**Analyzing multidrug resistance patterns across the food supply chain using association rule mining**

Joshua Glass1, Gayatri Anil1,2, Kristina M. Ceres2, Laura B. Goodman2, Casey L. Cazer1,2

1Department of Clinical Sciences, College of Veterinary Medicine, Cornell University

2Department of Public and Ecosystem Health, College of Veterinary Medicine, Cornell University

Corresponding Author: Joshua Glass

Email: [jg2527@cornell.edu](mailto:jg2527@cornell.edu)

**1. Introduction**

**2. Methods**

2.1 Data Preparation

Three datasets were analyzed in total: data from retail meat samples retrieved from the National Antimicrobial Monitoring System (NARMS)[CITE NARMS], data from cecal (slaughterhouse) samples retrieved from NARMS, and data retrieved from the National Animal Health Laboratory Network (NAHLN)[CITE NAHLN SOURCE]. For all datasets, only *E. coli* isolates associated with cattle were examined. The retail meats dataset included minimum inhibitory concentrations (MICs) for years 2002 to 2019, and genotypic information for years 2017 to 2019. The cecal dataset included MICs for years 2013 to 2021, and genotypic information for years 2017 to 2021. The NAHLN dataset included MICs and genotypic information for years 2018 to 2022.

MIC data was examined for the drugs listed in Table 1 and were interpretated as resistant or susceptible based on breakpoints from NARMS interpretive[CITE NARMS INTERPRETIVE CRITERIA] criteria or ECOFFS retrieved from EUCAST[CITE THE EUCAST ECOFFS SOURCE]. The genotypic information for each dataset was comprised of resistance genes as well as genetic mutations that were included due to their likelihood of conferring resistance (mistranslations were not considered).

The goals of this study were to examine how multidrug resistance (MDR) patterns change over time, how MDR patterns change across the food supply chain, and how well MDR patterns as indicated by phenotypic data correspond to those indicated by genotypic data. All datasets were stratified to produce the appropriate comparison sets. The three original datasets represent three different points in the food supply chain: comparing across these three data sources achieves the goal of quantifying change across the food chain. The three datasets were further divided based on year to quantify change across time. Finally, all datasets were divided based on resistance indicator (i.e., genotypic or phenotypic). Each source-year-indicator dataset was analyzed separately. Prevalence descriptives for phenotypic drug resistance and resistance genes/mutations can be found in Tables 2 and 3.

2.2 Association Mining

Association rule mining was implemented use the R package arules[CITE ARULES]. To appropriately format the data for association rule mining, each bacterial isolate was treated as a transaction. When mining phenotypic data, the set of drugs tested were treated as the possible items for the transaction: if the isolate was resistant to a drug, it was coded as 1 (the item is present in the transaction); and if the isolate was susceptible to a drug, it was coded as 0 (the item is not present in the transaction). Similarly, when mining genotypic data, the set of possible genes/mutations were treated as the items for the transaction: if a gene/mutation was present in the isolate, it was coded as 1; and if a gene/mutation was not present in the isolate it was coded as 0. The data was mined both at the basic level (mining drugs or genes/mutations) and at the class level. For phenotypic, data each drug was mapped to its corresponding class. If multiple drugs mapped to the same class, the largest value for that class for each isolate was retained (for example, if three drugs of the same class were coded as [0,0,1] for a particular isolate, then the final coding for the class would be 1 for that isolate). The same procedure was followed for the resistance genes/mutations. Classifications for genes/mutations were retrieved from [CITE WHERE GENE MAPPINGS CAME FROM]. In the case where a single gene conferred resistance to multiple classes of antimicrobials, an item for each of the classes was created with identical codings from that gene (for example if a gene that confers resistance to two classes was coded as 1, then both classes would be included and would be coded as 1).

2.2.1 Quality Measure Selection

If all possible rules are generated, it is likely that some will be spurious, and there may be some that are not spurious but that are uninformative. Therefore, it is important to be able to prune rules based on their quality (i.e., how interesting they are and how likely they are to describe meaningful associations), and there are numerous measures of rule quality available[57 (I think, check); CITE INTEREST MEASURES DOC].

