (like location). It is up to the submitter, however, to properly list each variable, for even if the information is available, it may not always be recorded. This provides a rough outline of how the data in the source was collected and where it came from.

We begin by explaining each covariate and the percentage of entries missing for that given variable (which accompanies each variable name in parentheses). The strain (17%) is given, along with its serovar (81%), the organism's group (not missing), the outbreak (not missing, assuming all empty entries are sporadic), the location of the isolate (12%), the isolation's source (22%), the date of collection (16%), and two minimum distance variables. The distance variables record the minimum distance from a given isolate to an isolate of the same strain (27% missing) and an isolate of a different strain (49% missing). Our plan is to conflate these two variables to gauge the minimum distance from a given isolate to an isolate of *any* strain, and so we took the minimum of these two variables and will use this new covariate as a way to examine this generalized form of distance. This new distance variable was missing to a slightly smaller degree (26% missing). We note that these distance variables denote the distance between to isolates in a given SNP cluster (an organization of the isolates by genetic information), which is a variable that we do not use directly, since its influence is felt through the use of the distances. Furthermore, we may also consider the number of contigs (genetic components of the isolates, where N50 is a related contig property that represents redundant information in our case) a given isolate has.

Naturally, there are many more variables in this dataset that are not of immediate use to us, either due to missingness of bureaucratic irrelevance. We will quickly go through these variables and justify their omission in our analysis. Which initiative (or bioproject) is recorded completely, but since any demographic information would be better represented through location and time data, we omit it. Software version variables and analysis type variables are not of any use to us since they are not tied to the biological properties of listeria or its origin. Enzyme pattern variables (93% for primary and 94% for secondary), host disease (87%), phenotype and genotype variables (100% for both), computed types (100%), the isolate's host (78%), stress genotype (65%), and IFSAC category (65%) are all missing in such profound percentages that using them as covariates would be impractical. Specifically, if we were to try to apply any method to replacing them (either through inverse probability weighting or imputation), we would either be working with an extremely small amount of data (in the case of weighting) or we would be engaging in a excessively and wastefully laborious process (in the case of imputation). As a general rule of thumb, we will only work (extensively) with covariates that are $33.\overline{3}\%$ missing or less (though we make **exploratory exceptions** for **serovar type** since this variable comes up immensely frequently in related literature; however, we may not be able to impute or adjust these values in future models). Furthermore, scientific name and organism name listings are uninformative and pleonastic, since we know that we are working with listeria in this dataset.

Already, it is evident that this data will be challenging to work with and analyze numerically. This is obviously due to the fact that so few of the variables we have to work with are numerical, which means direct quantitative analysis difficult. This is one of the primary limitations of the dataset that we will have to circumvent in our analysis, since visualization and statistical analysis dependent heavily on the numerical inputs of a given dataset. Another immediate limitation is the intense missingness of so many variables, as previously mentioned and discussed further below. Many of the variables we would like to consider are missing in drastic quantities, which render them unusable in most modeling contexts and even in some exploratory

contexts. As a result, we have to fix our focus on a specific set of key variables that we can explore and investigate, in order to get a better sense of any possible modeling avenues that seem fitting or any basic trends that are worth observing and addressing when it comes time to implement a model.

An immediate adjustment can be made for the sources of listeria, since we technically have two variables to work with. In addition to the source type listed above, there is another variable ("isolation type") that has a smaller percentage of missing entries (11%), but is far less informative since it only lists entries as either environmental or clinical. However, it is better to have *some* information on the source of isolation than none at all, so we can conflate these two variables to improve the missingness of the isolation data. Essentially, if any observation in the specific isolation variable is missing, but there is a general observation (either environmental or clinical), we fill in the empty entry with the general information available. Through this transformation, we reduced the missingness on the isolation source variable to only 7%, which is a significant improvement, and since there were some entries in this column that were already vaguely listed as either environmental or clinical to begin with, we have not lost a great deal of information by performing this transformation. Now that we have shaped, contextualized, and focused this dataset into a structure that is more manageable, we can begin to explore the trends of missingness between variables along with variable distributions that arise naturally among the different covariates.

An immediate concern we wanted to address was the potential difference in sources between data entries that had minimum distance entries and observations that *did not have* distance entries. Any pattern detected here could point to an association between the potential difficulties that come with collecting distance data and specific sources. After separating the data into these two subsets, we examine the top 10 most frequent sources for the subset with distance data and the subset without distance data. Table 1 below contains the information on the top ten sources of isolation for each subset. As we see, there is some repetition in the notation of sources, since several sources could easily be combined and/or generalized. For example, both food and food products could be combined into one observations, just as the environmental variables (environmental, environmental swab, and environmental swab sponge) could be collapse into one observation. As expected, the most common source reported for data entries that are missing distance observations was no source at all (that is to say the sources are quite often missing when distance observations are missing). This gives us a vague idea of how missingness operates in this dataset, but since only 26% of the data is missing distance values, the sizes of each of these subsets is are not of the same magnitude. Hence, it is more instructive to examine the proportions of these frequencies.

Table 1: Most Frequent Sources for Each Subset

Source (Distance)	Number of Occurances	Source (No Distance)	Number of Occurances
Food products	554	environmental	169
human listeriosis	941	clinical isolate	170
water sediment	1026	soil	212
environmental swab sponge	1432	environmental swab	218
human	1670	human listeriosis	295
blood	2102	human	319
environmental	2261	blood	978
food	3985	food	1056
environmental swab	4503	clinical	2435
clinical	5711	NA	3678

Taking these top ten sources, combining repetitive variables, and then examining the four remaining common sources between the two variables, we can get a better understanding of the patterns of missingness by distance. We found that about 21% of the subset with distance measurements came from environmental sources, while only about 4% of the observations without distance sources were environmental, while the frequencies for the distance subset and the no distance measurement subset were about 7% and 4% for human sources, 5% and 4% for blood sources, and 10% and 8% for food sources, respectively. An immediate distinction is present in the comparison of the presence of environmental type sources between the two

subsets. Entries with distance measurements seems to come from environmental sources with overwhelmingly greater frequency when compared to entries without distance measurements. This trend is of concern to us, since it could potentially impact the analysis of a possible strain to source relationship when adjusting for other covariates like distance from strain to strain. The other dominant sources seem to be relatively comparable, with general human sources and food sources be slightly more frequent for entries with distance data, while entries without distance data are more frequently sourced from blood when compared to their counterparts with distance entries. To explain the latter phenomenon, it may be due to the fact that when blood is drawn clinically, the distance to a given strain is not likely to be an immediately assessed factor, since other clinical observations and actions are likely to take precedence in such a setting. While this gives us a good idea of missingness mechanisms, it is also important to consider distributional structures for this dataset. Unfortunately, there are not many continuous variables here, which might make clustering and further analysis challenging. However, there a two main variables of interest that should be further explored.

