Predicting Drug Resistance and Susceptibility in Foodborne Illnesses with Random Forests

Breanna Richards, Yanru Liao, Jina Yang
Department of Biostatistics, Brown University School of Public Health

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.

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

1. Introduction

Foodborne illnesses and disease are preventable public health threats, with approximately one in 10 people worldwide becoming ill each year after eating contaminated food and killing more than 420,000 people (n.d.). Antibiotics play a vital role in the prevention and treatment of these infectious diseases. However, the excessive and uncontrolled use of antibiotics to treat infectious diseases in humans and animals has brought new challenges to preventing transmission, which has become a major area of human health concern in recent years. The main mechanism of antibiotic drug failure is antibiotic residues in daily food such as raw meat and raw milk. One of the main consequences of antibiotic residues in foods of animal origin is the proliferation of antibiotic-resistant bacteria. The presence of antibiotic-resistant pathogens in food may contribute to increased incidence, treatment failure, and disease severity of foodborne infections in humans. In addition, they can transfer resistance genes to other microorganisms through the food chain (Hassani et al. 2022). Foods of animal origin, such as meat and meat products, are a major vehicle for the transmission of foodborne zoonotic bacterial pathogens. Escherichia coli (E.coli) and Salmonella are considered major zoonotic bacterial pathogens. These pathogens have been linked to numerous human foodborne illnesses and deaths worldwide following the consumption of contaminated food (Hashempour-Baltork et al. 2019; Rajaei et al. 2021) 2021). Effective treatment of these infections requires an understanding of the organism's susceptibility to various antibiotics. However, the traditional way to obtain this information is to grow bacteria in clinical laboratories and then test for commonly used antibiotics, which is very time-consuming (Ren et al. 2022; Van Camp, Haslam, and Porollo 2020). High-throughput sequencing technologies based on bacterial genomic data have the potential to rapidly guide the clinical decision-making process. The target pathogens we isolated in this study included Salmonella enterica, E.coli and Shigella and Campylobacter jejuni. Our project focuses on identifying the impact of genomic and source information that contributes to resistance and susceptibility predictions—resistance genes and genetic variants as well as information about the source of disease—and evaluating the utility of a non-parametric prediction model. We predict antibiotic susceptibility and resistance in Salmonella, E. coli, and Campylobacter foodborne illnesses to create an effective drug recommendation system for treating new cases. In addition, we contrasted the source-only model with a mixed model of source and genetic variables, and compared the discriminative confounding metrics of the two models. Ultimately, we hope that our analysis can help healthcare clinicians and policymakers eliminate the spread of these human diseases before population-scale outbreaks of infectious diseases occur by utilizing the most informative and accurate predictors and models.

Van Camp et al. presents a method to computationally simulate the antimicrobial spectrum of eight drugs. Their study used the NCBI BioSample database to train and cross-validate eight XGBoost-based machine learning models to predict the response to Cefepime, Cefotaxime, Ceftriaxone tested in Acinetobacter baumannii, Escherichia coli, Enterobacter, Ciprofloxacin, Gentamicin, Levofloxacin, Meropenem, and Tobramycin-resistant cloacae, Klebsiella aerogenes, and Klebsiella pneumoniae. Shotgun sequencing technology has been used to generate whole genome sequencing (WGS) data on the coverage of known antibiotic-resistance genes. The dataset has been split into training and testing datasets (915), and demos (31) - used to illustrate how to implement new unseen samples through the pipeline - in antimicrobial filtering, discarding incomplete information, representation, bacteria and drugs of interest subset, and antimicrobial resistance categories balances from the original NCBI biological samples (6564). After processing and filtering (i.e. analysis of wgs data and clustering of similar antibiotic resistance genes), 2605 of 4579 ARDs in NCBI were identified in at least one of 946 samples. To account for performance biases caused by alternate inputs in machine learning models that depend on training and test data cutoffs, they trained 51 independent models for each antibiotic isolate across all species. These final models have a good predictive fit, e.g., accuracy, AUC, and R-squared above 0.8. XGBoost allows the detection of the most important ARGCs for each model to examine correlated feature ARGCs and assess the influence of individual features in the decision process. The top 5 most important features identified by ARGC extracted from the models tend to overlap. To test the stability of the important features ARGC, they implemented parametric models (including LASSO and Ridge regression) to identify the number of top 5 unique features for each antibiotic among 51 independent models. The higher the consistency in selecting important features across different independent models, the higher the robustness. Furthermore, they introduced a reliability index (RI) to additionally assess the predictive confidence of a given model. It is suggested that the majority of samples evaluated with the high-reliability index appear to have correct predictions. XGBoost has several advantages, for example, its sophisticated machine learning algorithm works on complex relationships between ARGs and phenotypes; it provides insight into the decision-making process by reporting the list of features ARGCs are most frequently used in building models as a proxy for key decision insights. Their work provides a tool for rapidly testing bacterial samples to obtain a preliminary antimicrobial profile, which can guide initial antibiotic selection for treatment.

