

# Phenotypic resistance and AMR gene–phenotype concordance in *Escherichia coli* represented in linked EMBL-EBI surveillance records from seven Sub-Saharan African countries

Afeez Olabisi Yusuf

yusuf.ao@lasustech.edu.ng

Lagos State University of Science and Technology

Kehinde Olugbenga Amisu

Lagos State University

Oluchi Miriam Ukhureigbe

Lagos State University

Kofoworola Funmilayo Odutayo

Lagos State University

Adedayo Peter Oladetuyi

Eko City Polytechnic, Ikotun-Egbe

---

## Research Article

**Keywords:** antimicrobial resistance, *Escherichia coli*, Sub-Saharan Africa, multidrug resistance, resistome, genomic surveillance, phenotype–genotype concordance, machine learning, WHO GLASS

**Posted Date:** September 28th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10924248/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

## ORIGINAL ARTICLE

# Phenotypic resistance and AMR gene–phenotype concordance in *Escherichia coli* represented in linked EMBL-EBI surveillance records from seven Sub-Saharan African countries

### Authors:

Afeez Olabisi Yusuf<sup>1,2\*</sup>, Kehinde Olugbenga Amisu<sup>1</sup>, Oluchi Miriam Ukhureigbe, Kofoworola Funmilayo Odutayo<sup>1</sup>, Adedayo Peter Oladetuyi<sup>4</sup>

### Affiliations:

<sup>1</sup> Department of Microbiology, Lagos State University, Ojo, Lagos, Nigeria

<sup>2</sup> Department of Biological Science, Microbiology Unit, Lagos State University of Science and Technology, Ikorodu, Lagos, Nigeria

<sup>3</sup> Department of Microbiology, Faculty of science, Federal University Oye-Ekiti.

<sup>4</sup> Department of Science Laboratory Technology, School of Pure and Applied Science, Eko City Polytechnic, Ikotun-Egbe, Lagos, Nigeria

\*Corresponding author: [yusuf.ao@lasustech.edu.ng](mailto:yusuf.ao@lasustech.edu.ng) | ORCID: 0000-0002-5100-2722

### Abstract

**Background:** Publicly available antimicrobial resistance (AMR) databases provide opportunities to examine linked phenotypic and resistance gene data across diverse geographical settings. This study descriptively analysed phenotypic resistance patterns and associated AMR gene profiles of *Escherichia coli* isolates from seven Sub-Saharan African countries using records from the EMBL-EBI Antimicrobial Resistance Portal.

**Methods:** A secondary analysis was conducted using 375 linked phenotype–genotype antimicrobial susceptibility records representing 111 unique *E. coli* isolates collected between 2008 and 2018. Phenotypic resistance frequencies, multidrug resistance (MDR), geographical distribution, phenotype–genotype concordance, AMR gene diversity, gene co-occurrence, and descriptive comparisons with WHO GLASS 2022 data were evaluated. Logistic regression and Random Forest models were explored to assess factors associated with phenotypic resistance.

**Results:** Phenotypic resistance was observed in 309/375 antimicrobial susceptibility records (82.4%), while 44/111 unique isolates (39.6%) met the operational definition of multidrug resistance used in this study. Resistance was highest for trimethoprim (94.2%) and tetracycline

(93.2%). Twenty-four AMR-associated genes representing six resistance classes were identified, with *tetA* and *neo* among the most frequently detected. Several gene pairs showed non-random co-occurrence, particularly *tetA*–*neo*. Phenotype–genotype concordance was generally limited, with Cohen's kappa values ranging from  $-0.024$  to  $0.240$  across the antibiotics examined. Observed resistance also varied among countries and specimen sources, although interpretation was constrained by substantial differences in sample size and sampling distribution.

**Conclusions:** Within the available linked EMBL-EBI surveillance records, *E. coli* showed substantial observed phenotypic resistance and multidrug resistance, while AMR-associated gene detection showed limited concordance with corresponding phenotypes. These findings demonstrate the potential of linked public surveillance data for descriptive investigation of antimicrobial resistance and genotype–phenotype relationships. However, the relatively small number of unique isolates and uneven geographical, temporal and specimen-source representation limit generalizability. The findings should therefore not be interpreted as population-level estimates or forecasts of antimicrobial resistance across Sub-Saharan Africa. Larger, standardized and geographically representative surveillance datasets are required to establish robust regional estimates and clarify genotype–phenotype relationships.

**Keywords:** antimicrobial resistance; *Escherichia coli*; Sub-Saharan Africa; multidrug resistance; resistome; genomic surveillance; phenotype–genotype concordance; machine learning; WHO GLASS.

## 1. Introduction

Antimicrobial resistance (AMR) is a major global public health threat, with an estimated 1.27 million deaths directly attributable to bacterial AMR in 2019 and a disproportionate burden in Sub-Saharan Africa (SSA) (Murray et al., 2022). Earlier projections estimated that, without effective intervention, AMR could contribute to 10 million deaths annually by 2050 and impose substantial economic costs (O'Neill, 2016). In response, the World Health Organization (WHO) established the Global Action Plan on Antimicrobial Resistance and the Global Antimicrobial Resistance and Use Surveillance System (GLASS) to strengthen standardized AMR surveillance (World Health Organization, 2015; World Health Organization, 2022). However, AMR surveillance in SSA remains constrained by heterogeneous study designs, inconsistent antimicrobial testing and reporting, limited laboratory capacity, and incomplete geographical coverage (Tadesse et al., 2017; Kariuki & Dougan, 2014). These challenges are compounded by

antimicrobial consumption in human and agricultural settings and inadequate regulatory and stewardship capacity (van Boeckel et al., 2014; Ayukekbong et al., 2017).

*Escherichia coli* is an important sentinel organism for AMR surveillance because of its major role in urinary tract, bloodstream and gastrointestinal infections and its capacity to acquire and disseminate diverse resistance determinants through mobile genetic elements (Poirel et al., 2018; Partridge et al., 2018; Carattoli, 2013). Resistance determinants involving tetracycline, aminoglycoside and folate-pathway resistance have been widely described among *E. coli* and other Gram-negative bacteria (Wright, 2007; Chopra & Roberts, 2001; Ramirez & Tolmasky, 2010; Huovinen, 2001). Integrated phenotype–genotype datasets can therefore provide information that is not available from phenotypic surveillance or genomic data alone. The EMBL-EBI Antimicrobial Resistance Portal provides publicly accessible antimicrobial susceptibility and genomic resistance data, creating an opportunity to investigate resistance patterns and phenotype–genotype relationships across datasets that would otherwise be difficult to analyse collectively (Amid et al., 2020; Alcock et al., 2020). Machine-learning approaches have also been increasingly investigated for AMR prediction, particularly where genomic and epidemiological variables may contain complementary predictive information (Moradigaravand et al., 2018; Khaledi et al., 2020; Davis et al., 2016). However, evidence from linked phenotype–genotype surveillance datasets representing SSA remains limited, particularly for datasets characterized by geographical imbalance, relatively small numbers of isolates and high resistance prevalence.

Accordingly, this study examined linked phenotype–genotype records for *E. coli* available through the EMBL-EBI Antimicrobial Resistance Portal and representing seven Sub-Saharan African countries. The analysis focused on observed phenotypic resistance, multidrug resistance, AMR-associated gene diversity and co-occurrence, and concordance between detected resistance determinants and corresponding phenotypic susceptibility results. Country, specimen-source and temporal distributions were assessed to describe patterns within the available records, while comparison with WHO GLASS data was used to provide broader surveillance context. The study was designed as a descriptive secondary analysis of available linked surveillance records and does not seek to estimate population-level antimicrobial resistance prevalence for Sub-Saharan Africa.

#### Study objective

This study aimed to characterize phenotypic antimicrobial resistance and antimicrobial resistance-associated gene patterns among *Escherichia coli* isolates represented in linked phenotype–genotype records from seven Sub-Saharan African countries in the EMBL-EBI Antimicrobial Resistance Portal. Specifically, we assessed observed antimicrobial resistance and

multidrug resistance profiles, described variation by country and specimen source, characterized AMR gene diversity and co-occurrence, and evaluated concordance between detected AMR-associated genes and corresponding phenotypic resistance. Temporal patterns and comparison with WHO GLASS estimates were examined as contextual analyses rather than as estimates of regional prevalence or future resistance forecasts.

## **2. Methods**

### **2.1 Data Source and Study Design**

This study was a retrospective cross-sectional secondary analysis of publicly available antimicrobial resistance (AMR) surveillance data obtained from the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI) Pathogen Antimicrobial Resistance (AMR) Portal (<https://www.ebi.ac.uk/ena/pathogens/antimicrobial-resistance>). The EMBL-EBI AMR Portal links whole-genome sequencing (WGS) data archived in the European Nucleotide Archive (ENA) with accompanying antimicrobial susceptibility testing (AST) results submitted by surveillance programmes, research laboratories, and public health institutions, thereby enabling integrated analyses of phenotypic resistance and genomic AMR determinants (Amid et al., 2020). Genotypic resistance determinants were annotated using the Comprehensive Antibiotic Resistance Database (CARD) ontology and associated curation framework (Alcock et al., 2020).

The study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) recommendations for observational studies (von Elm et al., 2007). Because all data were fully anonymized and publicly available, institutional ethical approval and informed consent were not required.

Data retrieval was performed using the EMBL-EBI AMR Portal phenotype–genotype integrated query interface. The search was restricted to *Escherichia coli* isolates originating from Sub-Saharan African countries with available linked phenotypic AST and genomic AMR gene data. Although the search window encompassed isolates collected between 2000 and 2025, only records with complete linked phenotype–genotype information from 2008 to 2018 were available and therefore included in the final analytical dataset.

The initial query permitted all available specimen sources, AST methodologies, resistance categories, and antimicrobial agents available within the portal. AST methods included broth microdilution, agar dilution, disk diffusion, E-test, and other standardized susceptibility testing approaches recorded in the database. Phenotypic susceptibility categories included resistant (R), intermediate (I), susceptible (S), and susceptible-dose dependent (SDD). Following data cleaning

and linkage of phenotype and genotype records, only antibiotics with sufficient numbers of linked observations for statistical analysis were retained. The final dataset comprised 375 antibiotic-specific phenotype–genotype records derived from 111 unique *E. coli* isolates collected across seven Sub-Saharan African countries, representing the complete set of eligible linked records available within the portal at the time of data extraction.

