Predicting Drug Resistance and Susceptibility in Foodborne Illnesses with Random Forests

Breanna Richards, Yanru Liao, Jina Yang

Abstract

In recent years, the excessive incorrect use of antibiotics for the treatment of infectious diseases in humans and animals has become a major area of concern for human health. When this behavior of misusage persists, bacteria are not killed but instead develop survival traits, or resistance, against these treatments, which they can pass on to even more bacteria. This often results in increased risks of severe infections, illnesses, and even death. The designation of effective treatment for these infections requires learning of organisms' susceptibility to various antibiotics, though the traditional clinical processes to do so are often time-consuming. High-throughput sequencing technology based on genomic data of bacteria has the potential to fast guide this decision-making process. Using data from the National Center for Biotechnology Information's (NCBI) Pathogen database, this study utilized binary random forest classification for antibiotic responses of Salmonella enterica, E.coli and Shigella, and Campylobacter jejuni isolates in consideration of two cases of data availability 1) source and genetic predictors and 2) source predictors only (to capture cases where genomic information is unavailable) to predict pathogens' susceptibility or resistance to prominent antibiotics. Observed patterns in these data were numerically assessed using multivariable logistic regression. Additionally, a system was developed to receive genetic and baseline information from new cases, to then return the most probable antibiotic those cases would garner susceptible responses to. Overall, both the models containing source and genetic predictors, and those only considering source predictors saw favorable performance with average sensitivity values of 0.98 and 0.79 respectively. Collection date, serovar, and isolation source variables were consistently integral in prediction. Upon utilizing the recommendation system on the testing data for the three species of interest, recommendations for E.coli and Shigella, and Campylobacter jejuni isolates saw the most agreement between the source and genetic predictor models and the source predictor only models. Prediction in Salmonella isolates were most dependent upon information about the presence/absence of antimicrobial resistant genes. The accurate and confident prescription of these drugs can help to target outbreaks at its source to quell the continued spread of disease without the risk of the unfavorable genetic outcomes that can result from mistreatment.

Introduction

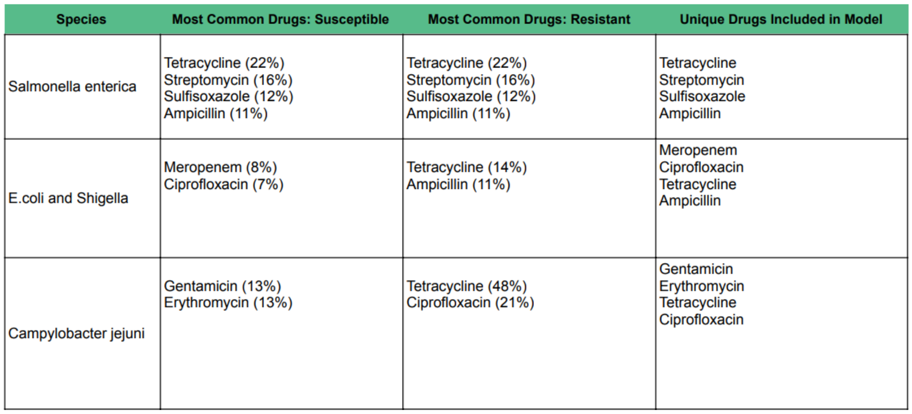
Here are two sample references: Feynman and Vernon Jr. (1963; Dirac 1953).

Methods

Results

A total of 16,049 isolates with antibiotic information (8,672 observations on Salmonella, 3,880 on E.coli, and 3,497 on the Campylobacter species) were retained and utilized for the purposes of this analysis. Data were separated by species and within each species, random forest models predicted susceptible/resistant outcomes on the most commonly recorded antibiotics recorded in **Table 1**.

Table 1: Antibiotics Considered in Modeling



Species	Most Common Drugs: Susceptible	Most Common Drugs: Resistant	Unique Drugs Included in Model
Salmonella enterica	Tetracycline (22%) Streptomycin (16%) Sulfisoxazole (12%) Ampicillin (11%)	Tetracycline (22%) Streptomycin (16%) Sulfisoxazole (12%) Ampicillin (11%)	Tetracycline Streptomycin Sulfisoxazole Ampicillin
E.coli and Shigella	Meropenem (8%) Ciprofloxacin (7%)	Tetracycline (14%) Ampicillin (11%)	Meropenem Ciprofloxacin Tetracycline Ampicillin
Campylobacter jejuni	Gentamicin (13%) Erythromycin (13%)	Tetracycline (48%) Ciprofloxacin (21%)	Gentamicin Erythromycin Tetracycline Ciprofloxacin

Data for each species were split 80-20 into a training set used to build that species' associated models, and a testing set in order to assess the performance of those models on data not directly used in the training process.