The methods and analyzes we performed in this study are an application of the paper published by Van Camp et al. in the 2020 Frontiers in Microbiology Journal (Van Camp, Haslam, and Porollo 2020). Specifically, we predicted whether a pathogen is resistant to Tetracycline, Streptomycin, Sulfisoxazole, Ampicillin, Ceftriaxone, etc. From previous studies, machine learning methods and penalized logistics outperform (Ren et al. 2022; Van Camp, Haslam, and Porollo 2020). We first converted all intermediate antibiotic resistance and proven resistance levels to 'resistant' and dose-dependent susceptibility levels to 'susceptible' to project the data to a binary classification problem. In terms of coverage of known antibiotic resistance genes, whole genome sequencing data were partial to our input data. Both source and genomic data were included in the model building in our analysis. The National Center for Biotechnology Information (NCBI) pathogen database has a high degree of sequence similarity (including polymorphisms in strains or sequences from closely related species) that comparison techniques cannot classify. We evaluated our model by comparing predicted antibiotic resistance and susceptibility with laboratory-confirmed drug susceptibility in measurements of AUC, accuracy, sensitivity, and specificity.

2. Methods

2.1 Data

The primary data source that we used for this study is the National Center for Biotechnology Information's (NCBI) Pathogen database. This database includes information about sequenced bacterial pathogens that originate from a food source. It records isolates with source and genetic information. Each row in the data is its own isolate, which is how we identify individual organisms that have little genetic mixing. We aimed to determine if we can learn from past isolates to develop a system to recommend antibiotics to new isolates. We wanted to observe the kind of patterns that we can extract from our current data about behaviors of drug susceptibility and resistance.

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

2.1.1 Variables

There are several variables in the NCBI dataset that we considered in our study (**Table 1**).

Table 1: Variables in Consideration for Random Forest Modeling

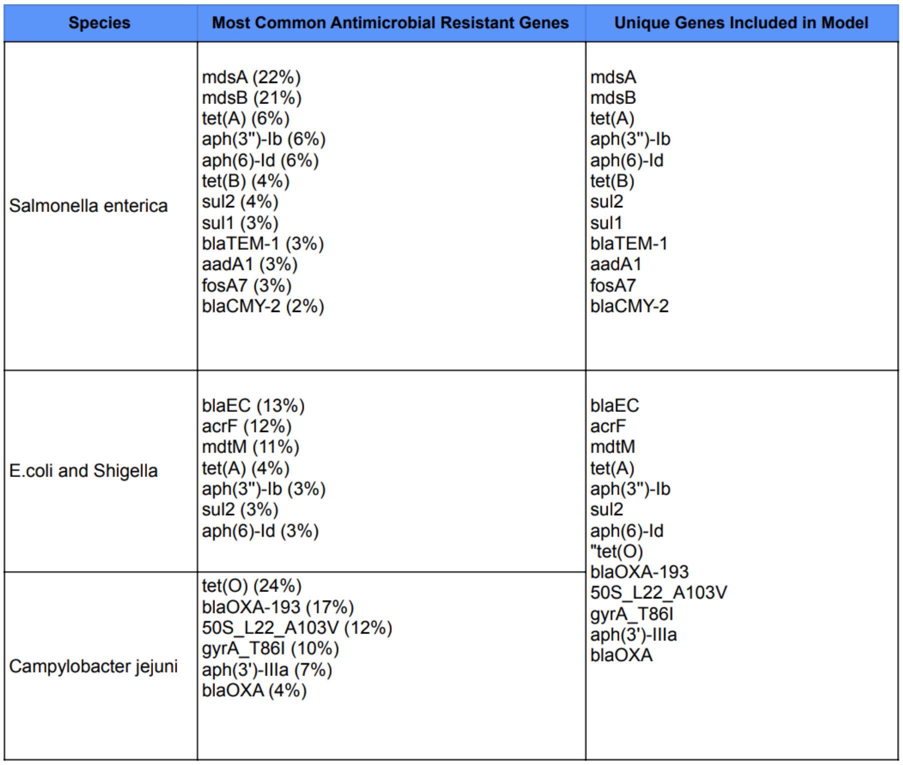
Variable Name(s)	Description
Organism Group	This variable refers to the taxonomy group that the isolate belongs to. For the purpose of our study this is going to take on one of three values: Salmonella enterica, E.coli Shigella, or Campylobacter jejuni.
Collection Date (Year)	Collection date gives the date that the sample was collected.
Location (Country)	The geographical origin of the sample.
Serovar	The serovar gives the distinct variation within a species of bacteria. This column groups isolates based on their surface antigens, allowing the classification of isolates to the subspecies level.

Variable Name(s)	Description
Isolation Source and **Isolation Type**	Isolation source describes the physical, environmental and/or local geographical source of the biological sample from which the sampled was derived. Examples include boneless beef, turkey, river, creek water, and water filtration membrane. Isolation type generalizes the isolation source into either clinical or environmental.
AST Phenotypes and **AMR Genotypes**	The AST Phenotype column gives the antibiotic resistant phenotype of the isolate. This variable is key to our study and lists antibiotics that the isolate has demonstrated resistance, susceptibility, intermediate responsiveness, and/or dose-dependent susceptibility to. AMR genotypes gives the antimicrobial resistant genes found in the isolate.
Antigen Formula	The antigen formula gives the detected presence of specific viral antigen which indicates viral infection.
Min-Same and **Min-Diff**	Min-same is the minimum SNP distance to another isolate of the same isolation type (clinical or environmental). Min-diff is the minimum SNP distance to another isolate of a different isolation type. For example, the minimum SNP difference from an environmental isolate to a clinical isolate.
Suscept_Sim and **Resist_Sim**	These are self-computed similarity scores between the drugs that isolates within the same clusters are susceptible and resistant to, respectively (computed using a pairwise string similarity function in R). Values range between 0 and 1 (inclusive), corresponding to perfect dissimilarity at a value of 0, and perfect similarity at a value of 1. Each isolate's value is the average similarity in antibiotics that they share for other isolates within their SNP cluster.