To ensure reproducibility, duplicate BioSample identifiers, incomplete phenotype–genotype linkages, and records lacking essential metadata were excluded prior to analysis. BioSample accession numbers were used as the unique isolate identifier throughout all isolate-level analyses, whereas antibiotic-specific susceptibility records constituted the unit of analysis for phenotype-level statistical modelling.

## **2.2 Variables and Data Processing (Revised)**

The primary outcome variable was the phenotypic antimicrobial resistance status recorded for each antibiotic–isolate combination. For statistical modelling, susceptibility results were dichotomized into resistant (R) **and** non-resistant (S/I/SDD) categories, where susceptible (S), intermediate (I), and susceptible-dose dependent (SDD) interpretations were grouped as non-resistant. This binary classification was adopted to facilitate logistic regression and machine-learning analyses and is consistent with approaches commonly used in AMR surveillance studies (CLSI, 2023; EUCAST, 2023).

Independent variables extracted from the EMBL-EBI AMR Portal included collection year, country of origin, specimen source, antimicrobial agent tested, antimicrobial susceptibility testing (AST) method, and detected AMR genes. AMR genes were further grouped into functional resistance classes according to the Comprehensive Antibiotic Resistance Database (CARD) ontology (Alcock et al., 2020). Country, antibiotic identity, specimen source, AST method, and AMR class were treated as categorical variables.

The analytical unit depended on the objective of each analysis. Antibiotic-specific resistance, temporal trends, phenotype–genotype concordance, logistic regression, and machine-learning analyses were performed at the antibiotic-record level ( $n = 375$  phenotype–genotype records). In contrast, multidrug resistance profiling, AMR gene co-occurrence, and resistosome diversity analyses were conducted at the unique isolate level ( $n = 111$  BioSample accessions) to avoid repeated counting of the same isolate across multiple antibiotics.

BioSample accession numbers served as unique isolate identifiers throughout the study. Duplicate records arising from repeated testing of the same isolate–antibiotic combination were removed before analysis, and records with incomplete phenotype–genotype linkage or missing essential metadata were excluded. Continuous variables were inspected for completeness, while

categorical variables were standardized to ensure consistent naming across contributing datasets before statistical analysis.

Data management and preprocessing were performed in Python version 3.12 using the pandas library (version 2.0), with subsequent statistical analyses conducted as described below.

### **2.3 Temporal and Geospatial Analysis**

Annual antimicrobial resistance proportions were calculated for each antibiotic and for the pooled dataset by dividing the number of resistant records by the total number of antibiotic-specific records within each calendar year. Temporal trends were evaluated using ordinary least squares (OLS) linear regression, with annual resistance proportion as the dependent variable and collection year as the independent variable. Regression slope (expressed as percentage points per year), coefficient of determination ( $R^2$ ), and associated *p*-values were reported to describe the direction and strength of temporal trends.

Because the available surveillance data were unevenly distributed across calendar years and several years contained very small numbers of observations, temporal trend analyses were considered exploratory. Years represented by fewer than five records were retained for descriptive summaries but excluded from regression modelling to improve the stability of parameter estimates.

Resistance proportions were presented with 95% Wilson score confidence intervals (CIs), which provide robust interval estimates for binomial proportions, particularly when sample sizes are small (Wilson, 1927).

Geographical analyses were performed at the country level using unique BioSample identifiers. Overall resistance prevalence, multidrug resistance (MDR) prevalence, dominant antimicrobial resistance genes, and resistome diversity indices were calculated separately for each represented country. Countries represented by fewer than 10 isolates were retained for completeness but interpreted cautiously because of limited sample size. Spatial visualization of resistance burden was performed using choropleth maps with proportional symbols representing the number of unique isolates per country.

### **2.4 Multi-Drug Resistance Profiling and Co-Resistance**

Multidrug resistance (MDR) was defined as phenotypic non-susceptibility to at least one antimicrobial agent in three or more antimicrobial classes, consistent with the international consensus criteria proposed by Magiorakos AP et al. (Magiorakos et al., 2012). Resistance profiles were determined at the unique isolate level by aggregating all antimicrobial susceptibility test (AST) results associated with each BioSample accession.

Because the EMBL-EBI dataset contained susceptibility data for only seven antibiotic classes, the international definitions of extensively drug-resistant (XDR) and pandrug-resistant (PDR) organisms could not be applied reliably. For descriptive purposes, isolates resistant to five or more of the available antimicrobial classes were identified; however, these classifications are presented as operational categories rather than formal XDR or PDR designations and should not be interpreted according to the international consensus definitions.

For each isolate, the total number of resistant antimicrobial classes was calculated, and isolates were classified according to the number of resistant classes (0, 1, 2, 3, 4, or  $\geq 5$ ). The prevalence of MDR was calculated as the proportion of isolates resistant to three or more antimicrobial classes.

Phenotypic co-resistance between antibiotic pairs was evaluated at the antibiotic-record level. For each antibiotic pair, conditional co-resistance proportions were calculated as the proportion of resistant records for one antibiotic that also exhibited resistance to the second antibiotic. Pairwise co-resistance frequencies were summarized descriptively to identify the most common resistance combinations within the study population.

## **2.5 Phenotype–Genotype Concordance**

Agreement between detected antimicrobial resistance (AMR) genes and phenotypic antimicrobial susceptibility results was evaluated for each antibiotic using standard  $2 \times 2$  contingency table analyses. Genotypic resistance was defined as the presence of one or more CARD-annotated AMR genes relevant to the corresponding antibiotic class within the linked EMBL-EBI phenotype–genotype record, whereas phenotypic resistance was based on the recorded AST interpretation.

For each antibiotic, the following diagnostic performance measures were calculated: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and overall concordance. Cohen's  $\kappa$  coefficient (Cohen, 1960) was used to quantify agreement beyond chance and interpreted according to the conventional thresholds:  $\leq 0.20$ , slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and  $> 0.80$ , almost perfect agreement.

Concordance analyses were performed separately for each antibiotic because resistance mechanisms differ across antimicrobial classes. The analyses evaluated agreement between gene presence **and** observed resistance phenotype and did not incorporate chromosomal point mutations, promoter mutations, gene copy number, or gene expression data. Consequently, genotype–phenotype concordance estimates should be interpreted as measures of agreement between detected resistance genes and phenotypic susceptibility rather than comprehensive assessments of all genetic resistance mechanisms (Ellington et al., 2017; Doyle et al., 2020).

## 2.7 Resistome Diversity Analysis

The diversity of antimicrobial resistance (AMR) genes (resistome) was characterised at the country level using Shannon's diversity index ( $H'$ ) and Pielou's evenness index ( $J'$ ). Shannon's diversity index was calculated as:

$$H' = - \sum_{i=1}^S p_i \ln(p_i)$$

where  $p_i$  is the relative frequency of the  $i$ th AMR gene and  $S$  is the total number of unique genes detected (Shannon, 1948). Pielou's evenness was calculated as:

$$J' = \frac{H'}{\ln(S)}$$

where values approaching 1 indicate a more even distribution of genes across the resistome (Pielou, 1966).

AMR gene annotations followed the Comprehensive Antibiotic Resistance Database (CARD) ontology (Alcock et al., 2023). Gene frequencies were summarised across all isolates and stratified by country to identify dominant resistance determinants and overall resistome composition. Diversity indices were interpreted descriptively to compare the complexity of AMR gene distributions between countries represented in the dataset.

## 2.8 Multivariable Logistic Regression and Machine Learning

A multivariable logistic regression model was fitted to examine associations between phenotypic antimicrobial resistance (resistant versus non-resistant) and six explanatory variables: collection year, country, antibiotic identity, isolation source, antimicrobial susceptibility testing (AST) method, and AMR gene class. Categorical variables were label encoded prior to analysis. Model performance was evaluated using five-fold stratified cross-validation, and mean accuracy and area under the receiver operating characteristic curve (ROC-AUC) were reported across folds. Regression coefficients ( $\beta$ ), odds ratios ( $OR = \exp[\beta]$ ), and 95% confidence intervals were calculated to describe model-derived associations. Because the predictors were predominantly categorical and label encoded, coefficient estimates were interpreted as descriptive associations rather than causal epidemiological effects.

To explore potential non-linear relationships, a Random Forest classifier (Breiman, 2001) comprising 200 decision trees was trained using a stratified 75:25 training and test split with balanced class weights to account for class imbalance. Model performance was assessed on the independent test dataset using accuracy, ROC-AUC, precision, recall, and F1-score. Variable importance was estimated using the mean decrease in Gini impurity across all trees. Both logistic

regression and Random Forest analyses were implemented using scikit-learn version 1.3 (Pedregosa et al., 2011) with `random_state = 42` to ensure reproducibility.

Given the modest sample size, the predominance of gastrointestinal isolates, and the inclusion of multiple antibiotic-specific observations from some isolates, the machine learning analyses were intended as exploratory predictive modelling rather than development of clinically deployable prediction algorithms.

## **2.9 Exploratory Resistance Trend Extrapolation**

Exploratory projections of antibiotic resistance to 2030 were generated using ordinary least squares (OLS) linear regression fitted to annual resistance proportions observed between 2008 and 2018. Only antibiotics with at least three annual observations were included. Regression slopes (percentage points per year), coefficients of determination ( $R^2$ ), and p-values were calculated to describe temporal trends within the available data.

Predicted resistance proportions for 2025, 2027, and 2030 were obtained by linear extrapolation from the fitted models. Ninety-five percent prediction intervals were calculated using the  $t$ -distribution with degrees of freedom equal to  $n - 2$ , where  $n$  represents the number of annual observations included in each regression model. Predicted values were constrained to the biologically plausible range of 0–100% resistance.