To determine which quality measure would be most appropriate for the data, rules were mined for each source-year-indicator dataset, and all available quality measures were calculated. Principal component analysis was performed on all quality measures[CITE PCA PACKAGE USED], and the first four principal components were examined. In each of the first four principal components the top five variable loadings were extracted for each source-year dataset, and the frequency with which each measure was present in the top five variable loadings was computed for each principal component.

In addition to this, the mathematical properties of the measures were considered. Since the data is sparse, two potential properties are desired to have in a measure: null-invariance (the strength of the association is influenced by the presence of items rather than their absence) and symmetry (the value of the measure is the same for A🡪B and B🡪A). From the measures that were found in the PCA procedure, four null invariant and symmetrical measures were chosen: cosine, jaccard, kulcszynski, and support. Support is a commonly used quality measure and simply reports how often a rule appears as a proportion of the sample size of the dataset. The formula for support can be seen in equation 1.

**Equation 1**

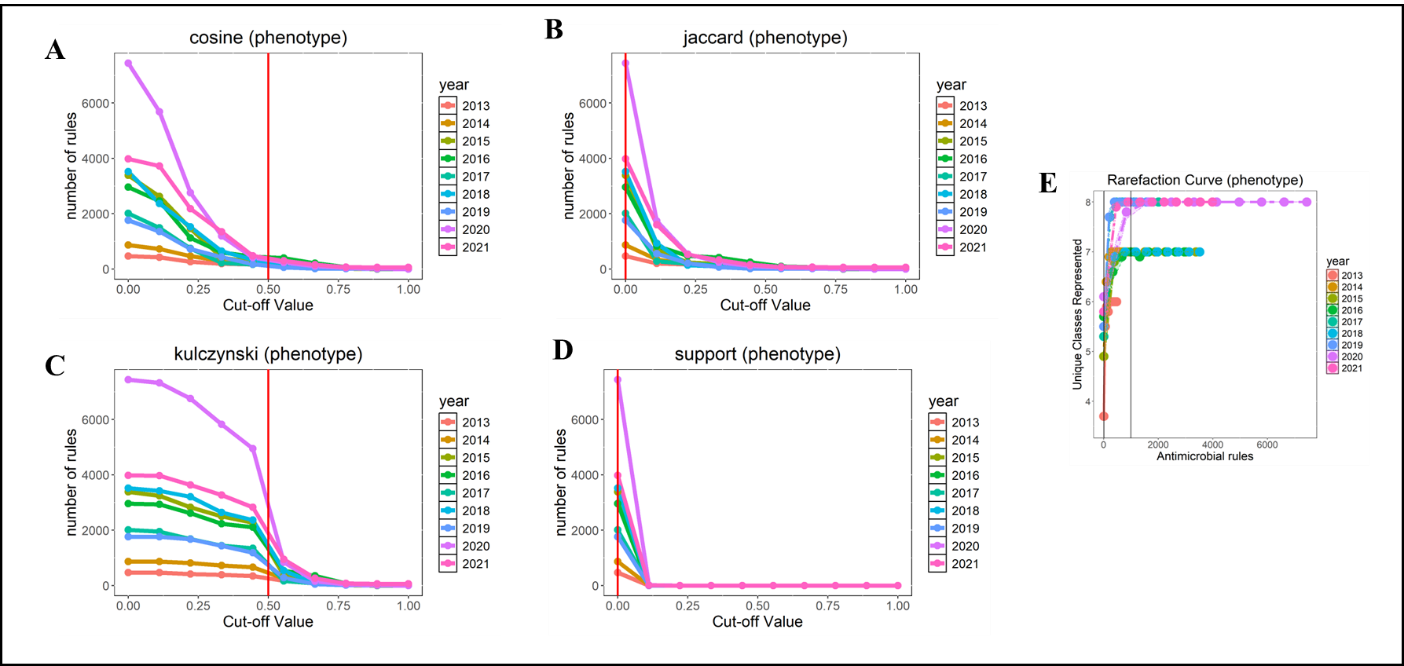
Kulcszynski is a modified version of confidence. Confidence represents the conditional probability that given the left-hand side of the rule, the right-hand side of the rule will be present as well (or vice verse). Confidence is not symmetrical. Kulcszynski is made symmetrical by taking both directions of the conditional probability and averaging them. Kulcszynski ranges from 0 to 1, with 0 meaning that the RHS and LHS never occur together, 1 meaning that they always occur together, and 0.5 suggesting that the rule is probably uninteresting. (see Equation 2).