As aforementioned, in order to reduce the dimensionality of our data, we utilized the Boruta algorithm on our training data before proceeding with applications to random forests (citation). Variable selection will only be performed for the models containing both source and genetic predictors.

In order to optimize the performance of our models, we used 10-fold cross validation (CV), a process in which our training data is split into 10 subsets and the model is iteratively fit 10 times, each time on 9 (k-1) of the folds with performance being evaluated on the k^{th} fold - a different subset of the data each time. Many iterations of the 10-fold CV process is performed, each with a different combination of our hyper-parameters of interest. For the purposes of this study, we tested different values for the number of trees (`ntree`) contributing to the overall prediction decisions, and the number of randomly sampled variables to be considered at each split (`mtry`) (cite). The combination that produced the lowest out-of-bag error, a metric used to measure the prediction error on data not utilized in a given bootstrap sample, will be the values used in the final random forest models.

Set $m =$ number of predictors in the data. We tested three common values for mtry: $\sqrt{m}$, $m/3$ (rounded down to the nearest integer), and $m/2$ (rounded down to the nearest integer) (cite). For number of trees, we tested values 250, 500, and 1,000. In these processes, we want enough trees to stabilize the error but not so many that the ensemble becomes correlated and causes the model to be overfit.

To account for any potential class imbalances within each model, we weight classes by $\frac{1}{prop\ of\ samples\ in\ class}$.

Variable Selection

Recall that we've identified the following covariates to be considered when predicting drug response for Salmonella enterica isolates: collection date, isolation type, minimum SNP distance to an isolate of the same isolation type, minimum SNP distance to an isolate of a different isolation type, serovar, country, isolation source, number of close isolates, antigen formula, drug susceptibility similarity to other isolates within the same cluster, and the presence or absence of 12 antimicrobial resistant genes within a given isolate for a total of 22 predictors.

When predicting antibiotic response for E.coli and Campylobacter isolates we also consider collection date, isolation type, minimum SNP distance to an isolate of the same isolation type, minimum SNP distance to an isolate of a different isolation type, country, isolation source, number of close isolates, and drug susceptibility similarity to other isolates within the same cluster. Additionally, we consider a variable that numerically captures the drug *resistance* similarity to other isolates within the same cluster, but unlike in Salmonella, we don't consider serovar and antigen formula due to that information only being retained on Salmonella isolates.

Lastly, we consider the presence or absence of 13 antimicrobial resistant genes for a total of 22 predictors. The 13 antimicrobial resistant genes in consideration as predictors in these models are highlighted in **Table 1**.

For our models containing only source predictors, we considered only the following covariates, kin to what the average person may be more likely to know about their own foodborne illness: collection date, isolation type, serovar, country, and isolation source. We decided to include serovar information in the 'source only' predictors because in our data, serovar information is often reduced to the geographical origin of illness, which may be known. Serovar information that isn't tied to a geographical origin on the other hand may be discovered through a doctor's appointment.

Salmonella enterica None of the variables within the salmonella dataset were deemed unimportant enough to exclude in any of the four random forest models considering both source and genetic predictors, i.e. all predictors (outcome Tetracycline model, outcome streptomycin model, outcome Ampicillin model, and outcome Sulfisoxazole model). And so, during hyper-parameter tuning, we consider all predictors in the data.

E.coli and Shigella In predicting response to Meropenem, the presence/absence of 7 antimicrobial resistant genes were deemed either unimportant (6) or tentatively unimportant (1) so we decided to remove all 7 of those genes from being predictors in that model (bla__oxa__193__complete, x50s__122__a103v__point, gyr__a__t86i__point, bla__oxa__complete, tet__o__complete, aph__3__ii__ia__complete, and bla__ec__complete).

In predicting antibiotic response to tetracycline in E.coli isolates, the presence/absence of 5 antimicrobial resistant genes were considered to be unimportant and were thus excluded in hyper-parameter tuning and final model runs for E.coli: aph__3__ii__ia__complete, bla__oxa__193__complete, gyr__a__t86i__point, tet__o__complete, x50s__122__a103v__point.

In predicting antibiotic response to Ampicillin and Ciprofloxacin, the same 5 attributes were confirmed unimportant.