Reformatting the Most Common Antimicrobial Resistant Genes We reformatted the prevalence of the antimicrobial-resistant genes found in each of our isolates. There are 499 unique resistant genes found in the isolates within our data. In order to include some level of gene resistant presence as predictors in our model, we chose to include binary variables for the most common resistant genes observed. These variables took a value of 1 if found within a given isolate, and a value of 0 is not found within a given isolate. The genes that we decided to include as predictors in each of the two separate models are displayed in **Table 2**. We decided to retain the genes with over 2% prevalence in the Salmonella isolates, and the genes with over 3% prevalence in either the E.coli isolates or the Campylobacter isolates.

Initially, we had decided to evaluate isolates from these three bacteria within joint modelling and analysis procedures. However, we found that Salmonella isolates had retained two additional pieces of information compared to the E.coli and Campylobacter isolates. This information are antigen formula, which gives the detected presence of a specific viral antigen indicating viral infection, and serovar which gives the distinct variation within a certain bacteria. Additionally, while isolate data for Salmonella were retained from 2002-2021, E.coli and Campylobacter records did not begin presenting themselves in our data until much later (2009 for E.coli and Shigella, and 2016 for Campylobacter). For these reasons, we split our dataset into three groups, one for each of the three species/organism groups. Notably, this separation was also convenient for the balancing of the different sample sizes within our datasets. Overall, we retained 8,672 observations on Salmonella, 3,880 on E.coli, and 3,497 on Campylobacter. By separating the data by species, we avoided a joint final model being trained heavily on the Salmonella group due to the relatively larger sample size and

Table 2: Antimicrobial Resistant Genes Considered in Modeling



Species	Most Common Antimicrobial Resistant Genes	Unique Genes Included in Model
Salmonella enterica	mdsA (22%) mdsB (21%) tet(A) (6%) aph(3")-Ib (6%) aph(6)-Id (6%) tet(B) (4%) sul2 (4%) sul1 (3%) blaTEM-1 (3%) aadA1 (3%) fosA7 (3%) blaCMY-2 (2%)	mdsA mdsB tet(A) aph(3")-Ib aph(6)-Id tet(B) sul2 sul1 blaTEM-1 aadA1 fosA7 blaCMY-2
E.coli and Shigella	blaEC (13%) acrF (12%) mdtM (11%) tet(A) (4%) aph(3")-Ib (3%) sul2 (3%) aph(6)-Id (3%)	blaEC acrF mdtM tet(A) aph(3")-Ib sul2 aph(6)-Id "tet(O) blaOXA-193 50S_L22_A103V gyrA_T86I aph(3')-IIIa blaOXA
Campylobacter jejuni	tet(O) (24%) blaOXA-193 (17%) 50S_L22_A103V (12%) gyrA_T86I (10%) aph(3')-IIIa (7%) blaOXA (4%)	

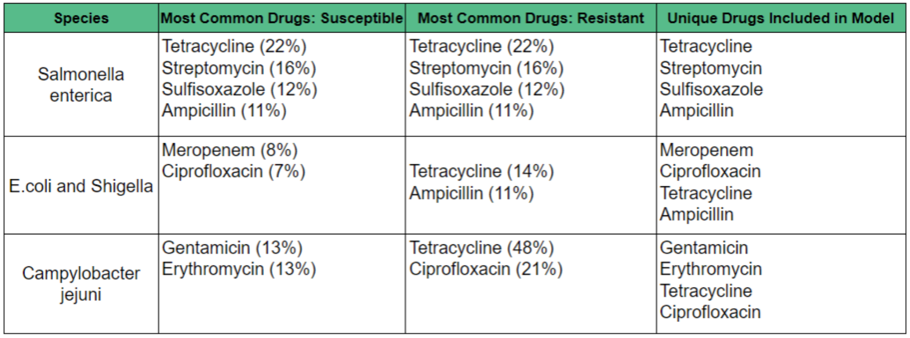
extra information.

The full scope of our data contains isolates that are resistant and/or susceptible to 79 unique antibiotics, 76 of which these isolates have shown resistance to, and 62 of which these isolates have shown susceptibility to. Evidently, whereas some isolates have shown resistance to a certain drug, other isolates have shown susceptibility to the same drug. The idea was to create a system or function to take genetic and baseline covariate information about isolates as input values and then within a subset of the most common drugs we've found the isolates to show the most response to, recommend the one that we've predicted with the most evidence that new isolate is susceptible to. Therefore, since this is a binary classification problem, on the data that we used to develop our models, we coded '1' as drug susceptibility, '0' as drug resistance, and NA if the data did not retain any information about that particular drug on a given isolate. The subset of most common antibiotics that we considered in our models fulfilled the following criteria: be within the top two (or top four in the result of ties) most common drugs that isolates of a given bacteria have shown susceptibility or resistance to. **Table 3** shows the most common antibiotics that isolates within each bacteria showed resistance and susceptibility to. Overall, each species' isolates were assessed to four different antibiotics.