**Equation 2**

Jaccard shows how strongly the RHS and LHS are associated by dividing the support of the intersection of RHS and LHS occurrences by the support of the union of RHS and LHS occurrences (Equation 3).

2.2.2 Cut-off Selection

To determine how many rules are needed to fully describe the data, random subsamples of rules were taken at various subsample sizes and the number of antimicrobial classes represented by each subsample was computed (figure 1E). This determined the minimum number of rules required to ensure that all the antimicrobial classes were represented in the rule set. To select cut-off values for each quality measure, rules were mined and pruned at various cut-off values. Cut-off values that achieved the target rule number were chosen.

****

**Figure 1: Quality measure cut-off selection.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Drug | Abbreviation | Resistant Breakpoint | In Dataset |
| AMINOGLYCOSIDE | Gentamicin | GEN | >=16 | Retail Meats; Cecal; NAHLN |
|  | Streptomycin | STR | >=32 | Retail Meats; Cecal |
|  | Neomycin | NEO | >8\* | NAHLN |
|  | Spectinomycin | SPE | >64\* | NAHLN |
| BETA-LACTAM | Amoxicillin–Clavulanic Acid | AMC | >=32 | Retail Meats; Cecal |
|  | Meropenem | MER | >=4 | Retail Meats; Cecal |
|  | Cefoxitin | FOX | >=32 | Retail Meats; Cecal |
|  | Ceftriaxone | AXO | >=4 | Retail Meats; Cecal |
|  | Ampicillin | AMP | >=32 | Retail Meats; Cecal; NAHLN |
|  | Ceftiofur | TIO | >=8 | NAHLN |
|  | Penicillin | PEN | NA | NAHLN |
| FOLATE-PATHWAY-INHIBITOR | Sulfisoxazole | FIS | >256 | Retail Meats; Cecal |
|  | Sulfamethoxazole | SMX | >256 | Retail Meats; Cecal |
|  | Trimethoprim/sulfamethoxazole | COT | >=4 | Retail Meats; Cecal; NAHLN |
|  | Sulphadimethoxine | SUL | NA | NAHLN |
| LINCOSAMIDE | Clindamycin | CLI | NA | NAHLN |
| MACROLIDE | Azithromycin | AZI | >=32 | Retail Meats; Cecal |
|  | Gamithromycin | GAM | NA | NAHLN |
|  | Tildipirosin | DIP | NA | NAHLN |
|  | Tilmicosin | TIL | NA | NAHLN |
|  | Tulathromycin | TUL | NA | NAHLN |
|  | Tylosin | TYL | NA | NAHLN |
| PHENICOL | Chloramphenicol | CHL | >=32 | Retail Meats; Cecal |
|  | Florfenicol | FFN | >16\* | NAHLN |
| PLEUROMUTILIN | Tiamulin | TIA | NA | NAHLN |
| QUINOLONE | Ciprofloxacin | CIP | >=1 | Retail Meats; Cecal |
|  | Nalidixic acid | NAL | >=32 | Retail Meats; Cecal |
|  | Danofloxacin | DAN | NA | NAHLN |
|  | Enrofloxacin | ENR | >0.125\* | NAHLN |
| TETRACYCLINE | Tetracycline | TET | >=16 | Retail Meats; Cecal; NAHLN |
|  | Chlortetracycline | LOR | NA | NAHLN |
|  | Oxytetracycline | OXY | NA | NAHLN |