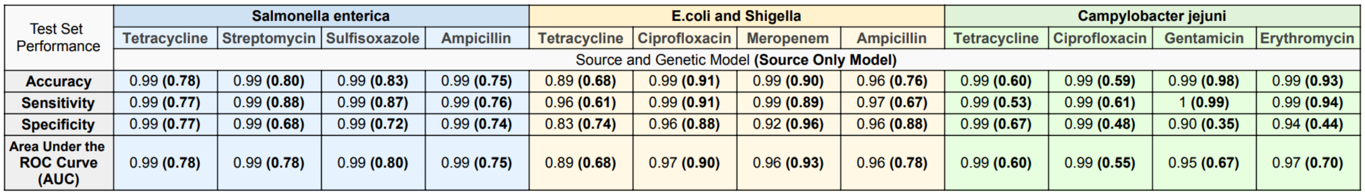
Campylobacter jejuni In predicting antibiotic response to tetracycline within Campylobacter isolates, seven antimicrobial resistant genes were deemed unimportant: acr__f__complete, aph__3__ib__complete, aph__6__id__complete, bla__ec__complete, mdt__m__complete, tet__a__complete, and sulf2__complete.

The same seven genes were deemed unimportant in predicting Gentamicin, Erythromycin, and Ciprofloxacin.

Test Set Performance

After hyper-parameter tuning (see 'Final Report' folder in GitHub repository), we assessed the performance of our models on testing sets for each species of interest (**Table 2**).

Table 2: Model Performance on Test Sets for Salmonella, E.coli, and Campylobacter



Test Set Performance	Salmonella enterica				E.coli and Shigella				Campylobacter jejuni			
	Tetracycline	Streptomycin	Sulfisoxazole	Ampicillin	Tetracycline	Ciprofloxacin	Meropenem	Ampicillin	Tetracycline	Ciprofloxacin	Gentamicin	Erythromycin
	Source and Genetic Model **(Source Only Model)**											
Accuracy	0.99 **(0.78)**	0.99 **(0.80)**	0.99 **(0.83)**	0.99 **(0.75)**	0.89 **(0.68)**	0.99 **(0.91)**	0.99 **(0.90)**	0.96 **(0.76)**	0.99 **(0.60)**	0.99 **(0.59)**	0.99 **(0.98)**	0.99 **(0.93)**
Sensitivity	0.99 **(0.77)**	0.99 **(0.88)**	0.99 **(0.87)**	0.99 **(0.76)**	0.96 **(0.61)**	0.99 **(0.91)**	0.99 **(0.89)**	0.97 **(0.67)**	0.99 **(0.53)**	0.99 **(0.61)**	1 **(0.99)**	0.99 **(0.94)**
Specificity	0.99 **(0.77)**	0.99 **(0.68)**	0.99 **(0.72)**	0.99 **(0.74)**	0.83 **(0.74)**	0.96 **(0.88)**	0.92 **(0.96)**	0.96 **(0.88)**	0.99 **(0.67)**	0.99 **(0.48)**	0.90 **(0.35)**	0.94 **(0.44)**
Area Under the ROC Curve (AUC)	0.99 **(0.78)**	0.99 **(0.78)**	0.99 **(0.80)**	0.99 **(0.75)**	0.89 **(0.68)**	0.97 **(0.90)**	0.96 **(0.93)**	0.96 **(0.78)**	0.99 **(0.60)**	0.99 **(0.55)**	0.95 **(0.67)**	0.97 **(0.70)**

Overall, our all models containing both source and genetic predictors performed exceptionally well on our test data. Most models including genetic predictors obtained accuracy, sensitivity, specificity, and area under the ROC curve values over 0.90, which shows great performance and utility. As the goal of the project was to create a recommendation system to apply to new cases of these species, we may want to prioritize sensitivity over accuracy or specificity since we want to reliably recommend an antibiotic that that new case would be susceptible to (susceptibility being defined as the positive '1' class in our models). Thus, using average sensitivity as a benchmark, our models for Campylobacter jejuni and Salmonella enterica perform the best with E.coli and Shigella not too far behind at an average of 0.96. Evidently, it seems that the models containing both source and genetic predictors may perform nearly perfectly, insinuating that there is information contained in these models that are highly correlated with the outcomes of antibiotic response.

As a result of this, we decided to investigate alternative models that included only 'source' predictors, i.e. reducing the predictor subspace down to just 5 variables in the salmonella models (serovar, collection date, isolation type, country, and isolation source) and 4 variables in both the E.coli and Campylobacter models (collection date, isolation type, country, and isolation source due to serovar information not being retained on E.coli and Campylobacter isolates). We also treat this scenario to be more aligned with the kind of information that we would except the average person to have about their case if they suspect they've contracted one of these foodborne illnesses, as well as the kind of information that would be easily accessible without having to go through the potentially tedious process of clinical testing to find genetic specific information. The question of interest then becomes: how accurate and useful can these models still be in the absence of less attainable genetic information? If we were to build a recommendation system for more public but safe use, can our models still be reliable to the average person?