2.1.2 Handling Missing Data: Multiple Imputation

In order to move forward with implementing our methods, we desired a way to handle the missingness in our data. Upon thorough investigation of the data which led us to believe that we were dealing with missingness at random (MAR), we decided to impute these missing values using Multiple Imputation. To do so, we utilized the `mice` function from the `mice` package in R (van Buuren and Groothuis-Oudshoorn 2011). Multivariate Imputation via Chained Equations (MICE) imputes on a variable-by-variable basis, meaning that a new imputation model is specified with each new variable containing missingness. Linear regression/predictive mean matching was used to predict missing values for continuous variables, and logistic regression was used to predict missing values for categorical variables. As a rule of thumb, we omitted the variables with over 80% missingness from our dataset first (which led to the omission of one of the self-made variables from exploratory data analysis assessing the measure of similarity between the antibiotics that isolates of the same SNP cluster show resistance to), and then imputed the remaining variables. For each of our three species of interest, we imputed five datasets and vertically stacked these imputed datasets for use in model implementation. These imputed datasets resulted in 43,360 total observations for Salmonella, 19,400 for E.coli, and 17,485 for Campylobacter.

Table 3: 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

2.2 Modeling Procedures

2.2.1 Variable Selection

In order to increase the accuracy of prediction and decrease computational time, we decided to consider variable selection. We used the feature selection wrapper algorithm, Boruta, to potentially reduce the dimensionality of our data before applying them to classification (`Boruta` function from the `Boruta` package in R) (Kursa and Rudnicki 2010). In use, the Boruta algorithm first adds randomness to the dataset of interest by creating shuffled copies of all of the features up for consideration. These new shuffled copies are called shadow features. Next, the function applies the random forest on the data and evaluates the importance of each of the features in the classification process. With each iteration of the classifier, the Boruta algorithm checks if the un-shuffled, original feature has a higher importance than the most important of its shuffled shadow features. If an original covariate is not deemed more important to classification than the best of that covariate's associated shadow features, then it is deemed unimportant and removed as a variable worth keeping in the modeling process. The Boruta algorithm stops when either all features are either confirmed or denied as being important, or alternatively when the maximum number of a specified number of random forest runs is reached.

2.2.2 Random Forest Classification

For the purpose of classification, we decided to use the Random Forest algorithm. In order to implement this process in R, we utilized the `randomForest` function from the `randomForest` package (Liaw and Wiener 2002). The Random Forest model's building components are decision trees. Decision trees are a sort of supervised learning in which input is continuously split based on certain parameters. Random Forest fits various trees with different bootstrap samples and splits with random feature subsets on each node. Predictions are made using the average of the results from each tree. To partition the data, the decision tree and random forest employ information gain and impurity to ensure that non-informative features are not chosen. To optimize the decision tree, a splitting criterion must be considered; it is critical to decide on splitting options to minimize errors caused by a few observations in a node, which would eventually lead to a lack of statistical significance. However, overfitting can occur when the tree is too deep, and underfitting can happen when the minimal number of samples to split is too small. In order to combat these potential issues in R, we considered tuning the following parameters for the algorithm on a training dataset: `ntree` indicating the number of trees, and mtry indicating the number of randomly sampled variables in each split. Different combinations of these parameters were explored in a grid search process using 10-fold cross validation, and chosen based on their resulting out-of-bag errors, a metric used to measure the prediction error of random forests on data points not utilized in a given bootstrap sample (akin to an internal testing set). More specifically, we chose the set of parameters that produced the lowest out-of-bag errors. The performance of our trained models were primarily assessed on withheld testing sets for each species. These models were also utilized in the system that was developed to recommend antibiotics to new isolates.

2.3 Variable Importance

After hyper-parameter tuning, we extracted the variables that our final random forest models deemed to be most important in the classification process. This was assessed using "Gini feature importance", which is commonly used as an indicator of feature relevance. This score provides a relative ranking of the spectral features, a by-product in the training of the random forest (Menze et al. 2009). It indicates how often a particular feature was selected for a split, and how large its overall value was in the classification problem, higher values indicating more importance.

2.4 Methods of Random Forest Evaluation

We assessed the performance of the models using accuracy, sensitivity, specificity, and area under the ROC curve (AUC). Moreover, gini importance captured strong predictors of drug response. We also use the area under the ROC curve (AUC) as a measure of classifier performance (Bradley 1997). The larger the AUC, the better the performance of the model.

2.5 Logistic Regression

One drawback of the random forest algorithm is that it does not make it accessible to extract quantifiable relationships between predictors and outcomes in terms of how magnitudes of changes in predictors affect the odds of observing the outcome. In order to remedy this, we used multivariable logistic regression to extract patterns between drug susceptibility and resistance for the most common drug in the data (tetracycline) and some of the most prominent variables to the prediction process, as observed in observations of variable importance. Logistic regression was conducted using the glm function from the stats package in R (R Core Team 2022).