The usage of dates throughout this dataset is spurious and quite challenging to work with. Some of the entries provide exact days and months, some only months, but most of the dates are either missing or provide only the year. However, enough provide yearly information that we can analyze the change in our data over time to a least some extent. To get an idea of the time dynamics of this data, we will examine the subsets of the data that simply put down the years 2010, through 2020 (and no other information). In future modeling, we want to restrict ourselves to data from the past ten years when conducting our analysis, so this setup makes sense. With this, we can examine the probability densities of our minimum distance variable over this ten year period. Figures 1 and 2 below show the densities of the minimum distance to another isolate by each yearly group from 2010 to 2020. As we see, the tails of each of these distributions behave quite similarly after distances of magnitude 10 or greater. but the initial behavior varies greatly, which is often specific interest to us. In the first half of the period, the year 2015 seems to present much higher probabilities of have isolates with lower minimum distances to other isolates. 2013 and 2014 seem to present comparable middle range probabilities of having low minimum distance, while 2010 and 2012 have even lower probabilities of having low minimum distances. The year 2011 presents itself as an outlier in this regard, due to its lower probability of having low minimum distances between isolates. Its distribution also exceeds the other distributions notably between distances of 5 to 10 in magnitude. Turning our attention to the other half of the period, we notice that 2018 had a similarly high probability of low minimum distances between isolates, while 2016 and 2017 trailed these two years, and 2020 presented low probabilities for low minimum distances. We note also that 2020 was impacted by the introduction of COVID-19 and that this may also play a part in its lower, wider distribution overall.

Figure 1: Probability Densities of Minimum Distances (2010 – 2015)

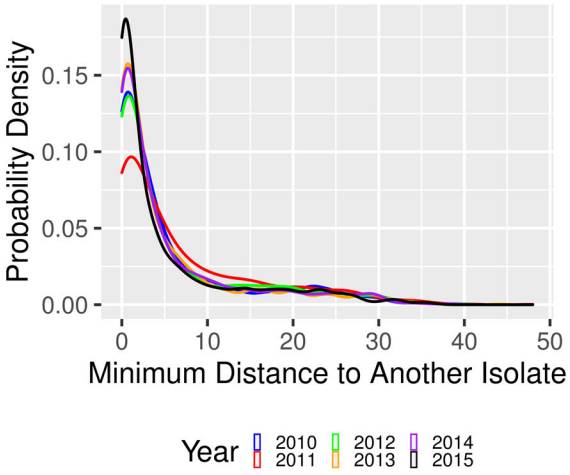
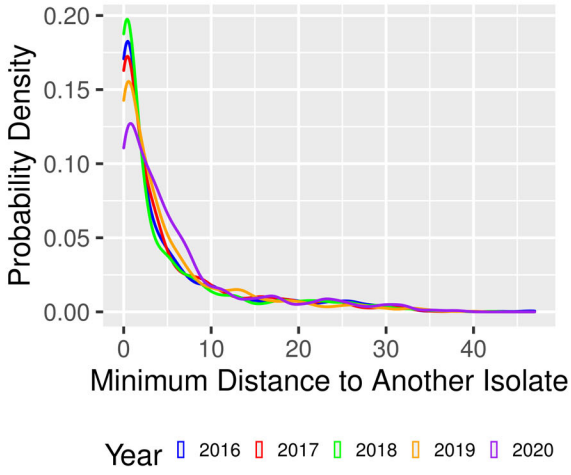


Figure 2: Probability Densities of Minimum Distances (2016–2020)



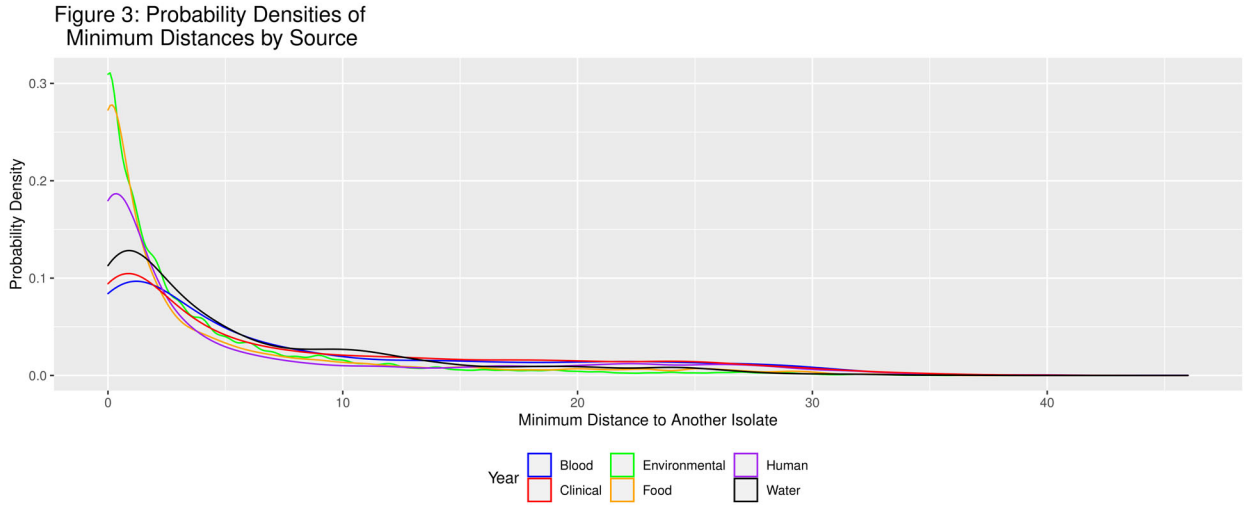
This information is important to us when we are planning our analysis approach of this data. Some of the highest probabilities for small minimum distances reach around 20% or higher (like in the cases of the

years 2018 and 2015), while others tend to exhibit middle ground low minimum probabilities (like 2019, 2013, and 2010). The lowest by far in terms of near-zero minimum distance probabilities occurs in 2011, which presents a wider distribution overall. These kinds of groupings are important to acknowledge when considering clustering methods, since they provide a rough outline of which subsets of the data resembled each other. Furthermore, since we want to examine the severity of strains of listeria, using minimum distance measurements is a good approach, since a high density of low minimum distances would suggest larger degrees of disease propagation, which could be used as one way to look at severity. Our main interest, however, is not the temporal effect on strain severity, but to examine the association between a strain's source and its severity. To do this, it may be more instructive to examine the densities of minimum distances that are grouped by similar sources. In doing so, we can examine the relationship between disease propagation and source, which is directly tied to our goal.

Using the left half of Table 1, we can boil down the top 6 collective sources of listeria and examine their minimum distance densities over each respective subset. We therefore consider clinical sources, environmental sources (which we take to mean "environmental", "environmental swab", and "environmental swab sponge" entries), human sources (taken to be "human" and "human listeriosis" entries), blood, food (taken to be "food" and "food processing" entries) and water sediment. Figure 3 captures our results below, while Table 2 provides a summary of the mean and standard deviation of the minimum distances by source.

Table 2: Descriptive Statistics of Minimum Distances by Source

Source	Mean	Standard Deviation
Environmental	3.714739	5.776042
Food	7.981446	9.183731
Human	6.061279	9.172677
Clinical	8.184031	9.269629
Water	5.628655	6.951085
Blood	7.981446	9.183731



There is a clear trend among the minimum distances and the source from which a given isolate originates. The probability of the minimum distance between an environmental isolate and another isolate being close to 0 is almost $33.\overline{3}\%$, which is a massive concentration of the data in comparison to clinical and water sediment isolates. Furthermore, food-borne listeria isolates also resemble this strong tendency, with over 26% being the probability of having a near zero minimum distance. This tells us a lot about the behavior of strains that come from specific sources, and the intrinsic point that is being made makes sense from a practical standpoint. Isolates from environmental or food-borne sources can easily spread through their medium of

extraction (for instance, if a field of soil has been widely contaminated, then the surrounding environment will likely yield isolates as well; similarly, if a company sends out an entire batch of food to a given area, multiple cases of listeria are likely to occur within a small area). Direct clinical isolates, blood samples, and water sediment samples are all more likely to have greater minimum distances between isolates, since the mediums through which these instances occur have greater ability to be separated or spread out from a given exposure. It makes sense that the human cases are in between these two groups, since isolation identification among human subjects would not seem as contained as a direct clinical isolate, but we would also not expect as widespread of an effect as environmental or food-borne contamination. Therefore, the trends we observe in Figure 3 make perfect sense in terms of the movement of listeria as a disease.