The results of our 'source only' models are bold in **Table 2**. Notably, removing the genetically associated predictors from our model generally decreases performance, but despite that, these models still relatively perform well. Taking sensitivity as the primary metric of interest, the 'source only' models for Salmonella perform the best with an average sensitivity of 0.82, followed by the models for E.coli with an average sensitivity of 0.77, and lastly our models for Campylobacter with an average sensitivity of 0.76. We observed and compared the variables most integral to predicting response in these antibiotics.

Variable Importance

Salmonella enterica

Figure 1 highlights the top 5 most important variables in all of the classification models for the Salmonella isolates, both for source and genetic predictors, and source predictors only. The variable importance results for all source and genetic predictor models are consistent in the fact that the top two most important variables are indicator variables for the presence or absence of antimicrobial resistant genes. More specifically, in predicting for susceptibility/resistance to tetracycline, the TET genes are most important to classification. The APH genes are most important to predicting response to streptomycin. The SUL2 genes are most important to predicting Sulfisoxazole, and the BLA genes are most important to classifying response to Ampicillin. When the genetic predictors are removed however, we see a consistent pattern of important variables for all 4 of the antibiotics of interest. Serovar information is most important and related to outcomes. This is followed by collection date, isolation source, isolation type, and country.

Figure 2 highlights similar information as **Figure 1** but for E.coli isolates' responses to the four most common antibiotics reported on for E.coli. While the TET and APH genes rank as the top 2 predictors for Tetracycline, in the source and genetic models for Ciprofloxacin, Meropenem, and Ampicillin, the absence/presence of antimicrobial resistant genes are not as important. In fact, the most important factors in these models share more similarities to the most important factors in their corresponding source predictors only models. In predicting response to Meropenem, country and collection date are particularly important. For Ampicillin, isolation source and isolation type are particularly important.

Figure 3 displays the most important covariates for the models involving the Campylobacter isolates. We observe the same pattern in predicting response to Tetracycline where information on the absence/presence of the TET and APH antimicrobial resistant genes are most important to prediction. In predicting susceptibility/resistance for Ciprofloxacin, the GYR gene seems to be very important, and at a much greater magnitude than all of the other predictors in the model. In predicting for both Gentamicin and Erythromycin, two of the variables that we created based on our data proved to be most important: the numeric evaluation of how similar the antibiotics that isolates are susceptible to within the same cluster are, and the numeric evaluation of how similar the antibiotics that isolates are resistant to within the same cluster are. In the models only concerning the source predictors, all but the model for tetracycline follow the importance pattern of isolation source, collection date, country, and isolation type.

Since serovar (within Salmonella isolates), isolation source and collection date are most often ranked as being the most important in source only models, we chose to briefly investigate under the hood to see if we could