**Table 1: Drug Information. Breakpoints with \* indicate that these are ECOFFs from EUCAST; all other breakpoints are from NARMS interpretative criteria.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Antimicrobial Level Phenotypic Resistance Prevalence | | | | | | | | | | | | | | | | | |
| Retail Meats | Year | N | AMC | AMP | AXO | AZI | CHL | CIP | COT | FIS | FOX | GEN | MER | NAL | SMX | TET | STR |
|  | 2002 | 295 | 2.03 | 6.1 | 0 | NA | 1.02 | 0 | 0.68 | NA | 0 | 0.34 | NA | 0 | 10.17 | 30.85 | NA |
|  | 2003 | 311 | 2.25 | 5.14 | 0.32 | NA | 2.25 | 0 | 0.32 | NA | 0 | 0.96 | NA | 0.96 | 10.29 | 25.08 | NA |
|  | 2004 | 338 | 3.85 | 5.33 | 1.48 | NA | 3.55 | 0 | 0.59 | 13.02 | 1.18 | 0.59 | NA | 1.48 | NA | 22.78 | NA |
|  | 2005 | 316 | 1.27 | 3.48 | 1.9 | NA | 1.58 | 0.32 | 0.63 | 6.96 | 0.95 | 0 | NA | 1.27 | NA | 16.46 | NA |
|  | 2006 | 295 | 2.37 | 9.15 | 1.69 | NA | 1.36 | 0 | 1.36 | 12.54 | 2.03 | 4.07 | NA | 0.68 | NA | 25.42 | NA |
|  | 2007 | 256 | 0.78 | 6.64 | 0.78 | NA | 3.91 | 0 | 1.17 | 9.38 | 0.78 | 0 | NA | 0.39 | NA | 21.88 | NA |
|  | 2008 | 250 | 2.4 | 6.4 | 1.6 | NA | 0.8 | 0 | 2 | 11.6 | 2.4 | 2 | NA | 0.4 | NA | 24 | NA |
|  | 2009 | 247 | 1.62 | 4.88 | 0.81 | NA | 2.43 | 0 | 2.02 | 7.69 | 1.62 | 0.81 | NA | 0.4 | NA | 18.62 | NA |
|  | 2010 | 269 | 1.12 | 4.83 | 1.12 | NA | 2.6 | 0 | 0.74 | 12.64 | 1.12 | 0.37 | NA | 0 | NA | 22.68 | NA |
|  | 2011 | 215 | 0.47 | 3.72 | 0.47 | 0 | 1.4 | 0 | 2.33 | 7.91 | 0.47 | 0.47 | NA | 0 | NA | 17.67 | NA |
|  | 2012 | 271 | 1.48 | 2.58 | 0 | 0 | 1.11 | 0 | 0.37 | 7.38 | 1.85 | 0.74 | NA | 1.48 | NA | 22.14 | NA |
|  | 2013 | 227 | 1.76 | 4.85 | 2.2 | 0 | 3.96 | 0 | 1.76 | 7.93 | 1.32 | 0 | NA | 0.44 | NA | 22.47 | NA |
|  | 2014 | 205 | 0.49 | 4.39 | 0.49 | 0 | 0.49 | 0 | 0.98 | 7.8 | 0.98 | 0.49 | NA | 0 | NA | 21.46 | NA |
|  | 2015 | 227 | 0.44 | 2.64 | 0.44 | 0 | 1.32 | 0 | 1.32 | 7.49 | 0.44 | 0.44 | NA | 0.44 | NA | 18.5 | NA |
|  | 2016 | 174 | 1.15 | 6.32 | 0.57 | 0 | 5.17 | 0 | 1.15 | 9.77 | 1.15 | 0.57 | 0 | 0 | NA | 23.56 | NA |
|  | 2017 | 271 | 0.37 | 6.27 | 0.37 | 0 | 2.58 | 0 | 1.11 | 6.27 | 0 | 0 | 0 | 1.85 | NA | 21.77 | 7.75 |
|  | 2018 | 133 | 0.75 | 3.76 | 0 | 0 | 3.76 | 0.75 | 0.75 | 10.53 | 0 | 0 | 0 | 1.5 | NA | 20.3 | 12.78 |
|  | 2019 | 286 | 0 | 2.1 | 0 | 0 | 2.45 | 0 | 0.7 | 6.64 | 0 | 0.7 | 0 | 0.7 | NA | 19.58 | 9.44 |
| Cecal | 2013 | 549 | 0.73 | 3.46 | 0.73 | 0 | 3.1 | 0 | 0 | 7.47 | 0.55 | 0.18 | NA | 0 | NA | 21.68 | NA |
|  | 2014 | 503 | 0.99 | 4.97 | 0.99 | 0 | 3.78 | 0.2 | 0.8 | 11.53 | 0.8 | 0.4 | NA | 0.4 | NA | 31.81 | NA |
|  | 2015 | 891 | 1.46 | 7.18 | 1.57 | 0 | 5.5 | 1.12 | 1.57 | 14.37 | 1.35 | 0.79 | NA | 2.02 | NA | 33.11 | NA |
|  | 2016 | 1188 | 1.26 | 6.06 | 1.18 | 0.17 | 5.81 | 0.25 | 0.59 | 13.55 | 1.18 | 0 | 0 | 0.93 | NA | 32.41 | NA |
|  | 2017 | 1382 | 0.87 | 5.79 | 0.94 | 0.22 | 5.14 | 0 | 1.16 | 13.89 | 0.72 | 0.07 | 0 | 0.65 | NA | 32.49 | 15.05 |
|  | 2018 | 1444 | 0.97 | 6.44 | 0.97 | 0 | 7.13 | 0.28 | 1.25 | 13.92 | 1.04 | 0.35 | 0 | 1.59 | NA | 31.65 | 15.65 |
|  | 2019 | 1181 | 0.59 | 5.33 | 0.85 | 0.08 | 5.76 | 0.17 | 1.02 | 11.77 | 0.59 | 0.25 | 0 | 1.35 | NA | 27.35 | 12.79 |
|  | 2020 | 779 | 4.24 | 16.82 | 5.01 | 0.77 | 10.65 | 1.03 | 8.1 | 19.77 | 4.36 | 1.16 | 0 | 1.54 | NA | 34.58 | 0 |
|  | 2021 | 429 | 1.86 | 9.56 | 3.96 | 0.93 | 8.16 | 0.47 | 3.96 | 15.85 | 2.1 | 1.4 | 0 | 1.4 | NA | 31 | NA |