We have clearly established a few important trends regarding the source of a given listeria isolate and its minimum distance to another isolate. Certain groupings of years demonstrate closer minimum distances, suggesting a greater deal of widespread propagation of listeria. A trend that is not visualized due to its relative obviousness is that sporadic outbreaks tend to have a much lower chance of having a low minimum distance. A simple student t-test between minimum distances for outbreak cases versus sporadic cases yielded a p-value of 2.2×10^{-16}, which means that there is a significant difference between the mean minimum distance for outbreak cases (which for this dataset had a magnitude of 2) and the mean minimum distance for sporadic cases (which had a magnitude of 5). Next, it will be important to investigate the possible relationships between the number of contigs and various other covariates in this dataset.

We can make use of our already created source subsets to investigate the contig distribution across sources. Table 3 displays the descriptive statistics related to the contigs counts for each source grouping. As we can see, there is a discernible trend in the differences between source groups. Human isolates and environmental isolates have the highest mean contig counts among the sources in the table, but they also have the highest standard deviations. The water sediment isolates had much lower mean contig counts with a much smaller standard deviation, and the remaining sources fell in between these two extremes.

Table 3: Descriptive Statistics of Contigs by Source

Source	Mean	Standard Deviation
Environmental	45.85695	48.69073
Food	32.62955	34.66450
Human	50.00558	50.17896
Clinical	35.06519	40.85366
Water	25.99734	21.13990
Blood	32.62955	34.66450

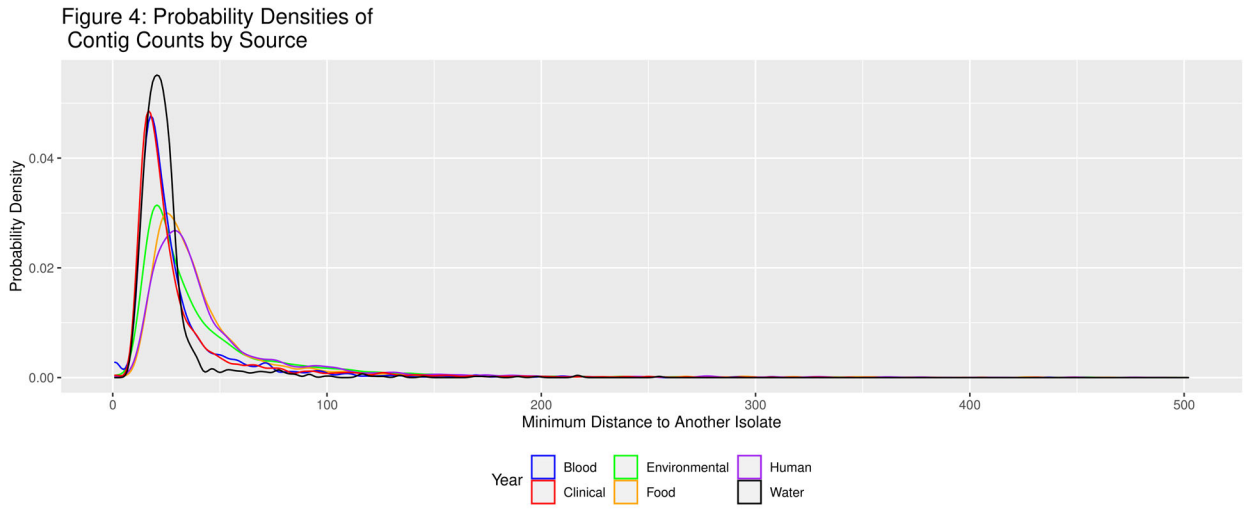


Figure 4 displays these densities above, which differ substantially from the distance densities. First, we note that there is an immense amount of skewing that occurs in the case of several of these distributions, most notably the distribution for contigs in water sediment observations. These outliers will undoubtedly influence the means present in Table 3. More importantly, the shape of these distributions is a bit more compelling than the shape of the distance distributions. While the minimum distance distributions followed shapes that looked vaguely exponential, or perhaps even Weibull-like in some cases, these distributions all strongly resemble variants of the F distribution. This is because all variables have positive support and maintain a quasi-normal looking center, but than quickly trail off with heavy tails. Water sediment isolates, blood isolates, and clinical isolates demonstrate very narrow distributions in accordance with their descriptive statistics, which would imply that a fitted F distribution with large, equal degrees of freedom would suit their forms. On the other hand, human, food, and environmental isolates demonstrate a wider range of spread. Lower valued, unbalanced degrees of freedom in an F distribution would therefore provide a more suitable form to these data subsets.

Combining this with what we know about the distributions of the minimum distances, it seems like minimum distance and high contig count could be negatively associated overall. Environmental, human, and food related isolates are all more likely to have higher contig counts and lower minimum distances from sources as suggested by the density graphs we've assembled, whereas water sediment, clinical, and blood isolates are all associated with lower contig values and higher minimum distances. This gives us a better idea into the roles of contig values in terms of listeria's capabilities and behavior in different sources. These kinds of relationships are essential when trying to create an analytical bridge between a given strain of listeria and its source, as they instrumentally tie potential severity to the source of isolation. Following this analogous comparison, we can compare the dynamics of contig values over the years by visualization their probability densities.

Figure 5: Probability Densities of Contigs by Year (2010 – 2015)

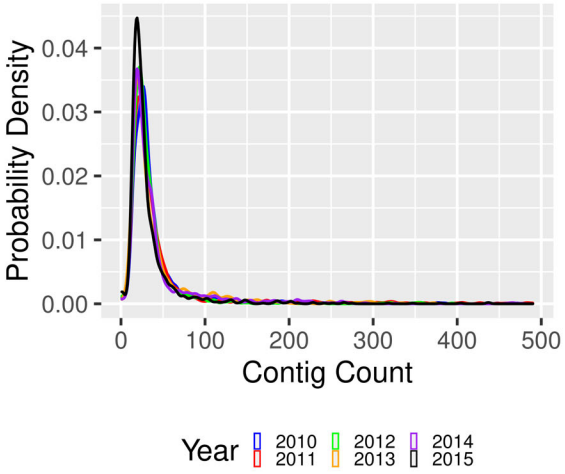
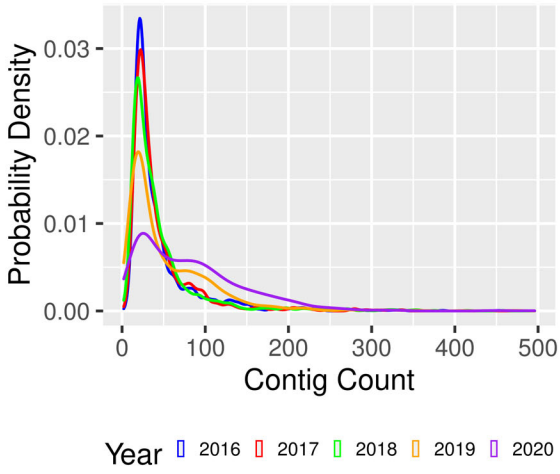


Figure 6: Probability Densities of Contigs by Year (2016 – 2020)



Figures 5 and 6 display densities for contigs by year. The trends we see here are reflect the current idea that contig count and minimum distance between isolates are negatively associated. Again we see that the distributions of contig values across the years look vaguely F distributed, with specific emphasis that the distributions for the years 2019 and 2020 show some degree of aberration in the form and geometric structure of their densities. We begin by addressing the distributions from the first half of this ten year period. Strangely enough, all of these distributions look quite similar, with one noticeable difference being that the mean contig count for the year 2015 seems to be much lower (with greater probability of having a lower contig count) than other years. However, the other distributions seem to be extremely similar, which is an odd contradiction of the previous phenomena we observed with the sources.