Figure 1: Top 5 Most Imporant Variables in Models for Salmonella

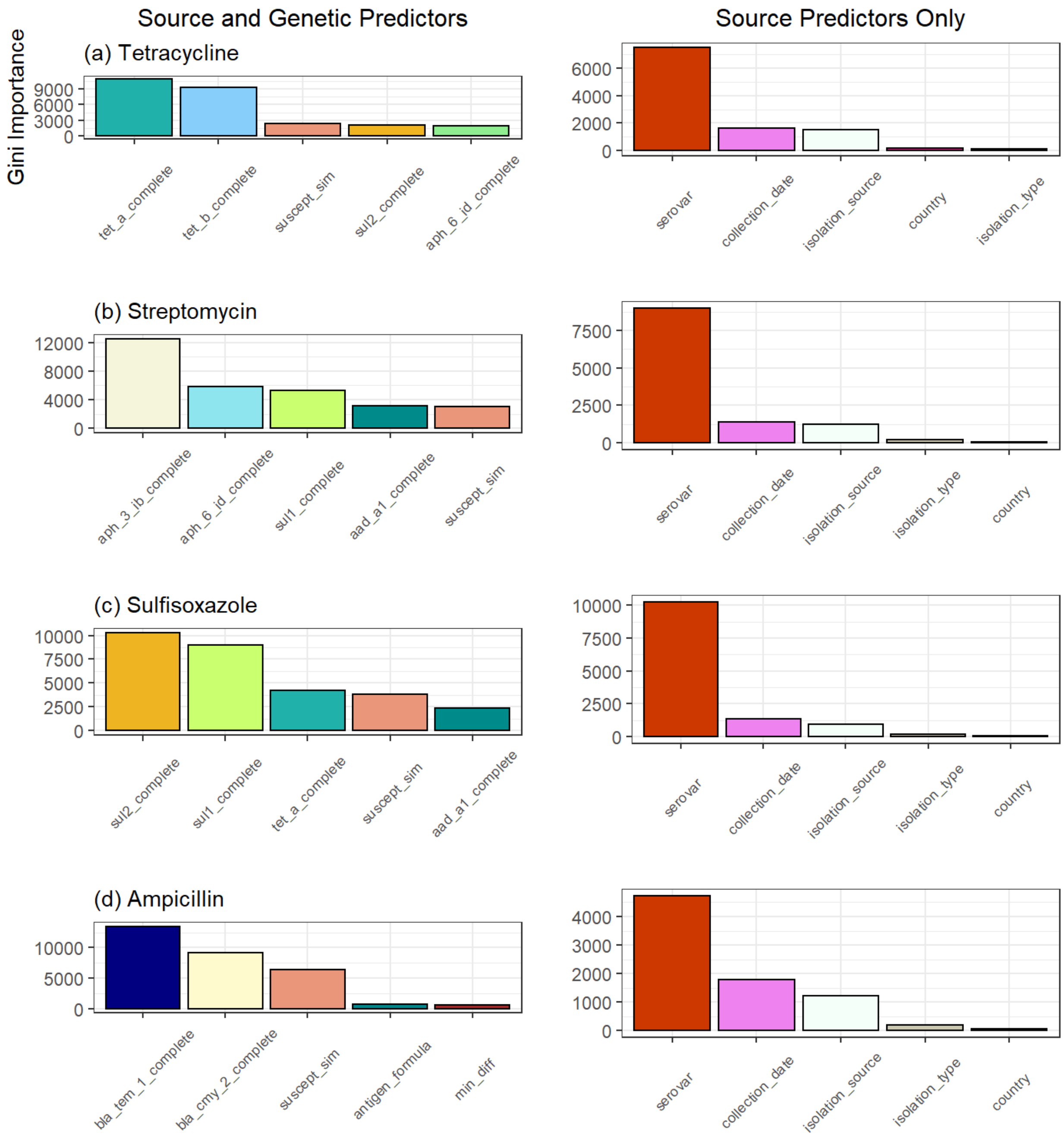