**Table 2: Phenotypic Resistance Prevalence Descriptives**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class Level Genotype Resistance Prevalence | | | | | | | | | | | | | | | |
| Retail Meats | Year | N | Aminoglycoside | Beta-lactam | Bleomycin | Colistin | Efflux | Folate-pathway-inhibitor | Fosfomycin | Lincosamide | Macrolide | Phenicol | Quinolone | Streptogramin | Tetracycline |
|  | 2017 | 148 | 8.78 | 0.68 | NA | 0 | NA | 5.41 | 2.03 | 0 | 0 | 2.7 | 0 | NA | 25 |
|  | 2018 | 133 | 17.29 | 0 | NA | 0.75 | NA | 10.53 | 2.26 | 0 | 100 | 3.76 | 0.75 | NA | 23.31 |
|  | 2019 | 286 | 11.54 | 0 | NA | 0 | NA | 6.99 | 1.4 | 0.35 | 100 | 2.45 | 0 | NA | 22.73 |
| Cecal | 2017 | 130 | 11.54 | 0.77 | 0 | 0 | 0 | 76.15 | 0 | 0.77 | 1.54 | 33.85 | 3.85 | 0.77 | 84.62 |
|  | 2018 | 593 | 6.41 | 0.34 | 0 | 0 | 0 | 29.85 | 0 | 0 | 0 | 16.36 | 3.37 | 0 | 49.24 |
|  | 2019 | 268 | 11.19 | 1.12 | 0 | 0 | 0 | 44.4 | 0 | 0 | 30.22 | 21.64 | 2.24 | 0 | 75 |
|  | 2020 | 289 | 21.11 | 4.15 | 0.35 | 0.35 | 7.96 | 44.29 | 8.3 | 2.08 | 2.08 | 22.84 | 2.08 | 0 | 61.25 |
|  | 2021 | 262 | 9.92 | 1.53 | 0 | 0 | 6.49 | 24.43 | 23.66 | 1.15 | 1.53 | 12.6 | 0.76 | 0.38 | 41.6 |