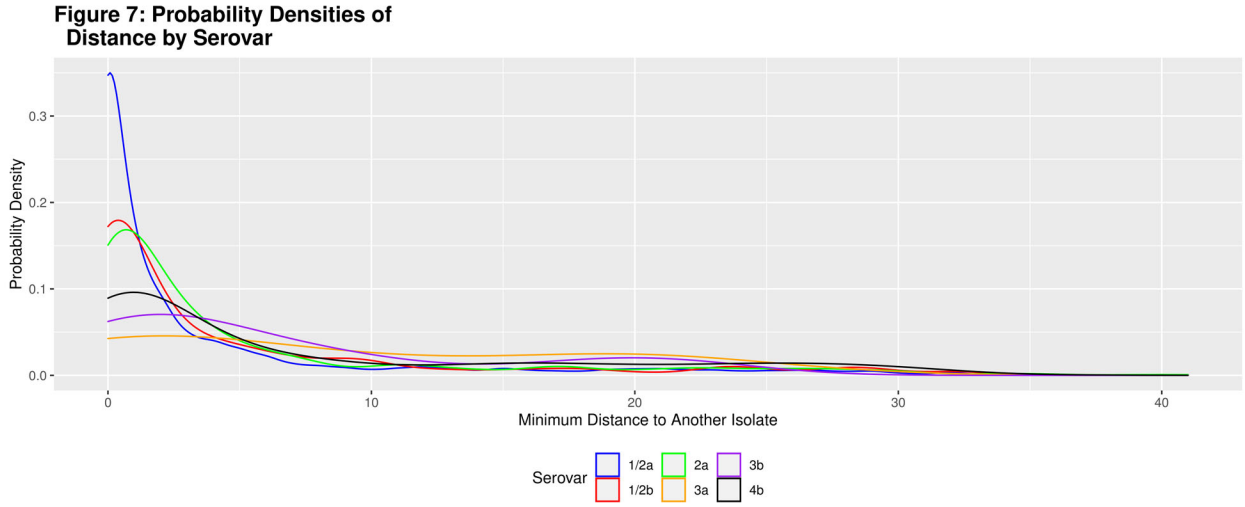
2020 and 2019 yielded greater spread when it came to contig counts (and thus had higher means), whereas the other years had narrower distributions and lower means overall. Unlike the previous half of the time interval, we notice that each year presents a distinguishable distribution, and even though some of the distributions are similar in structure (specifically 2016 through 2018), a clear difference in center and spread can be seen for all of the years to some degree. Furthermore, these visuals suggest that the association between contig count and minimum distance may not be as cut and dry as we had previously thought. 2020 had a one of the greatest levels of spread in terms of minimum distance, but it also has one of the greatest levels of spread in terms of contig count. This would suggest a potentially positive association between the two variables, whereas the source comparison suggested a negative association. In future modeling endeavors, we hope to gain a better understanding of these relationships through simple clustering models. Moreover, such relationships would not need to be explicitly determined to be fully incorporated into models with great degrees of statistical flexibility (e.g., nonparametric methods).

The final avenue that we want to pursue concerns the possible relationships between serovar and the numerical variables (contig count and minimum distance). Our analysis will be limited by the scarcity of available serovar information, but we hope to get at least some sense of the patterns and relationships between serovar type and the other variables, since serovar classification played a huge role in much of the literature we reviewed. We begin by examining how minimum distance is related to serovar type. Table 4 shows the (*approximate*) counts of the most common serovars of interest to us (keeping in mind that there are over 50000 data points, which further emphasizes the missingness of this variable.) These approximations were taken from the top 30 serovars in the dataset, for which there was noticeable overlap (for instance, serovar 2a was denoted by "2a" and "IIa"), so we combined any immediate overlapping data. We recall that 1/2a, 1/2b, 3a, 3b, and 4b were presented as particular strains of concern (Borcan et al. 2014), so we will narrow our focus down by examining these 5 serovars, along with serovar 2a due to its significantly larger presence in the dataset.

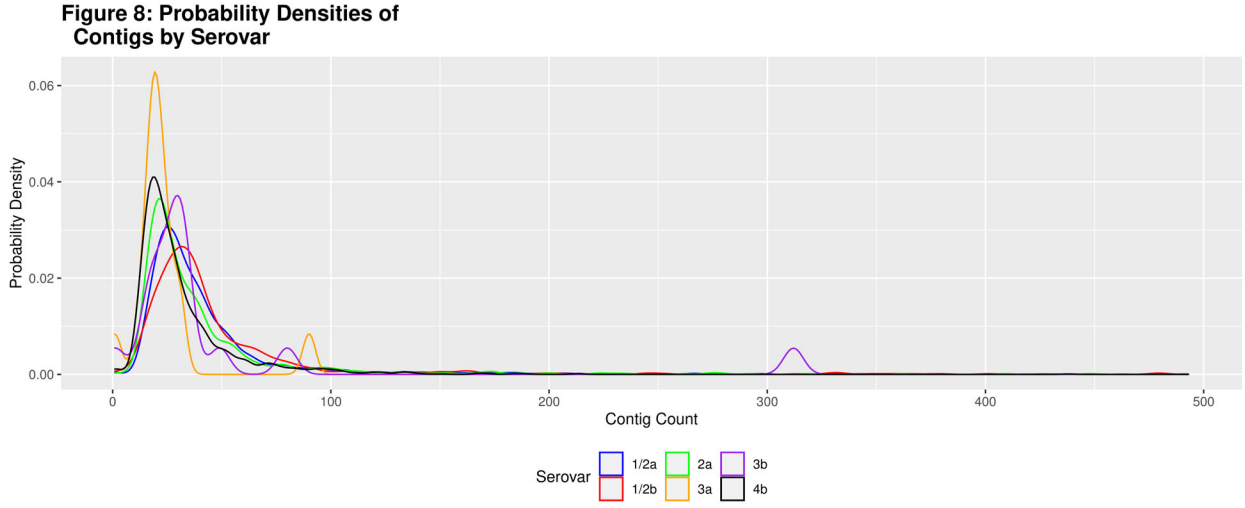
Table 4: Approximate Counts of Most Common Serovars

Serovar	Approximate Count	Approximate Frequency (%)
1/2a	3493	6.5000558
1/2b	691	1.2858685
1/2c	431	0.8020395
2a	778	1.4477651
2b	136	0.2530798
2c	48	0.0893223
3a	16	0.0297741
3b	16	0.0297741
3c	3	0.0055826
4a	8	0.0148870
4b	1593	2.9643827
4c	42	0.0781570

We begin once more by considering the relationship between minimum distance to another isolate and serovar. Figure 7 below shows the densities of minimum distance as grouped by the six serovars we previously mentioned (the method of combining repetitive counts is identical and still an approximation). Despite the fact that we are only approximating the serovar counts for the whole dataset, and the fact that most of the serovar information is missing, the figure provides a good idea of the behavior and biological character of each strain as characterized by its serovar. More specifically, since serovar type has been more thoroughly investigated through (basically all) of the sources in our literature review (see references), we can provide a more detailed and contextually rooted analysis of these strains and the distributions as shown below. Furthermore, it will be helpful to compare these serovar patterns to the other patterns we have noticed among other variables, which can better shape our approach for grouping these observations together.