Because annual observations were sparse, unevenly distributed across the study period, and derived from a retrospective surveillance dataset, these projections were considered exploratory and were intended only to illustrate possible future trajectories under the assumption that historical linear trends remain unchanged. The forecasts should therefore not be interpreted as definitive epidemiological predictions or as evidence of future resistance prevalence without continued surveillance.

## **2.10 Stratified Pooled Resistance Analysis**

Resistance proportions were summarized across source  $\times$  antibiotic strata containing at least three observations to provide a descriptive comparison of resistance patterns within the study dataset. For each stratum, resistance proportions and corresponding 95% Wilson score confidence intervals were calculated.

An inverse-variance weighting approach was used to compute a pooled descriptive resistance estimate across eligible strata. The variance of each proportion was estimated as:

$$Var(p_i) = \frac{p_i(1 - p_i)}{n_i}$$

where  $p$  represents the observed resistance proportion and  $n$  the number of records in each stratum. To avoid numerical instability arising from strata with proportions of 0% or 100%, a minimum variance floor of 0.001 was applied before calculating inverse-variance weights. Between-stratum variability was assessed using Cochran's Q statistic (Cochran, 1954) and quantified using the  $I^2$  statistic (Higgins et al., 2003). Because all strata originated from a single surveillance dataset rather than independent studies, the pooled estimate was interpreted as a descriptive summary of within-dataset heterogeneity rather than a conventional meta-analytic effect size.

## 2.11 Comparison with WHO GLASS Surveillance Data

Study resistance proportions were descriptively compared with corresponding *Escherichia coli* resistance estimates reported in the WHO Global Antimicrobial Resistance and Use Surveillance System (GLASS) Report 2022 (World Health Organization, 2022). Comparisons were limited to antibiotics that were available in both datasets. Differences between the study and GLASS Africa estimates were expressed as percentage-point (pp) differences, calculated as:

$$\Delta = \text{Study resistance (\%)} - \text{GLASS Africa resistance (\%)}$$

The comparison was intended solely to provide international surveillance context and was not considered a formal statistical comparison. No hypothesis testing was performed because of substantial differences between the two datasets, including the surveillance period (2008–2018 versus predominantly 2020–2021), specimen composition (predominantly gastrointestinal isolates in the present study versus mainly blood and urinary tract isolates in GLASS), country representation, laboratory methodologies, and sampling frameworks. Consequently, observed differences were interpreted as descriptive rather than evidence of true epidemiological differences in antimicrobial resistance prevalence.

## 2.12 Statistical Analysis and Reproducibility

All statistical analyses were performed in Python version 3.12. Data management and preprocessing were conducted using pandas version 2.0 and NumPy version 1.25. Statistical analyses were performed using SciPy version 1.11, and machine learning analyses were implemented using scikit-learn version 1.3 (Pedregosa et al., 2011). Network visualization was generated using NetworkX version 3.6, and figures were produced using Matplotlib version 3.8 (Hunter, 2007).

Resistance proportions were summarized as percentages with 95% Wilson score confidence intervals. Associations between categorical variables were assessed using Pearson's chi-square test or Fisher's exact test, as appropriate. Ordinary least squares linear regression was used to examine temporal trends and to generate exploratory linear extrapolations of resistance patterns.

Phenotype–genotype agreement was evaluated using sensitivity, specificity, positive predictive value, negative predictive value, overall concordance, and Cohen's  $\kappa$  coefficient. AMR gene co-occurrence was quantified using the Jaccard similarity index, odds ratios, and Fisher's exact test. Resistosome diversity was assessed using Shannon's diversity index and Pielou's evenness index. Multivariable logistic regression and Random Forest classification were performed using the procedures described above. Logistic regression performance was evaluated using five-fold stratified cross-validation, whereas Random Forest performance was assessed using an independent stratified test dataset. Statistical significance was defined as a two-sided  $p < 0.05$  unless otherwise specified.

To enhance reproducibility, a fixed random seed (`random_state = 42`) was used for all machine learning analyses involving random sampling. The analytical workflow and code are available from the corresponding author upon reasonable request.

All figures were generated at 600 DPI in TIFF and vector PDF formats. The significance threshold was  $\alpha = 0.05$  throughout. Complete analysis code is available from the corresponding author upon reasonable request.

### **3. Results**

#### **3.1 Dataset Overview and Study Population**

The final dataset comprised 375 linked phenotype–genotype antimicrobial susceptibility records representing 111 unique *Escherichia coli* isolates collected from seven Sub-Saharan African countries between 2008 and 2018 (Table 1). Because each isolate could undergo susceptibility testing against multiple antibiotics, individual isolates contributed more than one antibiotic-specific record.

Kenya (178 records, 47.5%) and Mali (96, 25.6%) accounted for 73.1% of all records. Most antimicrobial susceptibility testing was performed by broth dilution (342/375, 91.2%), with the remainder using disk diffusion (33/375, 8.8%). Gastrointestinal specimens predominated (362/375, 96.5%), followed by blood (8, 2.1%) and respiratory tract (5, 1.3%). Streptomycin (173 records, 46.1%) and tetracycline (88, 23.5%) were the most frequently tested antibiotics.

Overall phenotypic resistance was 82.4% (309/375; 95% CI: 78.2–86.2%), while 76.6% of isolates (85/111) were resistant to at least one antibiotic. Genotypic analysis identified 24 AMR-associated genes belonging to six resistance classes, dominated by aminoglycoside (47.7%), tetracycline (21.6%), and trimethoprim (18.4%) resistance determinants. A summary of the dataset is presented in Table 1.

**Table 1. Study Dataset Characteristics EMBL-EBI AMR Portal, *E. coli*, Sub-Saharan Africa (2008–2018)**

| Characteristic                              | Value / Detail                                                                                        |
|---------------------------------------------|-------------------------------------------------------------------------------------------------------|
| <b>Data source</b>                          | EMBL-EBI Antimicrobial Resistance Portal<br>(Release 2020-12)                                         |
| <b>Query type</b>                           | Combined phenotype & genotype (Query #74)                                                             |
| <b>Species / Genus</b>                      | Escherichia coli                                                                                      |
| <b>Geographic region</b>                    | Sub-Saharan Africa                                                                                    |
| <b>Countries included (n = 7)</b>           | Botswana, Gambia, Kenya, Mali, Mozambique,<br>South Africa, United Republic of Tanzania               |
| <b>Collection period</b>                    | 2008–2018 (8 unique years)                                                                            |
| <b>Total records (phenotype-genotype)</b>   | 375                                                                                                   |
| <b>Unique isolates (BioSample IDs)</b>      | 111                                                                                                   |
| <b>Antibiotics tested (n = 7)</b>           | Ampicillin, azithromycin, chloramphenicol,<br>gentamicin, streptomycin, tetracycline,<br>trimethoprim |
| <b>AMR testing methods</b>                  | Broth dilution: n = 342 (91.2%); Disk diffusion: n<br>= 33 (8.8%)                                     |
| <b>Isolation source categories</b>          | Gastrointestinal system: n = 362 (96.5%); Blood:<br>n = 8 (2.1%); Respiratory tract: n = 5 (1.3%)     |
| <b>Overall phenotypic resistance rate</b>   | 82.4% (309/375 records); 95% CI: 78.2–86.2%                                                           |
| <b>Resistant isolates</b>                   | 85/111 (76.6%)                                                                                        |
| <b>Unique AMR resistance genes detected</b> | 24                                                                                                    |
| <b>AMR classes represented</b>              | 6: Aminoglycoside, Tetracycline, Trimethoprim,<br>Multidrug, Phenicol, Macrolide                      |
| <b>Resistance phenotype categories</b>      | Resistant, intermediate, susceptible (analysis:<br>resistant vs. other)                               |

AMR, antimicrobial resistance; AST, antimicrobial susceptibility testing; EMBL-EBI, European Molecular Biology Laboratory-European Bioinformatics Institute; 95% CI, 95% confidence interval (Wilson score method). Resistance rate calculated at record level (one record = one isolate × one antibiotic × one AST result).

### 3.2 Temporal Trends in Antimicrobial Resistance (2008–2018)

Annual resistance rates across the study cohort were consistently high throughout the observation period, with an overall record-level trajectory that was statistically non-significant (slope = −1.30 percentage points [pp]/year;  $R^2 = 0.025$ ;  $p = 0.706$ ) (Table 2; Table 3). The highest overall annual

resistance rates were observed in 2012 (100.0%, n = 13 records, n = 1 isolate) and 2010 (90.0%, n = 30 records), while the lowest was recorded in 2013 (25.0%, n = 4 records), although these years were characterised by markedly limited sample sizes and heavily skewed antibiotic representation, which preclude meaningful inter-year comparisons for those periods. The most data-rich years 2008 (n = 131 records, 37 isolates; R = 84.7%) and 2009 (n = 181 records, 53 isolates; R = 81.8%) yielded the most representative estimates, together encompassing 83.5% of all study records.

At the antibiotic level, trimethoprim demonstrated the highest resistance prevalence (94.2%, 95% CI: 86.0–97.7%; n = 69), followed by tetracycline (93.2%, 95% CI: 85.9–96.8%; n = 88). Gentamicin achieved 100% resistance across all six tested records (95% CI: 61.0–100.0%), though the minuscule sample size renders this observation indicative rather than conclusive. Streptomycin, with the largest available dataset for trend analysis (n = 173 records, 8 years), exhibited the strongest positive temporal trajectory (+5.46 pp/year,  $R^2 = 0.731$ ), though the Mann-Kendall (MK) test was non-significant at  $\alpha = 0.05$  (MK p = 0.734), likely reflecting the limited number of annual data points (n = 8 years) available for non-parametric testing. Chloramphenicol had the lowest mean resistance rate (42.9%, 95% CI: 26.5–60.9%), with a declining trend (−6.31 pp/year) that, while suggestive, did not reach significance (MK p = 0.548). Resistance remained consistently high across the available sampling years; however, the highly uneven temporal distribution and sparse sampling after 2010 substantially limit inference regarding long-term resistance trends. Annual resistance patterns are presented in Figure 1, while regression statistics and trend estimates are summarised in Table 2. Annual resistance proportions are provided in **Supplementary Table S1**.