Figure 2: Top 5 Most Imporant Variables in Models for E.coli

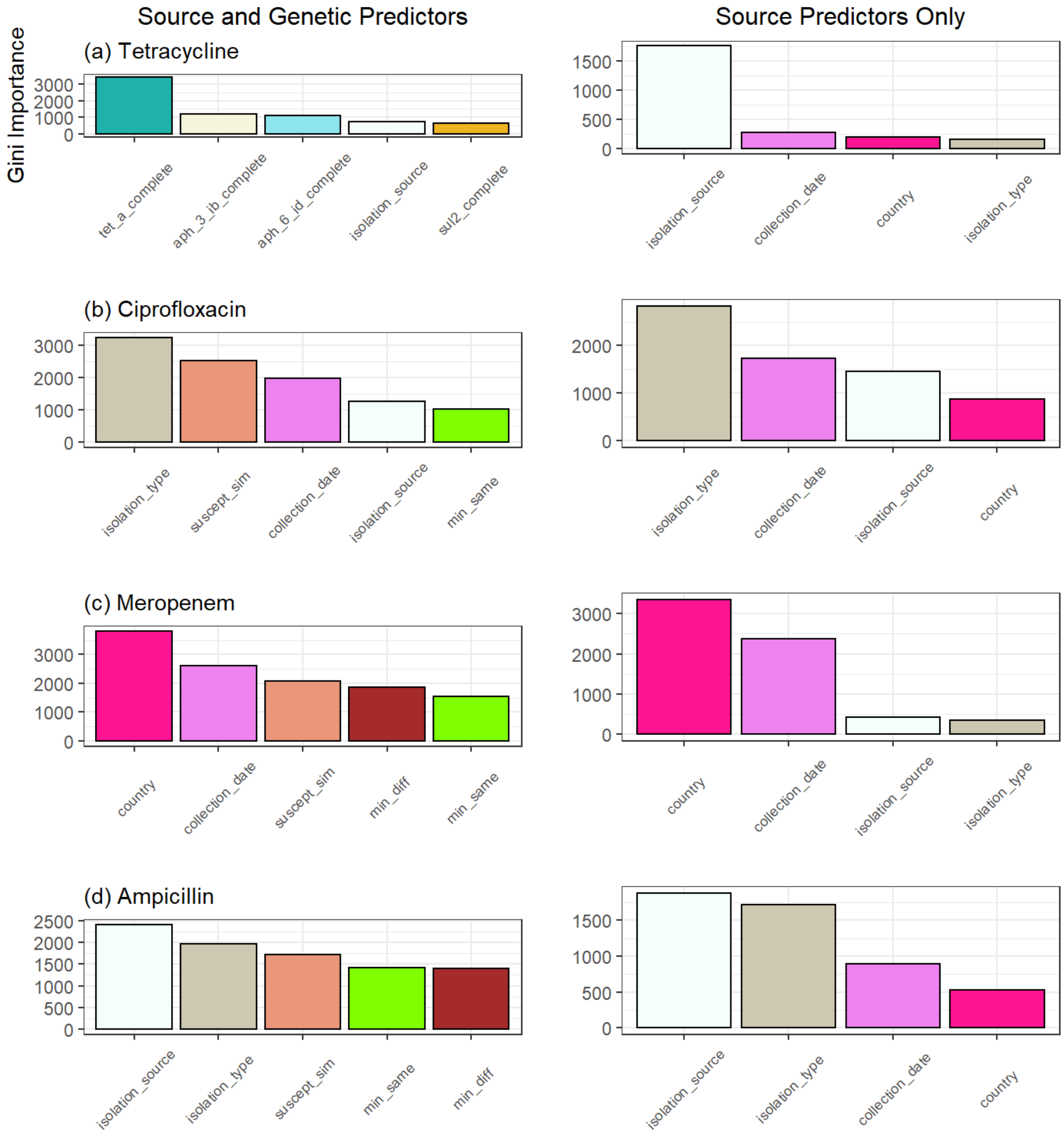
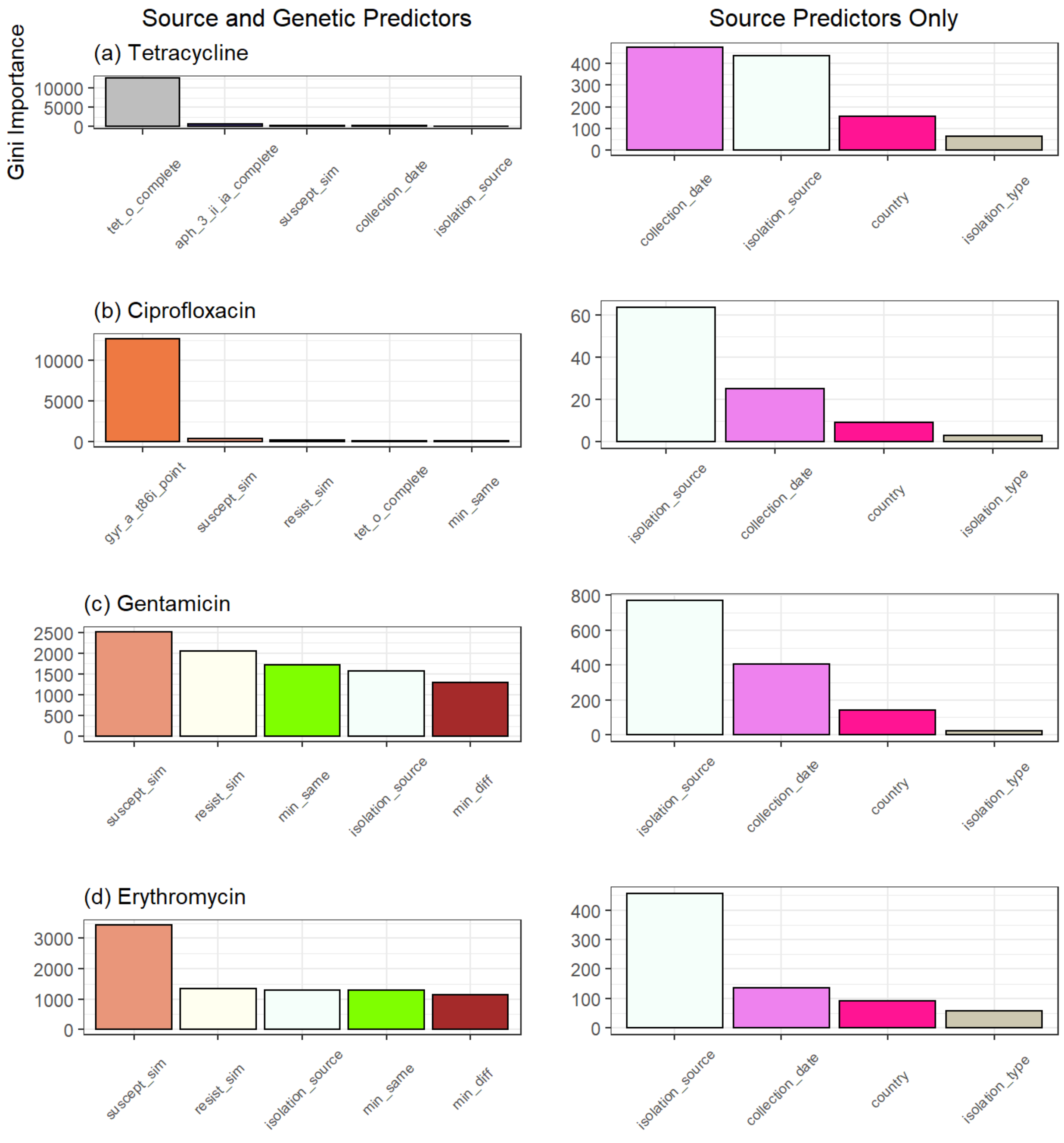