3. Results

A total of 80,245 isolates (as a result of multiple imputation) with antibiotic information (43,360 observations on Salmonella, 19,400 on E.coli, and 17,485 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 reported antibiotics in our data (**Table 3**).

Data for each species were split 80-20 into training sets used to build that species' associated models, and testing sets 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. Variable selection was only 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 was split into 10 subsets with the models being 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 were 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`) (Biau and Scornet 2016). 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, were the values used in the final random forest models.

Set $m =$ number of predictors in the data. We tested three common values for mtry: $\sqrt{m}$, $m/3$, and $m/2$ (all rounded down to the nearest integer) (Liaw and Wiener 2002). 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 weighted classes by $\frac{1}{prop\ of\ samples\ in\ class}$.

3.1 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 considered a variable that numerically captured the drug *resistance* similarity to other isolates within the same cluster, but unlike in Salmonella, we did not consider serovar and antigen formula due to that information only being retained on Salmonella isolates. Lastly, we considered 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 2**.

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 is not tied to a geographical origin on the other hand, may be discovered through visits to the doctor's office.

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 (models with outcome Tetracycline, outcome streptomycin, outcome Ampicillin, and outcome Sulfisoxazole). And so, during hyper-parameter tuning for Salmonella, we considered all predictors in the data.

E.coli and Shigella In predicting E.coli responses 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.

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

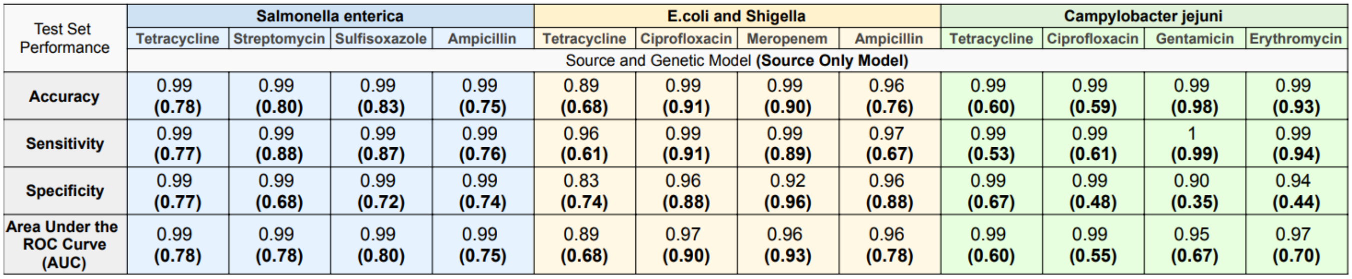
Campylobacter jejuni In predicting antibiotic response to tetracycline within Campylobacter isolates, seven antimicrobial resistant genes were deemed unimportant. The same seven genes were deemed unimportant in predicting Gentamicin, Erythromycin, and Ciprofloxacin.

See `report_data_processing.R` in the 'Final Report' folder in the GitHub repository for more details on the variable selection process.

3.2 Test Set Performance

After hyper-parameter tuning (see `hyperparameter_tuning.R` in the 'Final Report' folder in the GitHub repository for more details), we assessed the performance of our models on testing sets for each species of interest (**Table 4**).

Table 4: 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, all of our 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. 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 performed 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 and doctor-approved use, can our models still be reliable to the average person?

The results of our 'source only' models are bold in **Table 4**. Notably, removing the genetically associated predictors from our models 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.