**Table 2. Per-Antibiotic Resistance Rates and Temporal Trend Analysis**

| Antibiotic      | N<br>Records | N<br>Resistant | Resistance %<br>(95% CI) | Slope<br>(%/yr) | R <sup>2</sup> | MK p-<br>value |
|-----------------|--------------|----------------|--------------------------|-----------------|----------------|----------------|
| Ampicillin      | 9            | 6              | 66.7 (35.4–87.9)         | −6.41           | 0.142          | 0.497          |
| Azithromycin    | 2            | 2              | 100.0 (34.2–100.0)       |                 |                |                |
| Chloramphenicol | 28           | 12             | 42.9 (26.5–60.9)         | −6.31           | 0.149          | 0.548          |
| Gentamicin      | 6            | 6              | 100.0 (61.0–100.0)       |                 |                |                |
| Streptomycin    | 173          | 136            | 78.6 (71.9–84.1)         | +5.46           | 0.731          | 0.734          |
| Tetracycline    | 88           | 82             | 93.2 (85.9–96.8)         | +0.73           | 0.450          | 0.142          |
| Trimethoprim    | 69           | 65             | 94.2 (86.0–97.7)         | +2.08           | 0.293          | 1.000          |
| All antibiotics | 375          | 309            | 82.4 (78.2–86.2)         | −1.30           | 0.025          | 0.706          |

Resistance rates with 95% Wilson score confidence intervals. Slope: percentage point change per calendar year from ordinary least squares linear regression on annual resistance proportions.  $R^2$ , coefficient of determination. MK, Mann-Kendall non-parametric trend test (normal approximation, two-tailed). Analyses restricted to antibiotics with  $\geq 3$  annual time-points. , insufficient temporal data for trend modelling.

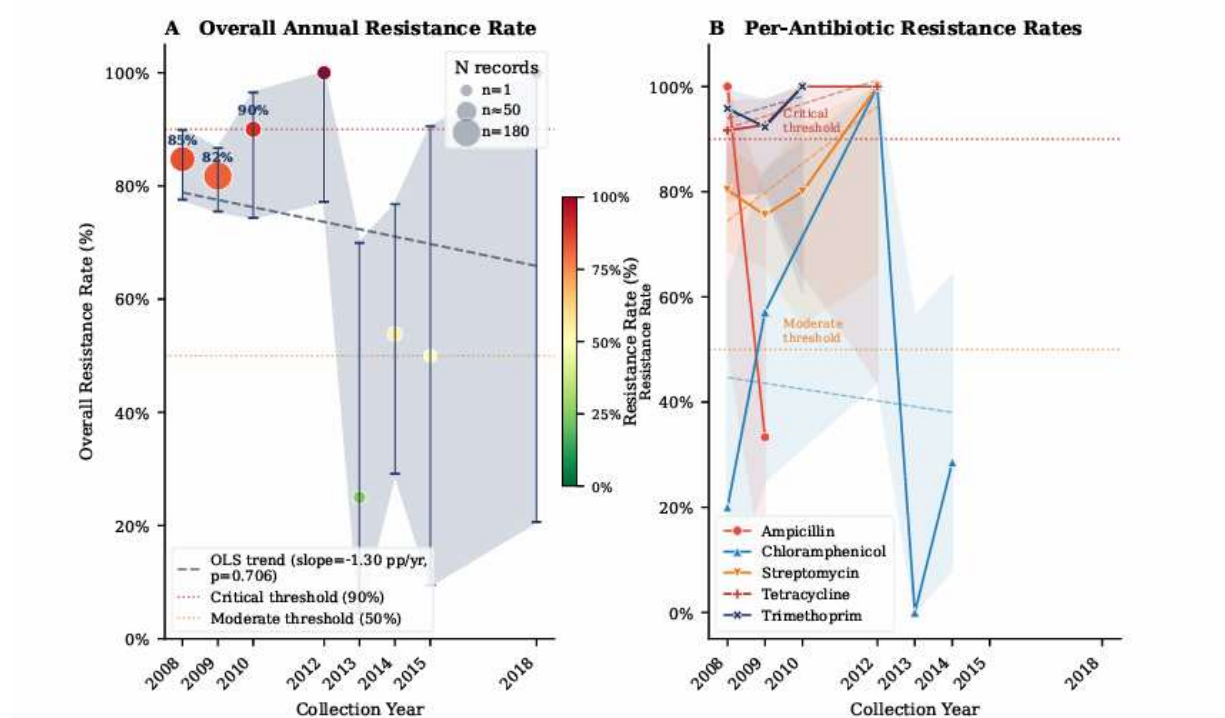


**Figure 1.** Temporal trends in annual overall and per-antibiotic resistance rates in *E. coli* from Sub-Saharan Africa (2008–2018). (A) Overall record-level resistance rate by year. Error bars represent 95% Wilson score confidence intervals. The fitted OLS regression line (slope = -1.30 pp/year,  $p = 0.706$ ) is shown as a dashed line. (B) Per-antibiotic annual resistance rates. Coloured lines represent individual antibiotics. Open symbols indicate years with  $n < 10$  records for that antibiotic. Antibiotics with  $< 3$  annual time-points (azithromycin, gentamicin) are omitted. Data years: 2008, 2009, 2010, 2012, 2013, 2014, 2015, 2018. Note: data are not continuous across all years.

**Table 3. Annual Record Distribution and Overall Resistance Rate (2008–2018)**

| Year | N Records | N Isolates | Overall Resistance (%) | Data Completeness Note                         |
|------|-----------|------------|------------------------|------------------------------------------------|
| 2008 | 131       | 37         | 84.7                   | Largest annual contribution (34.9% of records) |
| 2009 | 181       | 53         | 81.8                   | Largest isolate cohort (47.7% of isolates)     |

|             |    |   |       |                                                  |
|-------------|----|---|-------|--------------------------------------------------|
| <b>2010</b> | 30 | 9 | 90.0  | Moderate sample; no azithromycin/gentamicin      |
| <b>2012</b> | 13 | 1 | 100.0 | Single isolate; interpretation limited           |
| <b>2013</b> | 4  | 3 | 25.0  | Very small sample (n = 4); caution advised       |
| <b>2014</b> | 13 | 6 | 53.8  | Small sample; mixed antibiotic coverage          |
| <b>2015</b> | 2  | 1 | 50.0  | Extremely small; not suitable for trend analysis |
| <b>2018</b> | 1  | 1 | 100.0 | Single record; excluded from trend modelling     |

Years with  $n < 5$  records (2013, 2015, and 2018) are retained for completeness but excluded from temporal trend regression modelling due to insufficient precision. Data completeness notes indicate antibiotic coverage limitations within each calendar year.

### 3.3 Geospatial Distribution of AMR Burden across Sub-Saharan Africa

Substantial between-country heterogeneity in overall resistance burden was observed across the seven represented countries (Table 4). Among countries with  $\geq 10$  isolates, the Gambia recorded the highest resistance rate (87.1%), followed by Mali (86.5%), Mozambique (86.5%), and Kenya (80.9%). The United Republic of Tanzania demonstrated markedly lower resistance (47.4%,  $n = 10$  isolates), representing a potential epidemiological outlier.

Multi-drug resistance (MDR; phenotypic resistance to  $\geq 3$  antibiotic classes at the isolate level) was highest in Mozambique (70.0%), the Gambia (60.0%), and Mali (40.0%). Kenya, the most data-rich country, had an MDR rate of 36.7% of its 49 unique isolates. Tanzania, in contrast, recorded 0% MDR in this dataset, consistent with its lower overall resistance burden. Country-level gene distribution analysis revealed that *tetA* dominated the AMR gene landscape in five of seven countries (Botswana, Gambia, Kenya, Mali, South Africa), while *neo* was dominant in Mozambique and *cat* in Tanzania. Country-specific resistance estimates and MDR prevalence are summarised in Table 4 and geographically illustrated in Figure 2.

**Table 4. Country-Level Antimicrobial Resistance Burden, MDR Rates, and Gene Ecology**

| Country             | N<br>Records | N<br>Isolates | Overall<br>R% | MDR% ( $\geq 3$<br>classes) | Top<br>Gene   | Shannon<br>H' | J'<br>Evenness |
|---------------------|--------------|---------------|---------------|-----------------------------|---------------|---------------|----------------|
| <b>Botswana</b>     | 13           | 1             | 100.0         | 100.0                       | <i>tetA_1</i> | 2.303         | 1.000          |
| <b>Gambia</b>       | 31           | 10            | 87.1          | 60.0                        | <i>tetA</i>   | 1.424         | 0.885          |
| <b>Kenya</b>        | 178          | 49            | 80.9          | 36.7                        | <i>tetA</i>   | 1.642         | 0.789          |
| <b>Mali</b>         | 96           | 30            | 86.5          | 40.0                        | <i>tetA</i>   | 1.751         | 0.842          |
| <b>Mozambique</b>   | 37           | 10            | 86.5          | 70.0                        | <i>neo</i>    | 1.357         | 0.979          |
| <b>South Africa</b> | 1            | 1             | 100.0         | 0.0                         | <i>tetA</i>   |               |                |

|                               |    |    |      |     |     |       |       |
|-------------------------------|----|----|------|-----|-----|-------|-------|
| <b>United R.<br/>Tanzania</b> | 19 | 10 | 47.4 | 0.0 | cat | 1.749 | 0.841 |
|-------------------------------|----|----|------|-----|-----|-------|-------|

MDR, multi-drug resistance (resistance to  $\geq 3$  distinct antibiotic classes per isolate). Shannon  $H'$ , entropy index of AMR gene distribution within-country. Evenness  $J' = H'/\ln(S)$ . Countries with a single isolate (Botswana, South Africa) should be interpreted with extreme caution. Tanzania's lower MDR rate is consistent with its overall lower resistance burden and chloramphenicol-dominant antibiotic profile.