**Table 3: Genotypic Resistance Prevalence Descriptives**

**3 Results**

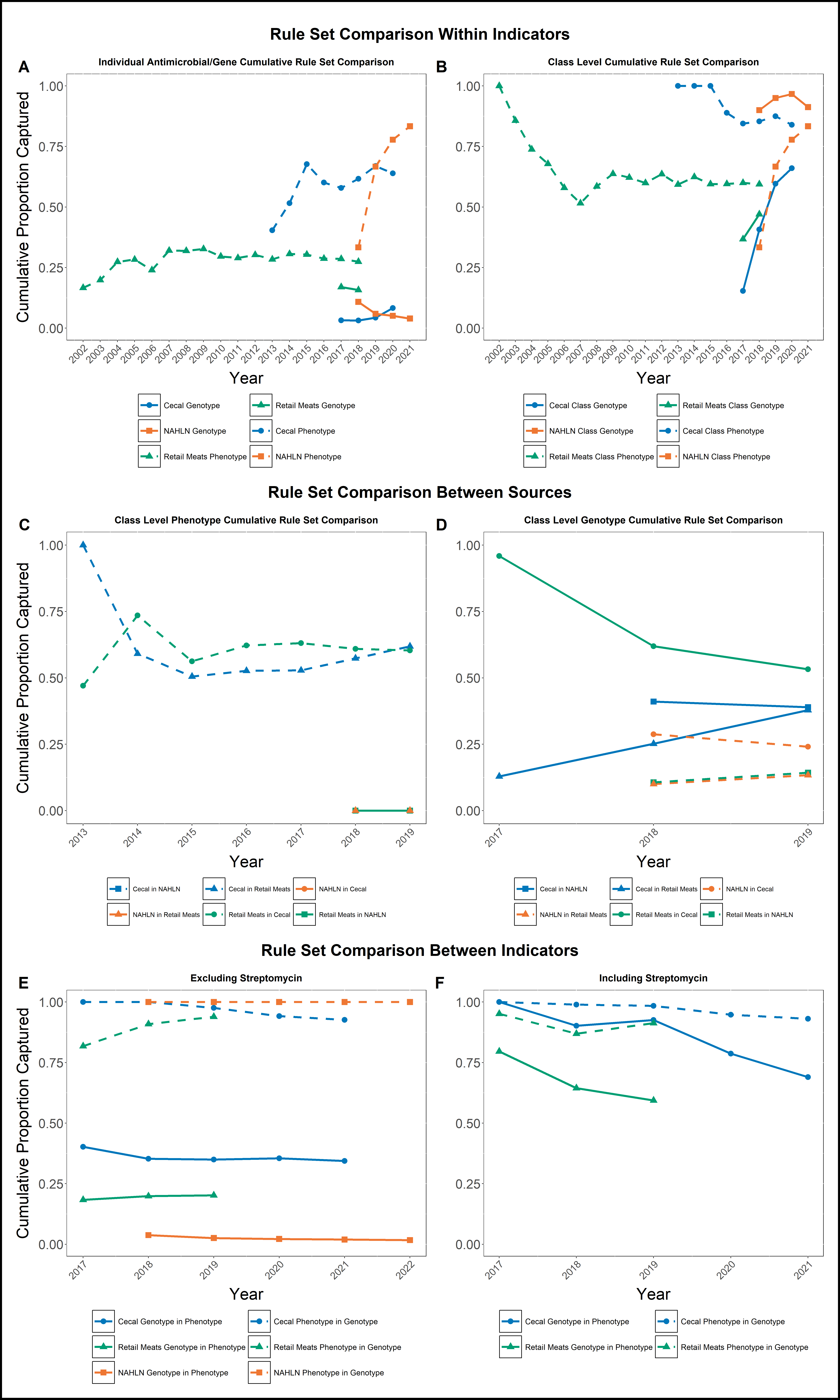
3.1 Rule Set Comparison

Rule sets were compared numerically and graphically. To compare the rule sets numerically, the cumulative percent captured was calculated for each rule set pair considered. Cumulative percent captured was calculated by dividing the magnitude of the intersection of two rule sets by the magnitude of the reference rule set (i.e., the rule set that is capturing the other rule set). This calculation was done for each year and progressively averaged across years making it cumulative. Three types of comparisons were made this way: within indicators comparisons, between source comparisons, and between indicators comparisons.

Within indicators comparisons were done by comparing a single indicator (phenotype or genotype) from a single source. This was done both at the individual antimicrobial/gene level and at the antimicrobial class level for each data source (retail meats, cecal, or NAHLN) across time (Figure 2A and 2B). These comparisons are meant to quantify how rule sets change over time.

For between sources comparisons, rule sets from the same year and indicator but different sources were compared. This was done both at the individual antimicrobial or gene level and at the antimicrobial class level for each data source (retail meats, cecal, or NAHLN) across time (Figure 2C and 2D). was done to quantify how rule sets change depending on where the samples were sourced from.

Comparisons between indicators needed to be done at the class level so that phenotypic resistance and genotypic resistance could be put into the same terms. Rule sets of the same year and source but different indicators were compared. Generally, the genotypic rules do a better job of capturing the phenotypic rules than the phenotypic rules do at capturing the genotypic rules. This asymmetry may be driven largely by drug specific aminoglycoside resistance genes. Overall, the data analysis did not include streptomycin due to the breakpoint changing in 2014 causing discrepancies in the data. However, to highlight the effect of the aminoglycosides on the asymmetry, this comparison was done both including and excluding streptomycin (Figure 2D and 2E).