3.3 Variable Importance

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

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

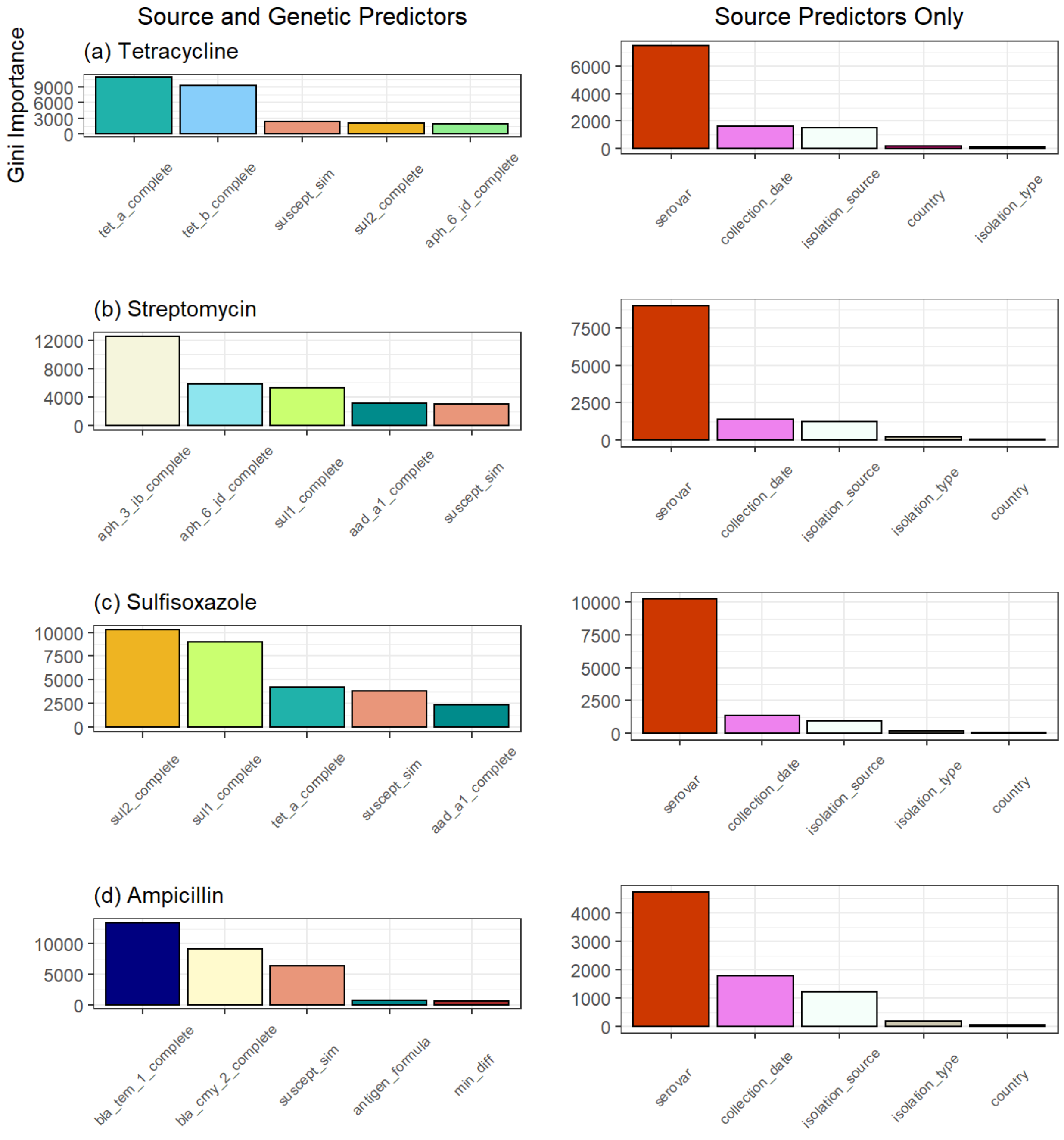
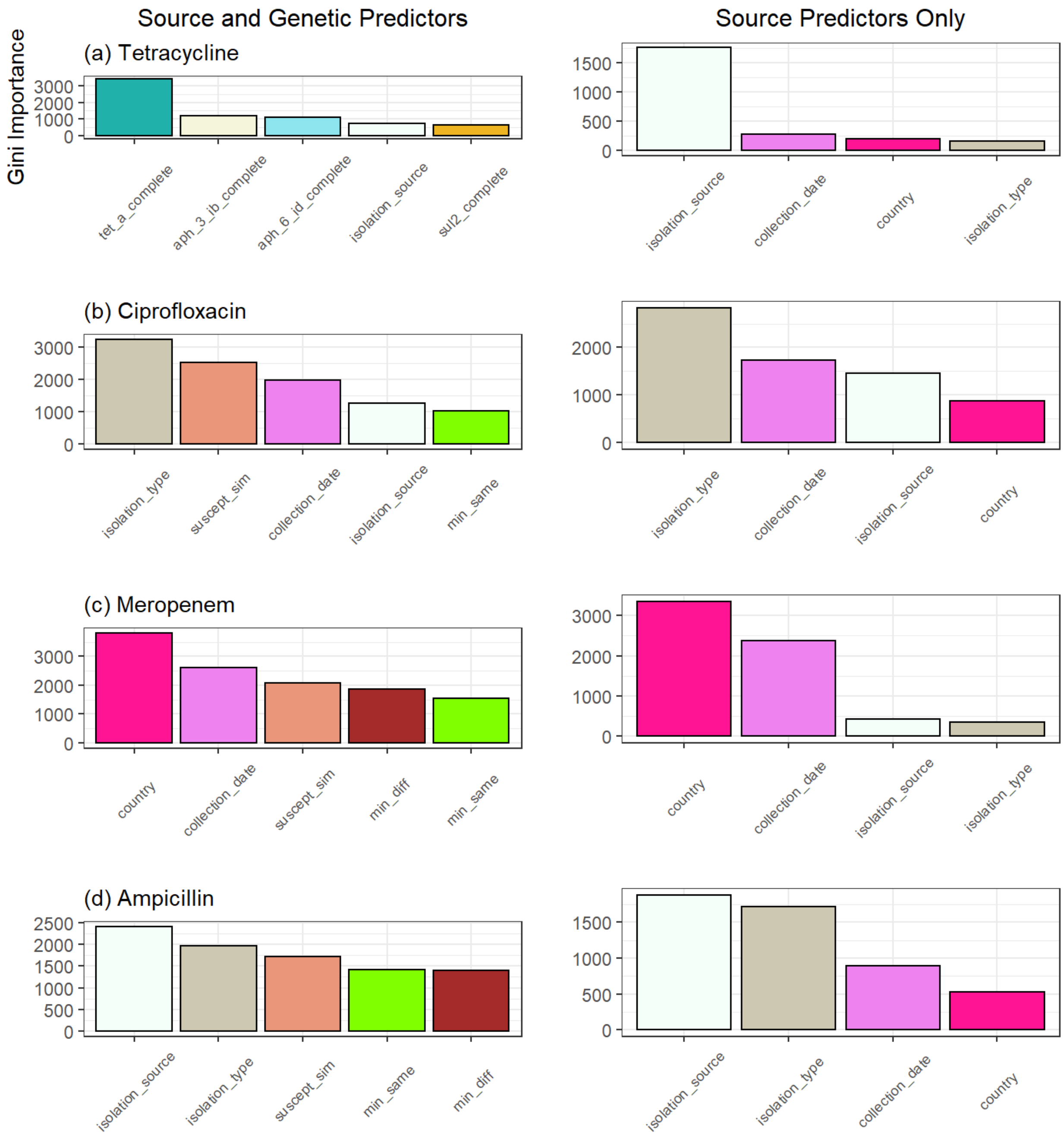