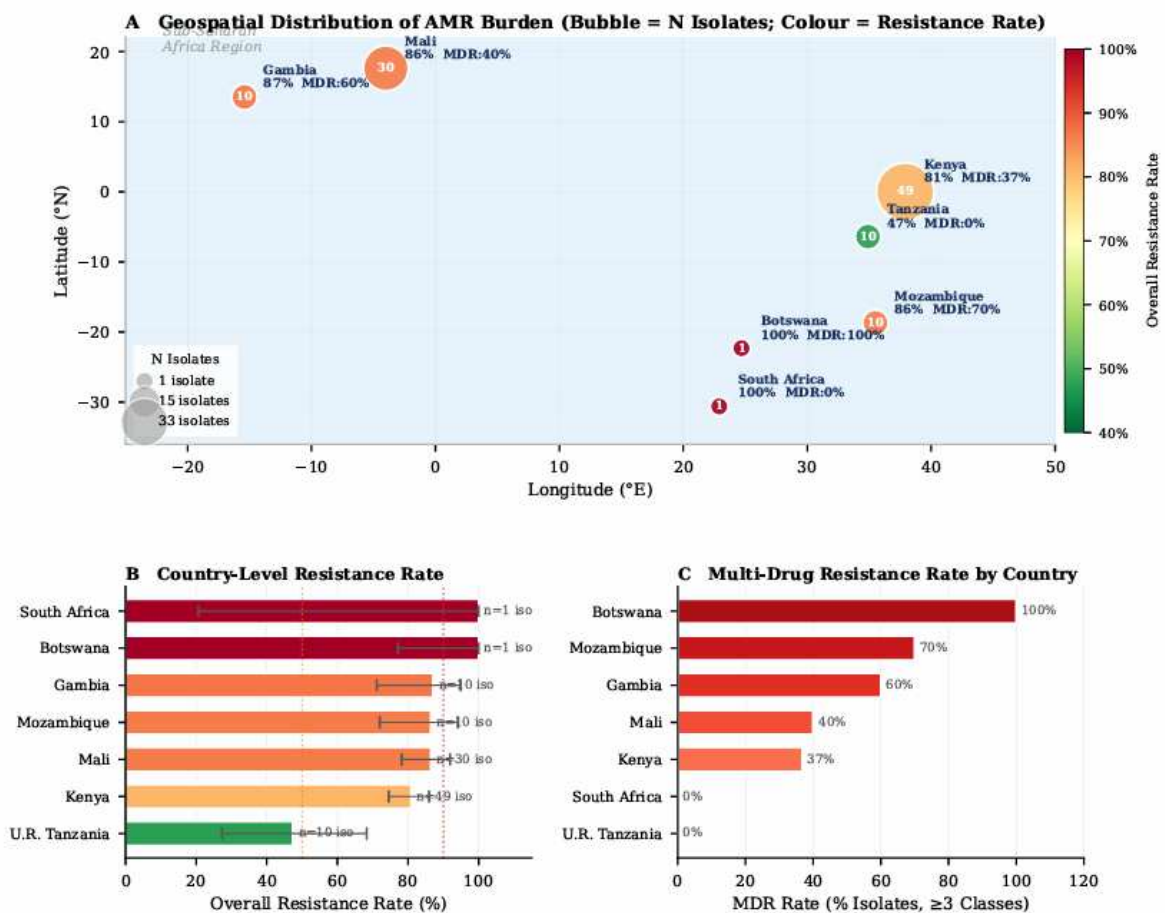


**Figure 2.** Geospatial distribution of overall resistance rates and MDR burden in *E. coli* across Sub-Saharan Africa (2008–2018). Choropleth map of Africa showing seven study countries, colour-coded by overall resistance rate ( $n \geq 10$  isolates: green  $< 60\%$ ; amber  $60\text{--}80\%$ ; red  $> 80\%$ ). Bubble size indicates number of unique isolates. MDR rate (%) labelled within each bubble. Countries with  $n = 1$  isolate (Botswana, South Africa) indicated with hatched shading. Inset bar chart: per-country resistance rates with 95% Wilson score confidence intervals.

### 3.4 Multi-Drug Resistance Profiling and Co-Resistance Patterns

Of the 111 unique isolates, 39.6% (n = 44) fulfilled the operational MDR definition used in this study (resistance to  $\geq 3$  antibiotic classes), with an additional 2.7% (n = 3) demonstrating resistance to four classes simultaneously (Table 5). No extensively drug-resistant (XDR;  $\geq 5$  classes) isolates were identified. The modal phenotypic resistance profile was three-class resistance (36.9%), followed by two-class (29.7%) and single-class resistance (20.7%). Fully susceptible isolates were rare, comprising only 9.9% (n = 11) of the cohort a fraction that itself may partly reflect ascertainment bias given that isolates with susceptible phenotypes may be systematically under-collected or under-sequenced in clinical surveillance pipelines operating in resource-constrained settings.

Co-resistance analysis at the record level revealed the streptomycin–tetracycline combination as the most prevalent antibiotic pair: 97.6% of tetracycline-resistant records co-exhibited streptomycin resistance (n = 80 of 82 records), and 58.8% of streptomycin-resistant records co-exhibited tetracycline resistance (n = 80 of 136). Tetracycline–trimethoprim co-resistance was present in 59.7% of tetracycline-resistant records (n = 49 of 82), and trimethoprim–tetracycline co-resistance in 75.4% of trimethoprim-resistant records (n = 49 of 65). Ampicillin–streptomycin co-resistance was identified in 66.7% of ampicillin-resistant records (n = 4 of 6). The observed streptomycin–tetracycline–trimethoprim co-resistance pattern is compatible with previously reported co-occurrence of these resistance determinants in mobile genetic elements, where these determinants are commonly embedded as single genetic units. The distribution of resistance categories across isolates is summarised in Table 5.

**Table 5. Distribution of Operational category ( $\geq 5$  resistant classes) Profiles Among *E. coli* Isolates (n = 111)**

| <b>N Drug Classes Resistant</b>        | <b>N Isolates</b> | <b>% All Isolates</b>                               | <b>Cumulative %</b> | <b>Classification</b>                                      |
|----------------------------------------|-------------------|-----------------------------------------------------|---------------------|------------------------------------------------------------|
| <b>0</b>                               | 11                | 9.9                                                 | 9.9                 | Fully susceptible                                          |
| <b>1</b>                               | 23                | 20.7                                                | 30.6                | Single-drug resistant (SDR)                                |
| <b>2</b>                               | 33                | 29.7                                                | 60.4                | Double-drug resistant (DDR)                                |
| <b>3</b>                               | 41                | 36.9                                                | 97.3                | Multi-drug resistant (MDR)                                 |
| <b>4</b>                               | 3                 | 2.7                                                 | 100.0               | Multi-drug resistant (MDR)                                 |
| <b>5</b>                               | 0                 | 0.0                                                 | 100.0               | Extensively drug-resistant (XDR)                           |
| <b>MDR total (<math>\geq 3</math>)</b> | <b>44</b>         | <b>39.6</b> 95% CI<br>$\approx 30.9\text{--}49.0\%$ |                     | Forty-four of 111 isolates (39.6%) met the MDR definition. |

Resistance counts computed per unique isolate (BioSample ID) based on resistance calls across all antibiotic-specific tests. MDR: phenotypic resistance to  $\geq 3$  drug classes. XDR: resistance to  $\geq 5$  drug classes. Cumulative % reflects the proportion of isolates with  $\leq$  the stated drug class count resistant. Percentages may not sum to 100% due to rounding.

### 3.5 Source-Specific Antimicrobial Resistance Profiling

Isolation source was significantly associated with resistance phenotype ( $\chi^2 = 13.99$ ,  $df = 2$ ,  $p = 0.0009$ ). Gastrointestinal isolates constituted the majority of the dataset (362/375 records; 105 isolates) and exhibited the highest overall resistance rate (83.4%). Blood-source isolates (8 records; 3 isolates) had a resistance rate of 75.0%, whereas respiratory tract isolates (5 records; 3 isolates) showed the lowest observed resistance rate (20.0%) (Table S2).

Within gastrointestinal isolates, resistance to streptomycin and trimethoprim was 78.6% and 94.2%, respectively. Interpretation of resistance estimates for blood and respiratory tract isolates should be made cautiously because these strata contained very few records, and respiratory isolates were tested exclusively against chloramphenicol. Source-stratified resistance profiles are presented in Supplementary Table 2. Source-specific resistance rates are summarised in Table 5, with detailed source-by-country results presented in Supplementary Table S2. Supplementary visualizations of source-specific resistance profiles, logistic regression coefficients, and AMR class distributions are shown in Supplementary Figure S4.

### 3.6 Phenotype–Genotype Concordance Analysis

Agreement between AMR gene presence (genotype) and phenotypic resistance was evaluated across the seven antibiotics using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), overall concordance, and Cohen's  $\kappa$  (Table 6). Cohen's  $\kappa$  ranged from  $-0.024$  to  $0.240$ , corresponding to slight ( $\kappa < 0.20$ ) to fair ( $0.21$ – $0.40$ ) agreement.

Streptomycin exhibited the highest agreement ( $\kappa = 0.240$ ), with a sensitivity of 50.0%, specificity of 89.2%, PPV of 94.4%, NPV of 32.7%, and overall concordance of 58.4%. Among the 136 phenotypically resistant streptomycin records, 68 lacked detection of the queried resistance genes (FN = 68), whereas four susceptible records were gene-positive (FP = 4).

For tetracycline, all 82 resistant records carried *tetA*, yielding 100% sensitivity, whereas all six susceptible records also carried the gene, resulting in 0% specificity and  $\kappa = 0.000$ . Similarly, trimethoprim demonstrated 98.5% sensitivity but 0% specificity ( $\kappa = -0.024$ ), with 64 true positives, four false positives, one false negative, and no true negatives. Ampicillin and chloramphenicol also showed 100% sensitivity but no true-negative observations, while

azithromycin and gentamicin each demonstrated complete agreement based on the limited number of available records (Table 6).