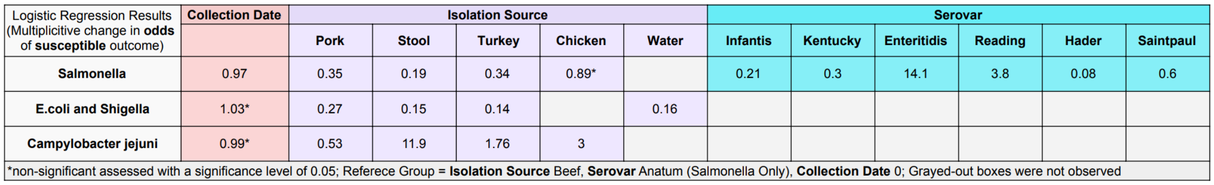
Figure 3: Top 5 Most Imporant Variables in Models for Campylobacter



quantify the relationships between these variables and response to tetracycline in all three of our species. Recall that we investigate this using logistic regression on our training data sets for all three species.

Logistic Regression Results

Table 3: Logistic Regression Results on Antibioitic Response to Tetracycline for Salmonella, E.coli, and Campylobacter isolates. Displays the multiplicative odds of having a susceptible response to Tetracycline.



Logistic Regression Results (Multiplicitive change in **odds** of **susceptible** outcome)	Collection Date	Isolation Source					Serovar					
		Pork	Stool	Turkey	Chicken	Water	Infantis	Kentucky	Enteritidis	Reading	Hader	Saintpaul
Salmonella	0.97	0.35	0.19	0.34	0.89*		0.21	0.3	14.1	3.8	0.08	0.6
E.coli and Shigella	1.03*	0.27	0.15	0.14		0.16						
Campylobacter jejuni	0.99*	0.53	11.9	1.76	3							

*non-significant assessed with a significance level of 0.05; Referece Group = **Isolation Source** Beef, **Serovar** Anatum (Salmonella Only), **Collection Date** 0; Grayed-out boxes were not observed

Table 3 displays the multiplicative odds constant associated with each variable of interest. Note that while isolation source and serovar are categorical variables and thus are indicators in nature, collection date is a continuous variable and so the odds constant represents with each 1-year increase, how much the odds of an isolate having a susceptible response is multiplied by. Values above 1 increase the odds while values less than 1 decrease the odds.

Most notably, serovar Enteritidis, and a stool isolation source have the largest *significant* magnitude of effect on the odds. While serovar Enteritidis drastically increases the odds of Salmonella isolates being susceptible to tetracycline, a stool isolation source decreases the odds the most, though not nearly as drastically. Additionally, the more recent the year, the slightly lower the odds of responding with susceptibility to tetracycline.

In the E.coli isolates, having a turkey isolation source lowers the odds of being susceptible to tetracycline the most, although an isolation source of stool or water have similar effects in terms of magnitude and direction.

Lastly, in the Campylobacter isolates, an isolation source of stool greatly increases the odds that the isolate will respond favorably to tetracycline.

Recommendation System

We next moved on to developing our antibiotic recommendation system. The system receives information about a new case as well as the species of that new case and pops out a recommendation based on the antibiotics that we have built models for, for each species. Recall that for each isolate, the antibiotic with predicted to have a susceptible response with the highest probability, will be the primary recommendation.If none of the drugs are predicted to garner a susceptible response, the system returns “No susceptibility”. We used the test data to illustrate the utility of this system.

For each isolate that it is fed, the algorithm recommends two antibiotics, one decided on by the source and genetic predictor model for that corresponding species, and the other decided by the source only model.