Figure 2: Top 5 Most Imporant Variables in Models for E.coli



Source and Genetic Predictors

Source Predictors Only

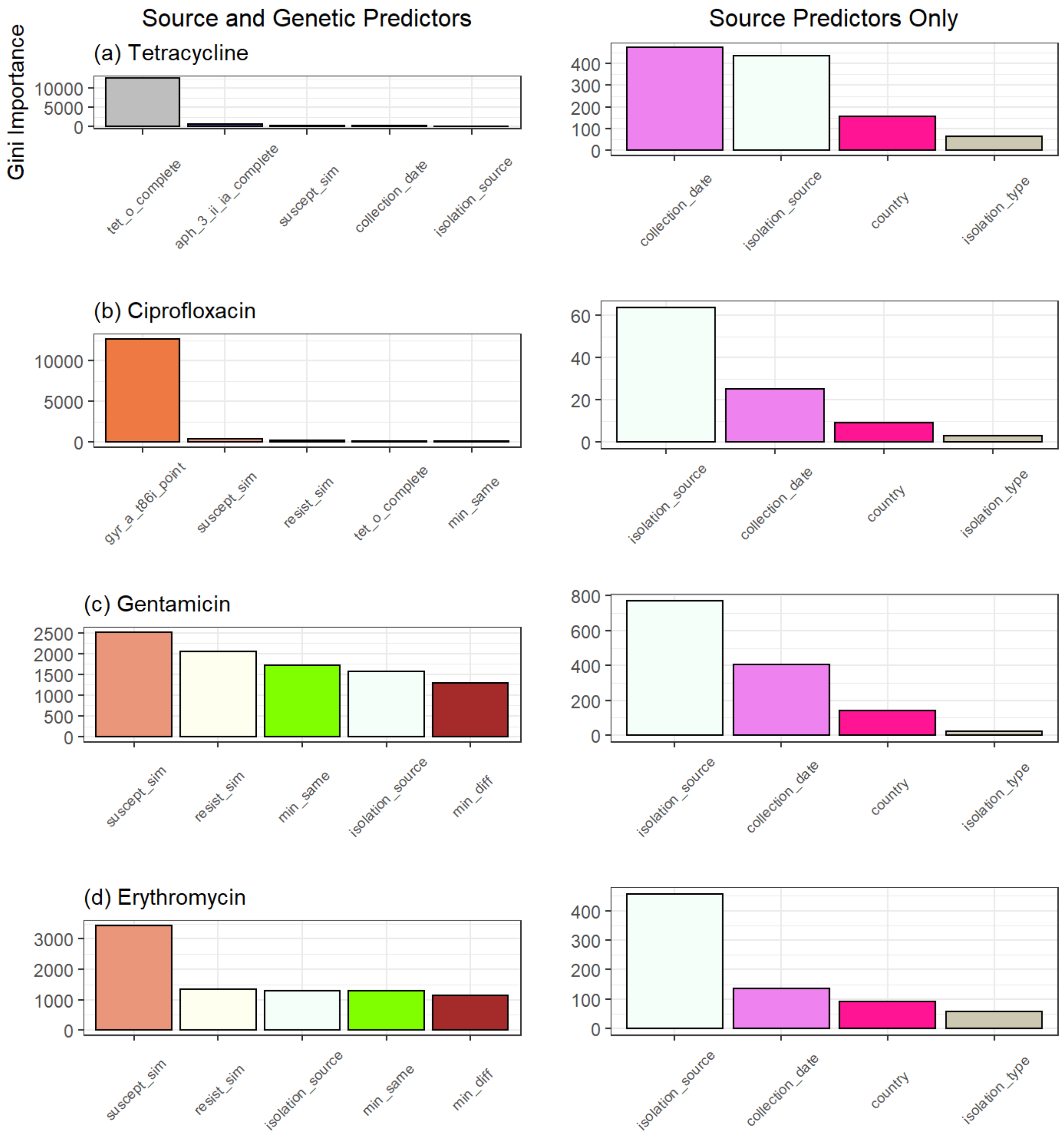
(a) Tetracycline

(b) Ciprofloxacin

(c) Meropenem

(d) Ampicillin

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



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.