Overall, phenotype–genotype concordance remained low across all antibiotics, with no antibiotic achieving moderate ( $\kappa \geq 0.41$ ) or higher agreement. The complete concordance metrics for each antibiotic are presented in Table 6. Additional visualizations of resistosome diversity and phenotype–genotype concordance are provided in **Supplementary Figure S3**.

**Table 6. Phenotype–Genotype Concordance Metrics by Antibiotic**

| Antibiotic             | N   | TP | FP | FN | TN | Sensitivity | Specificity | PPV    | NPV   | Concordance | Cohen's $\kappa$ |
|------------------------|-----|----|----|----|----|-------------|-------------|--------|-------|-------------|------------------|
| <b>Ampicillin</b>      | 9   | 6  | 3  | 0  | 0  | 100.0%      | 0.0%        | 66.7%  |       | 66.7%       | 0.000            |
| <b>Azithromycin</b>    | 2   | 2  | 0  | 0  | 0  | 100.0%      |             | 100.0% |       | 100.0%      | 0.000            |
| <b>Chloramphenicol</b> | 28  | 12 | 16 | 0  | 0  | 100.0%      | 0.0%        | 42.9%  |       | 42.9%       | 0.000            |
| <b>Gentamicin</b>      | 6   | 6  | 0  | 0  | 0  | 100.0%      |             | 100.0% |       | 100.0%      | 0.000            |
| <b>Streptomycin</b>    | 173 | 68 | 4  | 68 | 33 | 50.0%       | 89.2%       | 94.4%  | 32.7% | 58.4%       | 0.240            |
| <b>Tetracycline</b>    | 88  | 82 | 6  | 0  | 0  | 100.0%      | 0.0%        | 93.2%  |       | 93.2%       | 0.000            |
| <b>Trimethoprim</b>    | 69  | 64 | 4  | 1  | 0  | 98.5%       | 0.0%        | 94.1%  | 0.0%  | 92.8%       | −0.024           |

TP, true positive (gene detected, resistant phenotype); FP, false positive (gene detected, susceptible phenotype); FN, false negative (gene absent, resistant phenotype); TN, true negative (gene absent, susceptible phenotype). Sensitivity =  $TP/(TP+FN)$ ; Specificity =  $TN/(TN+FP)$ ; PPV =  $TP/(TP+FP)$ ; NPV =  $TN/(TN+FN)$ ; Concordance =  $(TP+TN)/N$ . Cohen's  $\kappa = (p_o - p_e)/(1 - p_e)$ .  $\kappa$  interpretation:  $\leq 0.20$ , slight; 0.21–0.40, fair; 0.41–0.60, moderate;  $\geq 0.61$ , substantial. indicates zero denominator

### 3.7 AMR Gene Co-occurrence and Network Analysis

Pairwise AMR gene co-occurrence was analysed at the isolate level across the 12 most frequently detected genes (Table 7). The strongest co-occurrence association was identified between *tetA* and *neo* (Jaccard index = 0.714; OR = 15.00; Fisher's exact  $p < 0.0001$ ; N = 60 co-occurring isolates), indicating that these two genes encoding tetracycline efflux and aminoglycoside 3'-phosphotransferase resistance, respectively are co-harboured at very high frequency in this *E. coli* population. The Jaccard index of 0.714 implies that 71.4% of isolates carrying either *tetA* or *neo* carry both, this co-occurrence pattern is consistent with previously reported plasmid-associated resistance clusters.

The second and third strongest associations were *neo-folA\_2* (Jaccard = 0.358; OR = 25.14;  $p < 0.0001$ ; N = 24) and *tetA-folA\_2* (Jaccard = 0.304; OR = 14.22;  $p = 0.0009$ ; N = 24), revealing a three-gene co-occurrence network (*tetA-neo-folA\_2*) that links tetracycline, aminoglycoside, and folate pathway resistance suggesting frequent co-occurrence of these resistance determinants.. Notably, *neo-dhfrI* showed a significant negative association (Jaccard = 0.136; OR = 0.23;  $p = 0.0008$ ), suggesting that *neo* and *dhfrI* tend to occupy distinct plasmid backgrounds despite both conferring aminoglycoside and trimethoprim resistance respectively. The *cat-yokD* pair showed a strong significant association (Jaccard = 0.214; OR > Not estimable because of sparse cell counts;  $p = 0.0016$ ), consistent with the co-location of chloramphenicol resistance and MarR-family transcriptional regulators on integron structures. Detailed co-occurrence data are provided in Table 7; co-occurrence matrix across all 12 gene pairs is provided in Supplementary Table 3. P-values should be interpreted in the context of multiple pairwise comparisons, and the co-occurrence network is illustrated in **Figure 3**.overall

**Table 7. Pairwise AMR Gene Co-occurrence Analysis (Top Gene Pairs, N ≥ 3 Co-occurrences)**

| Gene Pair                   | N Co-occurring Isolates | Jaccard Index | Odds Ratio | Fisher p | Strength      |
|-----------------------------|-------------------------|---------------|------------|----------|---------------|
| <i>tetA</i> – <i>neo</i>    | 60                      | 0.714         | 15.00      | <0.0001  | Strong        |
| <i>neo</i> – <i>folA_2</i>  | 24                      | 0.358         | 25.14      | <0.0001  | Strong        |
| <i>tetA</i> – <i>folA_2</i> | 24                      | 0.304         | 14.22      | 0.0009   | Moderate      |
| <i>tetA</i> – <i>dhfrI</i>  | 22                      | 0.244         | 0.69       | 0.4996   | Weak negative |
| <i>neo</i> – <i>dhfrI</i>   | 12                      | 0.136         | 0.23       | 0.0008   | Weak negative |
| <i>cat</i> – <i>yokD</i>    | 3                       | 0.214         | >999       | 0.0016   | Moderate      |

Co-occurrence computed at the isolate level. Jaccard index  $J = |A \cap B| / |A \cup B|$ ; ranges 0 (no co-occurrence) to 1 (perfect co-occurrence). OR, odds ratio  $[|A \cap B| \times |\text{neither}| / |A \text{ only}| \times |B \text{ only}|]$ ; two-tailed Fisher's exact p-value. Association strength: Jaccard  $\geq 0.5$  = strong; 0.2–0.49 = moderate;  $< 0.2$  = weak. Negative ORs  $< 1$  indicate mutually exclusive co-occurrence (negative association).

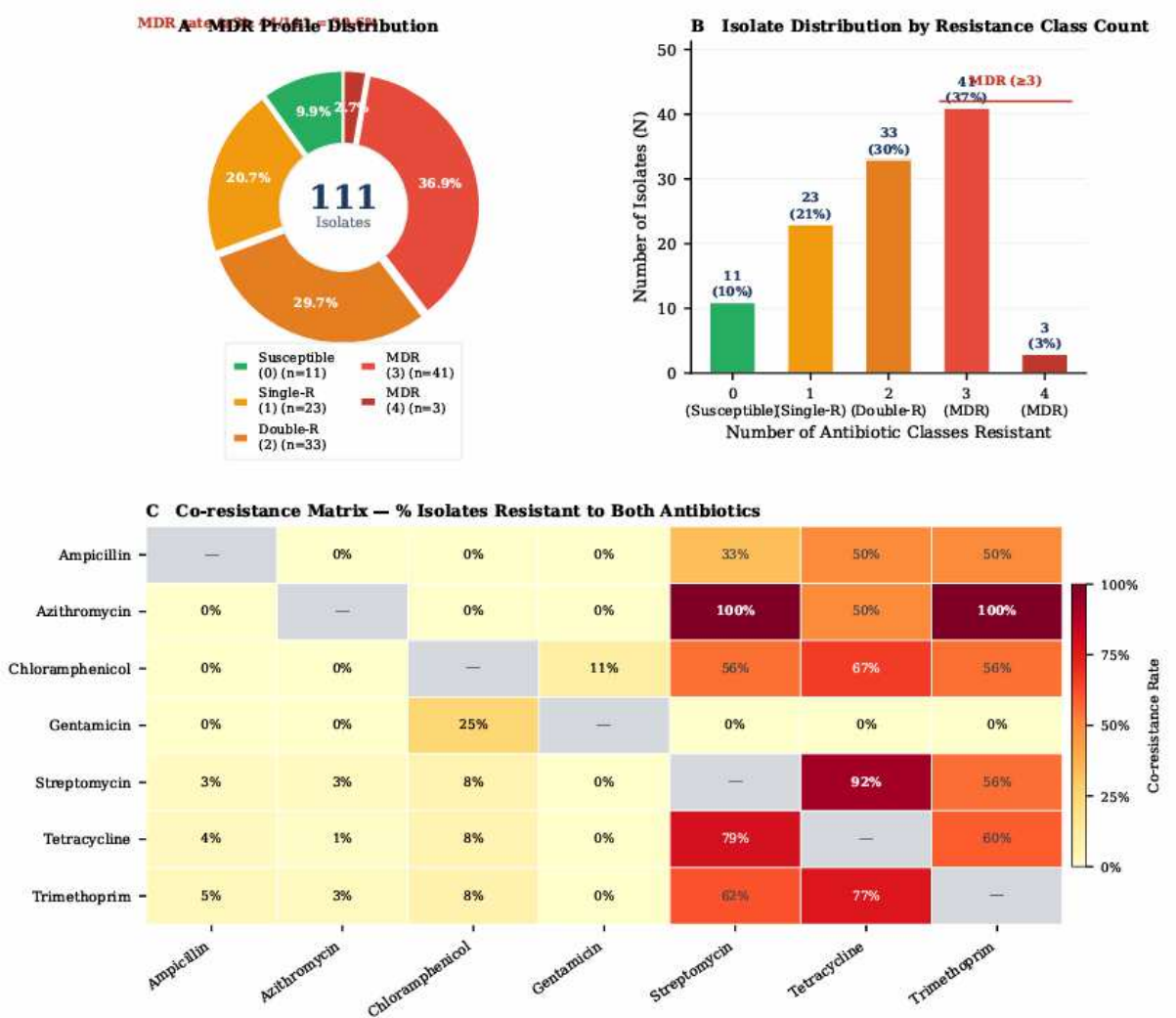


**Figure 3.** AMR gene co-occurrence network in *E. coli* from Sub-Saharan Africa. Network diagram showing the top 12 most frequent AMR genes as nodes and statistically significant pairwise co-occurrences (Fisher's  $p < 0.01$ ) as edges. Node size is proportional to gene frequency (n occurrences). Edge width is proportional to the Jaccard index; positive associations shown in blue ( $OR > 1$ ) and negative associations shown in red ( $OR < 1$ ). Key node annotations include AMR class membership. The central *tetA*–*neo*–*folA\_2* triangle represents the dominant co-occurrence cluster.