Table 4 picked 3 random isolates from each species that were fed into the algorithm and displays their recommendation results. From this table, we see examples of when the source and genetic models would agree, and when they would disagree. We compared the agreement between the source and genetic models and the source only models by finding the proportion of cases where the models agreed on the primary antibiotic recommendation for each species. For salmonella isolates, the two models agreed on ~41% of cases. For E.coli isolates, the two models agreed ~94% of the time, a huge leap in improvement from the salmonella isolates. Finally, the models on the Campylobacter isolates agreed ~84% of the time.

Table 4: Three randomly selected predictions from each species

	Species	Case number	Genetic + Source Prediction	Source Only Prediction
24388	Salmonella enterica	24388	Ampicillin	Streptomycin
43307	Salmonella enterica	43307	Streptomycin	Ampicillin
4050	Salmonella enterica	4050	Sulfisoxazole	Ampicillin
13499	E.coli and Shigella	13499	Meropenem	Meropenem
11571	E.coli and Shigella	11571	Meropenem	Meropenem
12257	E.coli and Shigella	12257	Meropenem	Meropenem
14262	Campylobacter jejuni	14262	Gentamicin	Gentamicin
13903	Campylobacter jejuni	13903	Ciprofloxacin	Gentamicin
9941	Campylobacter jejuni	9941	Gentamicin	Gentamicin

Discussion

Our models proved to perform significantly well on the NCBI data for Salmonella enterica, E.coli and Shigella, and Campylobacter isolates. It seems that even without retaining information for the presence/absence of antimicrobial resistant genes, our models still performed well, which may insinuate that ‘source information’ is often sufficient to prescribing antibiotic treatment if that is the desired root to tackle foodborne illnesses, especially when extracting that kind of genetic information is not feasible or timely enough. Overall, in our analysis, we noticed a pattern in that while the antimicrobial resistant genes were consistently important to predicting a susceptible or resistant response to tetracycline across all three of the species, even in the source and genetic models, genetic predictors were ranked as less important, although not ignorable. Generally, the TET, APH, SUL2, and BLA genes were the genes most imperative to have information about for predicting antibiotic response.

There were also some patterns that we were able to extract. In the source only models, serovar (for Salmonella only), isolation source, and collection date (which may indicate changes in resistance and susceptibility over time) were most commonly ranked as the most important predictors. Sources from pork, stool, and turkey were highly related to the antibiotic response of tetracycline, the most commonly reported antibiotic in the overall data. Serovar Enteritidis also had a very large effect on the odds of an isolate garnering a susceptible response to tetracycline. The inclusion of collection date as a top predictor also stirs some signal that antibiotic response is not consistent over time, and that perhaps due to prior mistreatment or even proper treatment patterns that have occurred involving treating these cases, certain genes have become more resistant to antibiotics that they might have been previously susceptible to. In future work, it may be of value to deeply examine patterns of when genes switch from being susceptible to resistant, and perhaps be able to predict when a gene will become resistant to certain antibiotics.

Lastly, in the assessment of our developed recommendation system, we overall observed the most agreement between decisions made by our source and genetic models and those made by our source only models, within the E.coli isolates (94% agreement), and in the Campylobacter isolates (84% agreement). This emphasizes the potential sufficiency of ‘source’ predictors without the extra assistance of antimicrobial resistant gene information for those species, but this should be validated further with the use of external data before any solid conclusions can be drawn.

Limitations

There are a few limitations of this study. Firstly, in order to predict responses to the antibiotics considered in our model, we considered a standard probability cut-off value of 0.5. While this value is commonly used, in practice, it is favorable to test for the best cutoff values, especially in thoughtful consideration of metric prioritization, i.e. the prioritization of sensitivity over specificity and vice versa. Additionally, 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 (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 were only helpful to refine our analysis for Salmonella but were not able to contribute to the other two bacteria of interest. 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

In the absence of traditional clinical-based procedures, we can turn to methods in machine learning to help prevent foodborne illnesses and outbreaks through the appropriate prescription of antibiotics to heterogeneous cases. Though the use of genetic information directly is most favorable to aid in this process, accessible source information still prove to have strong predictive power in select and non-negligible instances. The accurate and confident prescription of these drugs can help to target outbreaks at its source to quell the continued spread of disease without the risk of the unfavorable genetic outcomes that can result from mistreatment.

References

Dirac, P. A. M. 1953. "The Lorentz Transformation and Absolute Time." *Physica* 19 (1-–12): 888–96. https://doi.org/10.1016/S0031-8914(53)80099-6.

Feynman, R. P, and F. L Vernon Jr. 1963. "The Theory of a General Quantum System Interacting with a Linear Dissipative System." *Annals of Physics* 24: 118–73. https://doi.org/10.1016/0003-4916(63)90068-X.