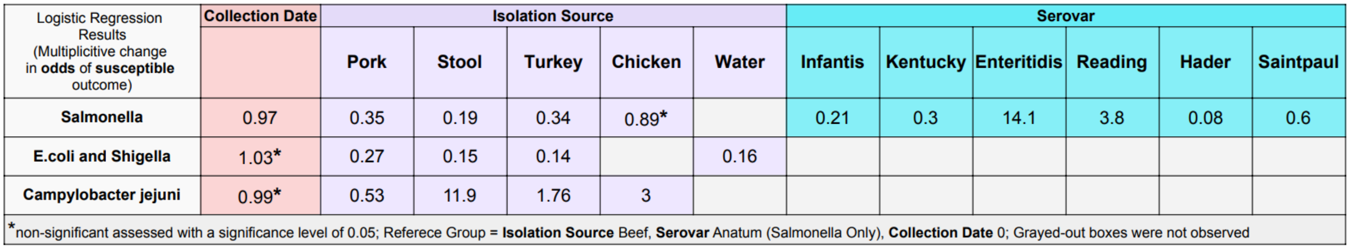
Figure 2 highlights similar information as **Figure 1** but for E.coli isolates' responses to the four most common antibiotics reported on. While the TET and APH genes rank as the top 2 predictors for Tetracycline, in the source and genetic models for Ciprofloxacin, Meropenem, and Ampicillin, the absence/presence 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 were particularly important. For Ampicillin, isolation source and isolation type were 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 (only within Salmonella isolates), isolation source and collection date were most often ranked as being the most important in the source only models, we chose to briefly investigate under the hood to see if we could quantify the relationships between these variables and response to tetracycline in all three of our species. Recall that we investigated this using mutlivariable logistic regression on our training data sets for all three species.

3.4 Logistic Regression Results

Table 5: 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 5 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 (farthest away from 1). While serovar Enteritidis drastically increases the odds of Salmonella

Table 6: Three randomly selected recommendations 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

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.

3.5 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 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 6 picked three 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 test 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.

4. 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. 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 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.

4.1 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 or 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 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. This may bring some nuance into the problem space of what we observe in our data. Additionally, the antigen formula and serovar variables only had information for Salmonella isolates. And so, these variables were only helpful to refine our analyses for Salmonella. 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 measure of what could be represented in data like this in future years.

4.2 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 sources to quell the continued spread of disease without the risk of the unfavorable genetic outcomes that can result from mistreatment.

References

n.d. *Foodborne Illness - Healthy People 2030*. Healthy People 2030. https://health.gov/healthypeople/objectives-and-data/browse-objectives/foodborne-illness.

Biau, Gérard, and Erwan Scornet. 2016. "A Random Forest Guided Tour." *Test* 25 (2): 197–227.

Bradley, Andrew P. 1997. "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms." *Pattern Recognition* 30 (7): 1145–59.

Hashempour-Baltork, Fataneh, Hedayat Hosseini, Saeedeh Shojaee-Aliabadi, Mohammadali Torbati, Adel Mirza Alizadeh, and Matin Alizadeh. 2019. "Drug Resistance and the Prevention Strategies in Food Borne Bacteria: An Update Review." *Advanced Pharmaceutical Bulletin* 9 (3): 335.

Hassani, Sima, Mir-Hassan Moosavy, Sahar Nouri Gharajalar, Seyed Amin Khatibi, Abolfazl Hajibemani, and Zahra Barabadi. 2022. "High Prevalence of Antibiotic Resistance in Pathogenic Foodborne Bacteria Isolated from Bovine Milk." *Scientific Reports* 12 (1): 1–10.

Kursa, Miron B., and Witold R. Rudnicki. 2010. "Feature Selection with the Boruta Package." *Journal of Statistical Software* 36 (11): 1–13. http://www.jstatsoft.org/v36/i11/.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

Menze, Bjoern H, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. 2009. "A Comparison of Random Forest and Its Gini Importance with Standard Chemometric Methods for the Feature Selection and Classification of Spectral Data." *BMC Bioinformatics* 10 (1): 1–16.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rajaei, Maryam, Mir-Hassan Moosavy, Sahar Nouri Gharajalar, and Seyed Amin Khatibi. 2021. "Antibiotic Resistance in the Pathogenic Foodborne Bacteria Isolated from Raw Kebab and Hamburger: Phenotypic and Genotypic Study." *BMC Microbiology* 21 (1): 1–16.

Ren, Yunxiao, Trinad Chakraborty, Swapnil Doijad, Linda Falgenhauer, Jane Falgenhauer, Alexander Goesmann, Anne-Christin Hauschild, Oliver Schwengers, and Dominik Heider. 2022. "Prediction of Antimicrobial Resistance Based on Whole-Genome Sequencing and Machine Learning." *Bioinformatics* 38 (2): 325–34.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in r." *Journal of Statistical Software* 45 (3): 1–67. https://doi.org/10.18637/jss.v045.i03.

Van Camp, Pieter-Jan, David B Haslam, and Aleksey Porollo. 2020. "Prediction of Antimicrobial Resistance in Gram-Negative Bacteria from Whole-Genome Sequencing Data." *Frontiers in Microbiology*, 1013.