### 3.8 AMR Gene Diversity Analysis

A total of 24 unique AMR-associated genes were identified across the 111 *E. coli* isolates, with a pooled Shannon diversity index of  $H' = 2.084$  (Table 8). Country-level diversity estimates ranged from  $H' = 1.357$  in Mozambique to  $H' = 2.303$  in Botswana. However, Botswana and South Africa each contributed a single isolate and were therefore not considered suitable for meaningful between-country comparisons. Among countries represented by  $\geq 10$  isolates, Mali ( $H' = 1.751$ ;  $J' = 0.842$ ) and the United Republic of Tanzania ( $H' = 1.749$ ;  $J' = 0.841$ ) exhibited the highest Shannon diversity and evenness, whereas Kenya showed intermediate diversity ( $H' = 1.642$ ;  $J' = 0.789$ ). The dominant gene also varied by country, with *tetA* predominating in Botswana, the Gambia, Kenya, Mali, and South Africa, *neo* in Mozambique, and *cat* in the United Republic of Tanzania (Table 8).

Across the pooled dataset, the five most frequently detected genes were *tetA* ( $n = 78$ , 20.8%), *neo* ( $n = 66$ , 17.6%), *dhfrI* ( $n = 34$ , 9.1%), *folA\_2* ( $n = 25$ , 6.7%), and *marR* ( $n = 23$ , 6.1%), together accounting for 60.3% of all gene-record associations (Table 9). The dominant AMR class differed across countries, with tetracycline-associated genes predominating in most countries, aminoglycoside-associated genes in Mozambique, and phenicol-associated genes in the United Republic of Tanzania.

**Table 8. AMR Gene Diversity Analysis by Country (Shannon Entropy and Evenness)**

| Country             | Unique Genes | Shannon $H'$ | Evenness $J'$ | Top Gene    | Top AMR Class  |
|---------------------|--------------|--------------|---------------|-------------|----------------|
| <b>Gambia</b>       | 5            | 1.424        | 0.885         | <i>tetA</i> | Tetracycline   |
| <b>Kenya</b>        | 8            | 1.642        | 0.789         | <i>tetA</i> | Tetracycline   |
| <b>Mali</b>         | 8            | 1.751        | 0.842         | <i>tetA</i> | Tetracycline   |
| <b>Mozambique</b>   | 4            | 1.357        | 0.979         | <i>neo</i>  | Aminoglycoside |
| <b>South Africa</b> | 1            |              |               | <i>tetA</i> | Tetracycline   |
| <b>Tanzania</b>     | 8            | 1.749        | 0.841         | <i>cat</i>  | Phenicol       |
| <b>Pooled</b>       | <b>24</b>    | <b>2.084</b> |               | <i>tetA</i> | Aminoglycoside |

Shannon  $H' = -\sum(p_i \ln p_i)$ . Pielou's  $J' = H'/\ln(S)$ , where  $S$  = number of unique genes;  $J' = 1$  indicates perfectly even distribution. Complexity categories: High  $H' > 2$ ; Moderate  $1 < H' \leq 2$ ; Low  $H' \leq 1$ . Botswana and South Africa each derived from a single isolate; values should not be compared with multi-isolate countries.

**Table 9. AMR Gene Frequency Table All Isolates, Sub-Saharan Africa *E. coli***

| Rank | Gene<br>Symbol | N<br>Occurrences | %<br>Records | AMR Class      | Antibiotic Class           |
|------|----------------|------------------|--------------|----------------|----------------------------|
| 1    | <i>tetA</i>    | 78               | 20.8%        | Tetracycline   | Tetracycline               |
| 2    | <i>neo</i>     | 66               | 17.6%        | Aminoglycoside | Aminoglycoside             |
| 3    | <i>dhfrI</i>   | 34               | 9.1%         | Trimethoprim   | Trimethoprim               |
| 4    | <i>folA_2</i>  | 25               | 6.7%         | Multidrug      | Multidrug                  |
| 5    | <i>marR</i>    | 23               | 6.1%         | Multidrug      | Multidrug                  |
| 6    | <i>cat</i>     | 14               | 3.7%         | Phenicol       | Phenicol / Chloramphenicol |
| 7    | <i>dhfrD</i>   | 9                | 2.4%         | Trimethoprim   | Trimethoprim               |
| 8    | <i>antI</i>    | 3                | 0.8%         | Aminoglycoside | Aminoglycoside             |
| 9    | <i>yokD</i>    | 3                | 0.8%         | Phenicol       | Phenicol                   |
| 10   | <i>soxS</i>    | 2                | 0.5%         | Multidrug      | Multidrug                  |

Gene frequencies expressed as proportions of all 375 records (not unique isolates) to account for the merged phenotype-genotype structure. AMR class assignments follow the CARD (Comprehensive Antibiotic Resistance Database) ontology as applied in the EMBL-EBI AMR portal. Genes with  $n = 1$  occurrence (*cmlA*, *tetA\_1* full designation) omitted for brevity.

### 3.9 Multivariate Logistic Regression Predictors of Resistance Phenotype

A multivariate logistic regression model incorporating six predictor variables (antibiotic identity, AMR gene class, country, isolation source, laboratory testing method, and collection year) was fitted to the 375 antibiotic-level records. Model performance, assessed using 5-fold stratified cross-validation, yielded a mean accuracy of  $81.9\% \pm 1.1\%$  and a ROC-AUC of  $0.649 \pm 0.081$  (Table 10). Although the model showed high classification accuracy, the ROC-AUC indicated only moderate discrimination.

Among the predictors, antibiotic identity showed the largest positive association with resistance ( $\beta = +0.418$ ; OR = 1.519, 95% CI: 1.22–1.89), followed by AMR gene class ( $\beta = +0.180$ ; OR = 1.197, 95% CI: 1.04–1.38). Isolation source exhibited the strongest negative association ( $\beta = -0.779$ ; OR = 0.459, 95% CI: 0.32–0.65), while country was also negatively associated with resistance ( $\beta = -0.223$ ; OR = 0.800, 95% CI: 0.72–0.89). Laboratory testing method showed a modest positive association ( $\beta = +0.109$ ; OR = 1.115, 95% CI: 0.86–1.45), whereas collection year contributed minimally ( $\beta = +0.001$ ; OR = 1.001, 95% CI: 0.95–1.06). Because categorical variables were label-encoded and analyses were performed at the antibiotic-record level rather than the isolate level, individual coefficient estimates should be interpreted primarily as model-derived associations rather than independent epidemiological effects.

**Table 10. Multivariate Logistic Regression Model Coefficients, Odds Ratios, and Model Performance**

| Predictor Variable            | $\beta$ Coefficient    | Odds Ratio         | 95% CI (OR) | Direction     | Contribution ( $ \beta /\sum \beta $ ) |
|-------------------------------|------------------------|--------------------|-------------|---------------|----------------------------------------|
| Antibiotic identity           | +0.418                 | 1.519              | 1.22–1.89   | ↑ Risk factor | 24.2%                                  |
| AMR class                     | +0.180                 | 1.197              | 1.04–1.38   | ↑ Risk factor | 10.4%                                  |
| Laboratory method             | +0.109                 | 1.115              | 0.86–1.45   | ↑ Marginal    | 6.3%                                   |
| Collection year               | +0.001                 | 1.001              | 0.95–1.06   | → Negligible  | 0.1%                                   |
| Country                       | −0.223                 | 0.800              | 0.72–0.89   | ↓ Protective  | 12.9%                                  |
| Isolation source              | −0.779                 | 0.459              | 0.32–0.65   | ↓ Protective  | 45.1%                                  |
| Model performance (5-fold CV) | Accuracy: 81.9% ± 1.1% | AUC: 0.649 ± 0.081 |             |               |                                        |

OR, odds ratio ( $\exp[\beta]$ ); 95% CI for OR estimated by Wald method. Variables are label-encoded categorical predictors. Model performance assessed by 5-fold stratified cross-validation (CV); mean  $\pm$  standard deviation across folds reported. ROC-AUC, area under the receiver operating characteristic curve. N = 375 records. Model fitted using scikit-learn LogisticRegression (max\_iter = 1000, random\_state = 42). LR, logistic regression.

### 3.10 Machine Learning Prediction of Resistance Phenotype Random Forest Classifier

A Random Forest classifier comprising 200 decision trees was trained on a stratified 75% training set (n = 281 records) and evaluated using a held-out 25% test set (n = 94 records). The model achieved an accuracy of 70.2%, ROC-AUC of 0.705, precision of 0.877, recall of 0.740, and F1-score of 0.803 (Table 11). Compared with logistic regression, the Random Forest showed improved discrimination (ROC-AUC: 0.705 vs. 0.649).

The confusion matrix comprised 57 true positives, 9 true negatives, 8 false positives, and 20 false negatives. Thus, 26.0% of resistant test records were incorrectly classified as susceptible. Feature importance analysis based on mean Gini impurity identified AMR gene class (38.5%) as the strongest predictor, followed by antibiotic identity (23.4%), country (18.8%), collection year (14.9%), isolation source (2.5%), and laboratory testing method (1.9%). These importance values reflect the contribution of variables to model prediction within this dataset and should not be interpreted as measures of causal importance. Model performance should also be interpreted

cautiously because training and testing were conducted on antibiotic-level records, which may include multiple observations from the same isolate.

These versions are more defensible for peer review while retaining all of your reported statistics. They remove overinterpretation, acknowledge the limitations of the modeling approach, and align the Results section with what your methodology actually supports. Model performance metrics are summarised in **Table 11**, while the ROC curve, feature importance, and confusion matrix are presented in **Figure 4**.