Our results were somewhat surprising, specifically in regard to the distributional properties of the minimum distances for the 4b serovar group. This distribution was noticeably flatter than other distributions (like the 1/2a, 1/2b, and 2a distributions), meaning that there is a greater chance of serovar 4b minimum distances being larger than minimum distances for other serovars. This observation is important, because it provides a point of conflict between two ways of identifying strain intensity. On one hand, if we use minimum distance between other strains as an indicator of intensity, which is not without limitations as we will discuss later, 4b would be less severe than 1/2a in this regard. Yet this conflicts with the clinical findings of the literature we have discussed (see for example (Borcan et al. 2014)), which means that minimum distance measurements may not be a good way to classify intensity. Furthermore, other distributions exhibit greater degrees of spread, and consequently, much larger mean minimum distances (like for instance the distribution of 3a), but this could very well be due to the limitations of our approximations and the data itself (specifically since we are only using 16 data points to construct the 3a distribution.) As a result, the serovar grouping distributions for minimum distance have shown that pre-existing knowledge of severity by serovar does not align in this data set and the ways we have processed that data with smaller minimum distances.



Analogously, we can examine the difference in contig distributions by serovar count to assess their relationship. Figure 8 above shows the densities of contigs grouped by the prominent serovars of interest. While the distributions still look *somewhat F* distributed, there are many more anomalies and interesting occurrences to discuss here. First, we must address the odd behavior of the serovar 3a and 3b distributions. Likely due to a lack of proper smoothing a result of a small sample, we see that the 3b distribution shows a small peak of

probabilities right past the 300 contig mark. A similar small peak exists right before the 100 contig mark for the 3a distribution, and both of these anomalies are a noticeable distance from the centers of their respective distributions. While the small sample sizes we are working with (both 16) clearly contribute to this, it is important to note that such radical outliers exist in small samples, even under the missingness of the data. Setting these distributions aside, we consider the serovar distributions with larger samples. The 4b serovar distribution has the lowest near-zero peak of all the non-small sample distributions, which would tie the 4b serovar samples to lower contig counts overall. The 2a, 1/2a, and 1/2b distributions demonstrate larger contig counts (increasing in that order) overall by comparison. Furthermore, all large sample estimated density curves do not show any anomalous behavior from lac of smoothing, obviously due to having large enough sample sizes. Interestingly enough, the densities would imply that strain intensity may be associated with number of contigs, since serovar 4b had the narrowest large sample contig distribution centered near zero. However, we previously speculated that *higher* number of contigs led to *more severe* strains, which appears to be defied by this distribution. Still, it is important to note that a great deal of the serovar data (about 81%) is missing, which could have a huge impact on the densities above.

Finally, we want to address the impact of location on serovar type, since the broader location of an isolate plays into the type of source from which this isolate was extracted in an indirect sense. Like many other variables, the naming of variables in this column is spurious and difficult to work with. Some entries specify country and state or country and province, but the most popular entries simply specify country alone. To provide a rough analysis of this data, we examined the top 20 most frequent kind of data entries for location and took the first 10 countries in this list (there were several repeats for the United States, due to the specification of specific states). These countries were the U.S., U.K., Canada, Germany, France, Italy, Norway, Australia, New Zealand, and the Netherlands. The Netherlands had only one case of an identifying serovar (3a), the U.K. had 13 cases (5 1/2a and 8 4b), Canada only had 4 4b cases, and Norway, Australia, and New Zealand (through our rough estimation) had no identifying serovars. Thus, the previous 6 countries are excluded from Table 5 below, which gives serovar counts for the remaining countries.

Table 5: (Very Rough) Approximate Counts of Most Common Serovars by Country

Country	1/2a	1/2b	2a	3a	3b	4b
U.S.	493	209	0	6	11	806
Germany	4	8	564	0	0	556
France	0	0	126	0	0	38
Italy	1	5	77	0	0	22

This gives us a basic idea of the serovar presence across five different countries, though the numbers presented here are *exceedingly rough* approximations of the true counts. Specifically, the U.S. numbers are most definitely underestimated since all of the state specific entries have not been processed (due to time constraints and technical limitations). However, we see that while serovar 4b has played a dominant role in the data for the U.S. and Germany, serovar 2a seems to be more present in France and Italy, with a significant presence in Germany as well. Hence, we see that the type of serovar and its frequency among the different locations in the data differs greatly overall, which would be an important consideration to make when performing any kind of modeling analysis on this data set. It may be best to simply focus on one specific location and address the propagation of listeria therein.

Naturally, the clinical effects of given serovar presences in the different countries vary greatly from place to place and are influenced by external variables (food service legislature, variable reporting methods, clinical practices etc.). The severity of certain strains as classified by their serovars has been an important focus in this exploration, and clearly, 4b plays a huge role in severity and listeria's propagation as a whole, particularly in the U.S. and Germany. However, other strains mentioned in the literature (like (Borcan et al. 2014)) such as 1/2a and 2a seem to also play a prominent role in the movement and severity of listeria. An efficient way to classify a strain as "severe" or not might be to use a serovar grouping method, but as we have seen, this might not always capture the true intended effect.

Discussion of Limitations and Brief Conclusion

Both our analysis thus far and the dataset as a whole are subject to a vast array of substantial, severe, and often erroneous limitations. We will begin by addressing the dataset, then move to what we have done. This dataset possess such an incomprehensible naming system that a complete synthesis of all overlapping data is virtually impossible in efficient time. Whether it is the naming of serovars, locations, or sources, the dataset exhibits intense fluctuations in name specificity and reliable consistency which makes assessing true data patterns intensely frustrating. For instance, singular isolate sources are occasionally listed in great detail (e.g. "ice cream"), while the majority of sources are listed vaguely (i.e. "food"), and to systematically combine all specific sources into one general category would be horrendously inefficient. This vexation came up numerous times throughout our analysis, like for example in the last section, where we intensely underestimated the number of specific serovar occurrences in the United States (and other countries) simply because we would have had to sort through all of the specific state entries one by one to get a closer aggregation. If there is anything the unreliable naming of this dataset says, it is that the only way to provide a close to accurate analysis in to focus in on a narrow subsection of the dataset and analyze that specific portion. Thinking ahead to our future analysis, we will almost surely focus our attention on the United States alone, so that we only need to synthesize repetitive data for (at most) fifty states, as opposed to repeating the process for all provinces and territorial subdivisions internationally.

The next biggest issue with this dataset is the profound degree of missingness that exists for several important variables but specifically serovar type. The literature we have reviewed strongly suggests that serovar type is good indicator of intensity, but with over 81% of this information missing, we cannot reliably use this variable in a complete dataset analysis. We could focus our efforts on the subset consisting of the remaining 19% of the data, and perhaps we will, but we will be ignoring over four fifths of the data available to us. Imputation and missing data methods would likely be challenging in this case as well, given the extent of missingness, but we are considering using propensity scoring methods to amend this issue if necessary. Alternative methods of determining severity were also considered (like minimum distance and contig counts), but the potentially instrumental relationship between these variables and severity (by way of serovar) is not precisely clear.