****

**Figure 2: rule set comparisons.**

3.2 Network Graph Analysis

See powerpoint.

3.3 Tabulation Comparison

The mined rules were compared to resistance patterns generated by tabulation. It was found that almost all the tabulated resistance patterns were captured by the association rules. In addition to producing the relevant resistance patterns, the association rules provide additional information about the strength of the associations represented by such patterns. Table 4 shows the top ten most frequently occurring resistance patterns.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Resistance Pattern | Frequency | Recovered Resistance Pattern | avg\_cosine | avg\_jaccard | avg\_kulczynski | avg\_support |
| FIS; TET | 506 | Yes | 0.612058562 | 0.383148606 | 0.680821588 | 0.187871759 |
| AMP; CHL; FIS; TET | 177 | Yes | 0.609642497 | 0.404687916 | 0.65098744 | 0.082892825 |
| CHL; FIS; TET | 139 | Yes | 0.615500823 | 0.391843387 | 0.677938812 | 0.111249879 |
| AMP; TET | 100 | Yes | 0.573386093 | 0.335048004 | 0.657710314 | 0.283871282 |
| AMP; FIS; TET | 29 | Yes | 0.616043034 | 0.39887251 | 0.671127967 | 0.164037633 |
| AMP; CHL; COT; FIS; TET | 27 | Yes | 0.600242555 | 0.396845151 | 0.639724026 | 0.074347442 |
| CHL; FIS; NAL; TET | 19 | Yes | 0.615500823 | 0.391843387 | 0.677938812 | 0.111249879 |
| AMC; AMP; AXO; FOX | 17 | Yes | 0.792221706 | 0.659873112 | 0.802678286 | 0.012311695 |
| CHL; COT; FIS; TET | 17 | Yes | 0.609400169 | 0.389584313 | 0.667530929 | 0.102080526 |
| AMC; AMP; AXO; CHL; FIS; FOX; TET | 14 | Yes | 0.653318631 | 0.45718924 | 0.695248592 | 0.017263242 |

**Table 4: Tabulation patterns recovered by rule mining.**

**4 Discussion**

**5 Conclusions**