Finally, our exploration may have shown us a lot about the data, but it is extremely limited in several cases. We have already mentioned that several estimates (shown in tables 2, 3, 4, *and especially* 5) are likely inaccurate, and in some cases grossly inaccurate, due to the difficulty we had combining variables that were repetitive. Some of our distributions were constructed using very small sample sizes (like in the cases of serovars 3a and 3b), which may have yielded inaccurate results due to lack of smoothing. Our conflation of the two original distance variables resulted in the loss of some information (specifically whether or not the isolate within distance is of the same strain), but this did not hinder the results of our analysis much. Furthermore, we omitted a wide variety of variables due to intense missingness which skewed the impact of our results and limits the possibility of using these variables as covariates in future models.

Collectively, we have learned a great deal about this dataset and the work that has been done on the topic of the propagation of listeria. The literature has focused on a wide variety of approaches to dealing with listeria, but most of these approaches focus on a small number of sources or face experimental hindrance. ((Borcan et al. 2014) (Tonner et al. 2020)) Serovar 4b seems to be the center of most attention in the literature, proving to be a serovar indicative of medical concern. (Muchaamba et al. 2021) Our analysis has revealed interesting trends between this serovar and minimum distance and contig count. As a result, it's use as a reference point of severity will be considered when we engage in future modeling. Different years play an important part in the fluctuation of certain numerical covariate (minimum distance and contig) count as shown in our visuals, and similar trends have been shown for contigs as well. The effect of missingness on sources is also evident as shown in Table 1, and we feel that while this dataset is most certainly not missing completely at random, it is arguable that the data could be missing at random, as opposed to missing not at random. Clearly, the types of serovars and their influence vary by country as well, though our estimates of this phenomenon are rough at best. Our goal moving forward will be to focus our attention on the U.S. alone and attempt to engage in a nonparametric Bayesian clustering approach along with a standard clustering approach, whereby we organize the strains (as given by serovar or another variable potentially) by sources. We will only consider the most prominent sources and strains in our analysis, but organize and examine the data in a more robust and less erroneous manner going forward.

References

Borcan, A. M. et al. 2014. "Listeria Monocytogenes – Characterization of Strains Isolated from Clinical Severe Cases." *Journal of Medicine and Life* 7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4391364/.

Deligentte-Muller, M. L. et al. 2006. "Use of Bayesian Modelling in Risk Assessment: Application to Growth of Listeria Monocytogenes and Food Flora in Cold-Smoked Salmon." *International Journal of Food Microbiology* 106. https://www.sciencedirect.com/science/article/pii/S0168160505004332?via/%3Dihub.

Luth, S. et al. 2020. "Backtracking and Forward Checking of Human Listeriosis Clusters Identified a Multiclonal Outbreak Linked to Listeria Monocytogenes in Meat Products of a Single Producer." *Emerging Microbes and Infections* 9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7473094/.

Muchaamba, F. et al. 2021. "Different Shades of Listeria Monocytogenes: Strain, Serotype, and Lineage-Based Variability in Virulence and Stress Tolerance Profiles." *Frontiers in Microbiology* 12. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8764371/.

NLM:NCBI. 2022. "Pathogen Detection Listeria Monocytogenes."

Pouillot, R. et al. 2003. "Estimation of Uncertainty and Variability in Bacterial Growth Using Bayesian Inference. Application to Listeria Monocytogenes." *International Journal of Food Microbiology* 81. https://www.sciencedirect.com/science/article/pii/S0168160502001927?via/%3Dihub/#BIB46.

Psareva, E. et al. 2021. "Diversity of Listeria Monocytogenes Strains Isolated from Food Products in the Central European Part of Russia in 2000–2005 and 2019–2020." *Foods* 10. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8617672/.

Rogalla, D., and P. Bomar. 2022. "Listeria Monocytogenes." https://www.ncbi.nlm.nih.gov/books/NBK534838/.

Tonner, P. et al. 2020. "A Bayesian Non-Parametric Mixed-Effects Model of Microbial Growth Curves." *PLOS Computational Biology* 16. https://journals.plos.org/ploscompbiol/article/citation?id=10.1371/journal.pcbi.1008366.

Wang, H. et al. 2021. "Machine Learning Prediction of Foodborne Disease Pathogens: Algorithm Development and Validation Study." *JMIR Medical Informatics* 9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7872834//#.

Ward, T. et al. 2004. "Intraspecific Phylogeny and Lineage Group Identification Based on the prfA Virulence Gene Cluster of Listeria Monocytogenes." *Journal of Bacteriology* 186. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC451661/.

Link to Repository: https://github.com/NTProvost/PHP-2550-Project

Code Appendix

```
#Libraries
library(tidyverse)
library(ggplot2)
library(tableone)
library(table1)
library(knitr)
library(IRdisplay)
library(kableExtra)

#Removes Unneeded Variables
#Load in Data
df<-subset(df,select=-Create.date)
df<-subset(df,select=-Organism.group)

#Evaluates Missing Data
df.matrix<-as.matrix(df)
df.row<-dim(df.matrix)[1]
df.col<-dim(df.matrix)[2]
percent.missing<-rep(0,df.col)
for(k in 1:df.col){
  NA.count<-sum(is.na(df.matrix[,k]))
  empty.count<-sum(df.matrix[,k]=="")
  if(is.na(empty.count)==T){empty.count<-0}
  missing.count<-NA.count+empty.count
  missing.percent<-missing.count/df.row
  percent.missing[k]<-missing.percent
}
Missing.Table<-data.frame(Variable=names(df),
  Percent_Missing=percent.missing*100)
print(Missing.Table)

#Labels Sporadic Cases
for(k in 1:nrow(df)){
  if(df$Outbreak[k]==""){
    df$Outbreak[k]<-"Sporadic"
  }
}

#Conflates Distance Variables
min.dist<-rep(0,nrow(df))
for(k in 1:nrow(df)){
  if(is.na(df$Min.same[k])==T & is.na(df$Min.diff[k])==T){
    min.dist[k]<-NA
  }
  if(is.na(df$Min.same[k])==F & is.na(df$Min.diff[k])==T){
    min.dist[k]<-df$Min.same[k]
  }
  if(is.na(df$Min.same[k])==T & is.na(df$Min.diff[k])==F){
    min.dist[k]<-df$Min.diff[k]
  }
  if(is.na(df$Min.same[k])==F & is.na(df$Min.diff[k])==F){
    min.dist[k]<-min(df$Min.same[k],df$Min.diff[k])
```

```
  }
}
df$min.dist<-min.dist

#Conflates Source Variables and Removes Unneeded Variables
for(k in 1:nrow(df)){
  if(df$Isolation.source[k]=="" & df$Isolation.type[k]!=""){
    df$Isolation.source[k]<-df$Isolation.type[k]
  }
}
df.adj<-subset(df,select=-c(Min.same,Min.diff,Isolation.type))

#Separates Data into Subsets with/without Distance Data
#Using a Distance Indicator
dst<-rep(0,nrow(df.adj))
for(k in 1:nrow(df.adj)){
  if(is.na(df.adj$min.dist[k])==F){
    dst[k]<-1
  }
  if(is.na(df.adj$min.dist[k])==T){
    dst[k]<-0
  }
}
df.adj$dst<-dst
df.dist<-df.adj%>%filter(dst==1)
df.nodist<-df.adj%>%filter(dst==0)

#Computes Top 10 Isolates for Each Subset as a Table
dist.sources<-as.data.frame(tail(sort(table(df.dist$Isolation.source)),10))
nodist.sources<-as.data.frame(tail(sort(table(df.nodist$Isolation.source)),10))
names(dist.sources)<-c("Source (Distance)","Number of Occurances")
names(nodist.sources)<-c("Source (No Distance)","Number of Occurances")
nodist.sources$`Source (No Distance)`[10]<-NA

#Render Table 1 in Latex
kable(list(dist.sources,nodist.sources),
  "latex",caption="Most Frequent Sources for Each Subset")%>%
  kable_styling(latex_options="HOLD_position")

#Computes Frequency of Sources by Distance Subsets
#Numbers had to be inserted manually from df.dist and df.nodist
dist.compare<-data.frame(Source=c("Human","Blood",
  "Environmental","Food"),
  Frequency.With.Distance=c(941+1670,2102,2257+4503+1432,3984)/nrow(df.dist),
  Frequency.Without.Distance=c(295+319,980,169+218+106,1056)/nrow(df.nodist))

#Subsets the Data by Year
df10<-df.adj%>%filter(Collection.date=="2010")
df11<-df.adj%>%filter(Collection.date=="2011")
df12<-df.adj%>%filter(Collection.date=="2012")
df13<-df.adj%>%filter(Collection.date=="2013")
df14<-df.adj%>%filter(Collection.date=="2014")
df15<-df.adj%>%filter(Collection.date=="2015")
```

```
df16<-df.adj%>%filter(Collection.date=="2016")
df17<-df.adj%>%filter(Collection.date=="2017")
df18<-df.adj%>%filter(Collection.date=="2018")
df19<-df.adj%>%filter(Collection.date=="2019")
df20<-df.adj%>%filter(Collection.date=="2020")

#Figure 1: Densities of Distance 2010 to 2015
ggplot()+geom_density(aes(x=min.dist,color="2010"),
  data=df10,na.rm=T)+
  geom_density(aes(x=min.dist,color="2011"),
  data=df11,na.rm=T)+
  geom_density(aes(x=min.dist,color="2012"),
  data=df12,na.rm=T)+
  geom_density(aes(x=min.dist,color="2013"),
  data=df13,na.rm=T)+
  geom_density(aes(x=min.dist,color="2014"),
  data=df14,na.rm=T)+
  geom_density(aes(x=min.dist,color="2015"),
  data=df15,na.rm=T)+
  scale_color_manual(
  values=c("blue","red","green","orange","purple","black"),
  name="Year")+theme(legend.position="bottom",
  legend.text=element_text(size=7),
  legend.key.size=unit(0.1,"cm"),
  plot.title=element_text(size=10))+
  labs(x="Minimum Distance to Another Isolate",
  y="Probability Density",title="Figure 1: Probability Densities of
  Minimum Distances (2010 - 2015)")

#Figure 2: Densities of Distance 2016 to 2020
ggplot()+geom_density(aes(x=min.dist,color="2016"),
  data=df16,na.rm=T)+
  geom_density(aes(x=min.dist,color="2017"),
  data=df17,na.rm=T)+
  geom_density(aes(x=min.dist,color="2018"),
  data=df18,na.rm=T)+
  geom_density(aes(x=min.dist,color="2019"),
  data=df19,na.rm=T)+
  geom_density(aes(x=min.dist,color="2020"),
  data=df20,na.rm=T)+
  scale_color_manual(
  values=c("blue","red","green","orange","purple"),
  name="Year")+theme(legend.position="bottom",
  legend.text=element_text(size=7),
  legend.key.size =unit(0.1,"cm"),
  plot.title=element_text(size=10))+
  labs(x="Minimum Distance to Another Isolate",
  y="Probability Density",title="Figure 2: Probability Densities of
  Minimum Distances (2016-2020)")

#Source Subsets
dfenv<-df.adj%>%
  filter(Isolation.source=="environmental" |
```

```
  Isolation.source=="environmental swab sponge" |
  Isolation.source=="environmental swab")
dffood<-df.adj%>%
  filter(Isolation.source=="food" |
  Isolation.source=="Food Processing")
dfhuman<-df.adj%>%
  filter(Isolation.source=="human" |
  Isolation.source=="human listeriosis")
dfclin<-df.adj%>%filter(Isolation.source=="clinical")
dfwat<-df.adj%>%filter(Isolation.source=="water sediment")
dfblood<-df.adj%>%filter(Isolation.source=="blood")

#Table 2: Contig Descriptives
t2<-data.frame(x=c("Environmental","Food","Human",
  "Clinical","Water","Blood"),y=c(mean(dfenv$min.dist,na.rm=T),
  mean(dfblood$min.dist,na.rm=T),mean(dfhuman$min.dist,na.rm=T),
  mean(dfclin$min.dist,na.rm=T),mean(dfwat$min.dist,na.rm=T),
  mean(dfblood$min.dist,na.rm=T)),z=c(sd(dfenv$min.dist,na.rm=T),
  sd(dfblood$min.dist,na.rm=T),sd(dfhuman$min.dist,na.rm=T),
  sd(dfclin$min.dist,na.rm=T),sd(dfwat$min.dist,na.rm=T),
  sd(dfblood$min.dist,na.rm=T)))
names(t2)<-c("Source","Mean","Standard Deviation")
kable(t2, "latex",
 caption="Descriptive Statistics of Minimum Distances by Source")%>%
 kable_styling(latex_options="HOLD_position")

#Figure 3: Densities by Source
ggplot()+geom_density(aes(x=min.dist,color="Environmental"),
  data=dfenv,na.rm=T)+
  geom_density(aes(x=min.dist,color="Food"),
  data=dffood,na.rm=T)+
  geom_density(aes(x=min.dist,color="Blood"),
  data=dfblood,na.rm=T)+
  geom_density(aes(x=min.dist,color="Human"),
  data=dfhuman,na.rm=T)+
  geom_density(aes(x=min.dist,color="Clinical"),
  data=dfclin,na.rm=T)+
  geom_density(aes(x=min.dist,color="Water"),
  data=dfwat,na.rm=T)+
  scale_color_manual(
  values=c("blue","red","green","orange",
  "purple","black"),
  name="Year")+theme(legend.position="bottom",
  plot.title=element_text(size=15))+
  labs(x="Minimum Distance to Another Isolate",
  y="Probability Density",title="Figure 3: Probability Densities of
  Minimum Distances by Source")

#Outbreak/Sporadic Subsets and T Test
df.out<-df.adj%>%filter(Outbreak!="Sporadic")
df.spor<-df.adj%>%filter(Outbreak=="Sporadic")
t.test(df.out$min.dist,df.spor$min.dist)
```

```
#Table 3: Contig Descriptives
t3<-data.frame(x=c("Environmental","Food","Human",
  "Clinical","Water","Blood"),y=c(mean(dfenv$Contigs),
  mean(dfblood$Contigs),mean(dfhuman$Contigs),
  mean(dfclin$Contigs),mean(dfwat$Contigs),
  mean(dfblood$Contigs)),z=c(sd(dfenv$Contigs),
  sd(dfblood$Contigs),sd(dfhuman$Contigs),
  sd(dfclin$Contigs),sd(dfwat$Contigs),
  sd(dfblood$Contigs)))
names(t3)<-c("Source","Mean","Standard Deviation")
kable(t3, "latex",
 caption="Descriptive Statistics of Contigs by Source")%>%
 kable_styling(latex_options="HOLD_position")

#Figure 4: Densities by Source
ggplot()+geom_density(aes(x=Contigs,color="Environmental"),
  data=dfenv,na.rm=T)+
  geom_density(aes(x=Contigs,color="Food"),
  data=dffood,na.rm=T)+
  geom_density(aes(x=Contigs,color="Blood"),
  data=dfblood,na.rm=T)+
  geom_density(aes(x=Contigs,color="Human"),
  data=dfhuman,na.rm=T)+
  geom_density(aes(x=Contigs,color="Clinical"),
  data=dfclin,na.rm=T)+
  geom_density(aes(x=Contigs,color="Water"),
  data=dfwat,na.rm=T)+
  scale_color_manual(
  values=c("blue","red","green","orange",
  "purple","black"),
  name="Year")+theme(legend.position="bottom",
  plot.title=element_text(size=15))+
  labs(x="Minimum Distance to Another Isolate",
  y="Probability Density",title="Figure 4: Probability Densities of
 Contig Counts by Source")

#Figure 5: Densities of Contigs 2010 to 2015
ggplot()+geom_density(aes(x=Contigs,color="2010"),
  data=df10,na.rm=T)+
  geom_density(aes(x=Contigs,color="2011"),
  data=df11,na.rm=T)+
  geom_density(aes(x=Contigs,color="2012"),
  data=df12,na.rm=T)+
  geom_density(aes(x=Contigs,color="2013"),
  data=df13,na.rm=T)+
  geom_density(aes(x=Contigs,color="2014"),
  data=df14,na.rm=T)+
  geom_density(aes(x=Contigs,color="2015"),
  data=df15,na.rm=T)+
  scale_color_manual(
  values=c("blue","red","green","orange","purple","black"),
  name="Year")+theme(legend.position="bottom",
  legend.key.size =unit(0.1,"cm"),
```

```
  legend.text=element_text(size=7),
  plot.title=element_text(size=10))+
  labs(x="Contig Count",
       y="Probability Density",title="Figure 5: Probability Densities of
  Contigs by Year (2010 - 2015)")

#Figure 6: Densities of Contigs 2016 to 2020
ggplot()+geom_density(aes(x=Contigs,color="2016"),
  data=df16,na.rm=T)+
  geom_density(aes(x=Contigs,color="2017"),
  data=df17,na.rm=T)+
  geom_density(aes(x=Contigs,color="2018"),
  data=df18,na.rm=T)+
  geom_density(aes(x=Contigs,color="2019"),
  data=df19,na.rm=T)+
  geom_density(aes(x=Contigs,color="2020"),
  data=df20,na.rm=T)+
  scale_color_manual(
  values=c("blue","red","green","orange","purple"),
  name="Year")+theme(legend.position="bottom",
  legend.key.size =unit(0.1,"cm"),
  legend.text=element_text(size=7),
  plot.title=element_text(size=10))+
  labs(x="Contig Count",
  y="Probability Density",title="Figure 6: Probability Densities of
  Contigs by Year (2016 - 2020)")

#Table of Most Common Serovars (Some Numbers were inserted manual from the
#command below) (Table 4)
#head(sort(table(df.adj$Serovar),decreasing=T),30)
t.sero<-data.frame(Serovar=c("1/2a","1/2b","1/2c","2a","2b","2c","3a","3b","3c",
  "4a","4b","4c"),ACount=c(2846+647,589+102,379+52,602+176,84+52,48,16,16,3,
  8,1136+440+17,42),AFreq=c(2846+647,589+102,379+52,602+176,84+52,48,16,16,3,
  8,1136+440+17,42)*100/nrow(df.adj))
names(t.sero)<-c("Serovar","Approximate Count","Approximate Frequency (%)")
kable(t.sero, "latex",
 caption="Approximate Counts of Most Common Serovars")%>%
 kable_styling(latex_options="HOLD_position")

#Serovar Subsets
df1.2a<-df.adj%>%filter(Serovar=="1/2a" | Serovar=="Serotype 1/2a")
df1.2b<-df.adj%>%filter(Serovar=="1/2b" | Serovar=="Serotype 1/2b")
df2a<-df.adj%>%filter(Serovar=="2a" | Serovar=="IIa")
df3a<-df.adj%>%filter(Serovar=="3a")
df3b<-df.adj%>%filter(Serovar=="3b")
df4b<-df.adj%>%filter(Serovar=="4b" | Serovar=="IVb" | Serovar=="Serotype 4b")

#Figure 7: Densities of Distance by Serovar
ggplot()+geom_density(aes(x=min.dist,color="1/2a"),
  data=df1.2a,na.rm=T)+
  geom_density(aes(x=min.dist,color="1/2b"),
  data=df1.2b,na.rm=T)+
  geom_density(aes(x=min.dist,color="2a"),
```

```
  data=df2a,na.rm=T)+
  geom_density(aes(x=min.dist,color="3a"),
  data=df3a,na.rm=T)+
  geom_density(aes(x=min.dist,color="3b"),
  data=df3b,na.rm=T)+
  geom_density(aes(x=min.dist,color="4b"),
  data=df4b,na.rm=T)+
  scale_color_manual(
  values=c("blue","red","green","orange","purple","black"),
  name="Serovar")+theme(legend.position="bottom",
  plot.title=element_text(face="bold",size=15))+
  labs(x="Minimum Distance to Another Isolate",
  y="Probability Density",title="Figure 7: Probability Densities of
  Distance by Serovar")

#Figure 8: Densities of Contigs by Serovar
ggplot()+geom_density(aes(x=Contigs,color="1/2a"),
  data=df1.2a,na.rm=T)+
  geom_density(aes(x=Contigs,color="1/2b"),
  data=df1.2b,na.rm=T)+
  geom_density(aes(x=Contigs,color="2a"),
  data=df2a,na.rm=T)+
  geom_density(aes(x=Contigs,color="3a"),
  data=df3a,na.rm=T)+
  geom_density(aes(x=Contigs,color="3b"),
  data=df3b,na.rm=T)+
  geom_density(aes(x=Contigs,color="4b"),
  data=df4b,na.rm=T)+
  scale_color_manual(
  values=c("blue","red","green","orange","purple","black"),
  name="Serovar")+theme(legend.position="bottom",
  plot.title=element_text(face="bold",size=15))+
  labs(x="Contig Count",
  y="Probability Density",title="Figure 8: Probability Densities of
  Contigs by Serovar")

#Table of Most Serovars by Location
#tail(sort(table(df.adj$Location),descending=T),20)
t.loc<-data.frame(Country=c("U.S.","Germany","France","Italy"),
  halfa=c(sum(df1.2a$Location=="USA"),sum(df1.2a$Location=="Germany"),
  sum(df1.2a$Location=="France"),sum(df1.2a$Location=="Italy")),
  halfb=c(sum(df1.2b$Location=="USA"),sum(df1.2b$Location=="Germany"),
  sum(df1.2b$Location=="France"),sum(df1.2b$Location=="Italy")),
  twoa=c(sum(df2a$Location=="USA"),sum(df2a$Location=="Germany"),
  sum(df2a$Location=="France"),sum(df2a$Location=="Italy")),
  threea=c(sum(df3a$Location=="USA"),sum(df3a$Location=="Germany"),
  sum(df3a$Location=="France"),sum(df3a$Location=="Italy")),
  threeb=c(sum(df3b$Location=="USA"),sum(df3b$Location=="Germany"),
  sum(df3b$Location=="France"),sum(df3b$Location=="Italy")),
  fourb=c(sum(df4b$Location=="USA"),sum(df4b$Location=="Germany"),
  sum(df4b$Location=="France"),sum(df4b$Location=="Italy"))
  )
```

```
#Renders Table 5
names(t.loc)<-c("Country","1/2a","1/2b","2a","3a","3b","4b")
kable(t.loc, "latex",
 caption="(Very Rough) Approximate Counts of Most Common Serovars by Country")%>%
 kable_styling(latex_options="HOLD_position")
```