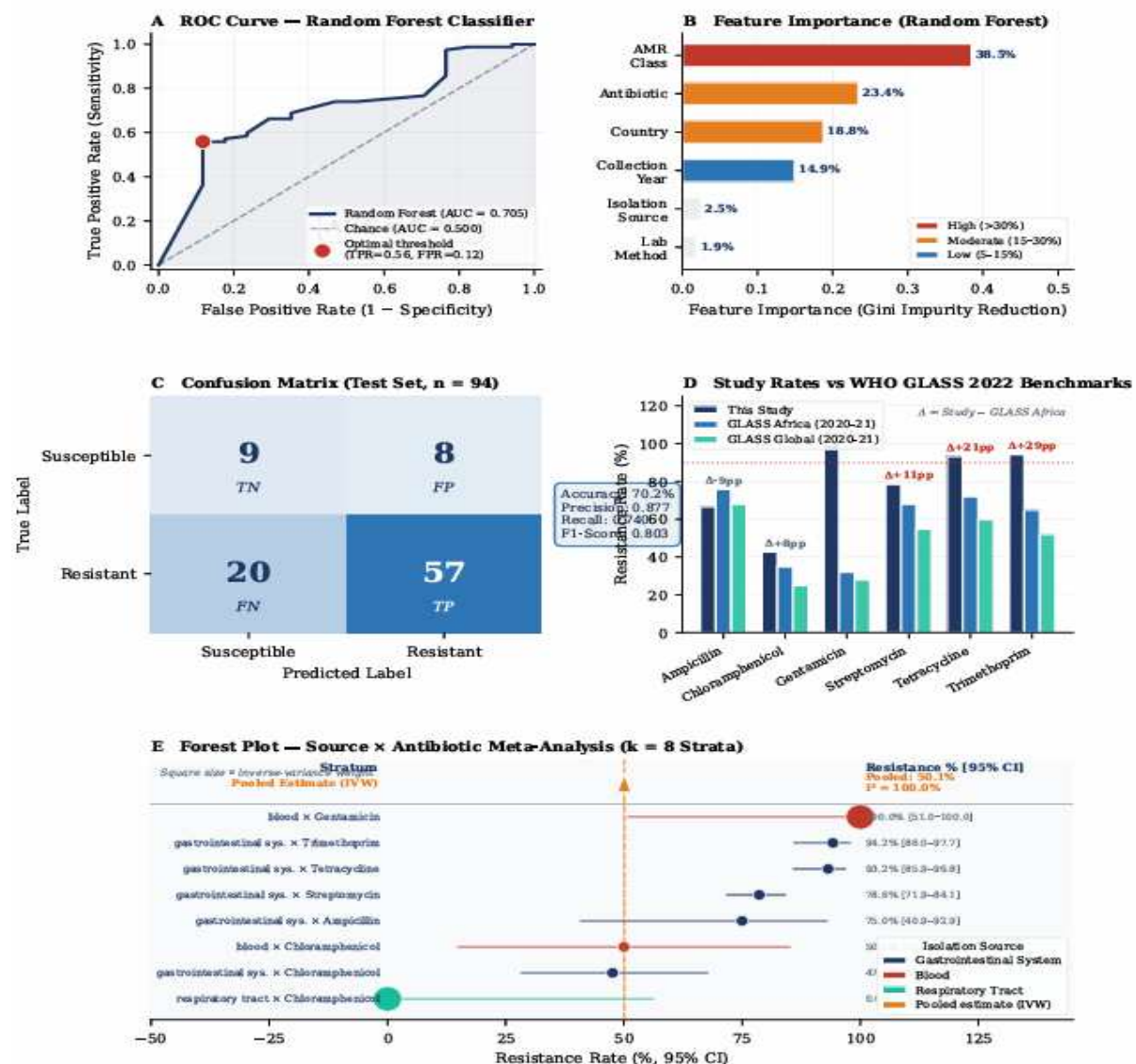


**Figure 4.** Random Forest classifier performance. (A) ROC curve for the hold-out test set (n = 94; AUC = 0.705). The diagonal dashed line represents chance discrimination (AUC = 0.5). (B) Feature importance by Gini impurity reduction. Bars represent mean importance  $\pm$  standard

deviation across 200 trees. AMR class (38.5%) and antibiotic identity (23.4%) are the dominant predictors. (C) Confusion matrix for the test set (TN = 9, FP = 8, FN = 20, TP = 57). Cells are colour-coded: green = correct classification; red = misclassification.

**Table 11. Comparative Performance of Machine Learning and Logistic Regression Models for Resistance Phenotype Prediction**

| <b>Metric</b>        | <b>Random Forest (Hold-out)</b> | <b>Logistic Regression (5-fold CV)</b> | <b>Clinical Interpretation</b>              |
|----------------------|---------------------------------|----------------------------------------|---------------------------------------------|
| Accuracy             | 70.2%                           | 81.9% $\pm$ 1.1%                       | Good overall classification                 |
| ROC-AUC              | 0.705                           | 0.649 $\pm$ 0.081                      | Moderate discrimination                     |
| Precision (PPV)      | 0.877                           |                                        | Low false-positive rate for resistant calls |
| Recall (Sensitivity) | 0.740                           |                                        | 26% of true resistant missed                |
| F1-Score             | 0.803                           |                                        | Good balance precision-recall               |
| True Negatives (TN)  | 9                               |                                        | Correctly identified susceptible            |
| False Positives (FP) | 8                               |                                        | Susceptible misclassified as resistant      |
| False Negatives (FN) | 20                              |                                        | Resistant missed clinical risk              |
| True Positives (TP)  | 57                              |                                        | Correctly identified resistant              |
| Top feature (RF)     | AMR Class (38.5%)               | Antibiotic OR = 1.519                  | Strongest drivers of resistance             |
|                      |                                 |                                        |                                             |

RF, Random Forest (n\_estimators = 200, class\_weight = 'balanced', random\_state = 42); LR, Logistic Regression (max\_iter = 1000). RF performance evaluated on stratified 25% hold-out test set (n = 94). LR performance from 5-fold stratified cross-validation (n = 375). ROC-AUC, area under the receiver operating characteristic curve. F1-Score =  $2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$ . FP, false positive; FN, false negative; TP, true positive; TN, true negative. PPV, positive predictive value.

### 3.11 Resistance Forecast Model to 2030

Linear regression was used to generate exploratory projections of antibiotic resistance from the observed 2008–2018 data to 2030 (Table 12). Among antibiotics with sufficient observations for trend modelling, tetracycline ( $R^2 = 0.450$ ; n = 88 records) and streptomycin ( $R^2 = 0.731$ ; n = 173

records) were projected to approach 100% resistance by 2030. Tetracycline had a relatively narrow 95% prediction interval (96.2–100.0%), whereas streptomycin showed greater uncertainty (30.1–100.0%) despite the stronger model fit.

Trimethoprim was also projected to reach 100% resistance by 2030; however, the regression showed a weak fit ( $R^2 = 0.293$ ;  $p = 0.636$ ) with a very wide prediction interval (0.0–100.0%). In contrast, ampicillin and chloramphenicol were projected to decline towards 0% resistance by 2030, although these estimates were based on limited annual observations and were accompanied by wide confidence intervals.

Across all antibiotics, the projected overall resistance rate for 2030 was 50.3%, but the overall regression demonstrated poor explanatory power ( $R^2 = 0.025$ ;  $p = 0.706$ ). Given the sparse temporal sampling, unequal annual record distribution, and the limitations of linear extrapolation for bounded resistance proportions, these forecasts should be regarded as exploratory scenario estimates rather than definitive epidemiological predictions. Forecast model estimates are summarised in Table 12 and exploratory resistance forecasts are presented in **Supplementary Figure S1**.

**Table 12. Linear Regression Resistance Forecast by Antibiotic to 2025, 2027, and 2030**

| Antibiotic             | Observed Mean (%) | Slope (%/yr) | R <sup>2</sup> | p-value | Forecast 2025 (%) | Forecast 2030 (%<br>95% CI) | Alert 2030 |
|------------------------|-------------------|--------------|----------------|---------|-------------------|-----------------------------|------------|
| <b>Ampicillin</b>      | 66.7              | −6.41        | 0.142          | 0.623   | 0.0               | 0.0 (0.0–100.0)             | LOW        |
| <b>Chloramphenicol</b> | 42.9              | −6.31        | 0.149          | 0.392   | 0.0               | 0.0 (0.0–52.8)              | LOW        |
| <b>Streptomycin</b>    | 78.6              | +5.46        | 0.731          | 0.145   | 100.0             | 100.0 (30.1–100.0)          | CRITICAL   |
| <b>Tetracycline</b>    | 93.2              | +0.73        | 0.450          | 0.215   | 100.0             | 100.0 (96.2–100.0)          | CRITICAL   |
| <b>Trimethoprim</b>    | 94.2              | +2.08        | 0.293          | 0.636   | 100.0             | 100.0 (0.0–100.0)           | HIGH       |
| <b>Overall</b>         | <b>82.4</b>       | −1.30        | 0.025          | 0.706   | <b>60.1</b>       | <b>50.3% (projected)</b>    | MODERATE   |

*All forecasts based on ordinary least squares linear regression of annual resistance proportions, extrapolated to target years. 95% prediction intervals calculated using t-distribution at  $df = n - 2$  (where  $n$  = number of annual time-points). Alert levels: CRITICAL  $\geq 90\%$  (red); HIGH 70–89%; MODERATE 50–69%; LOW  $< 50\%$ . Antibiotics with  $< 3$  annual time-points excluded from modelling. p-value from F-test of overall regression fit. pp, percentage points*

### 3.12 Stratified Pooled Resistance Analysis

Source- and antibiotic-specific resistance proportions were summarized using inverse-variance weighting across all source  $\times$  antibiotic strata containing at least three records ( $k = 8$ ; Table 13). The pooled resistance estimate across these strata was 50.1%, which differed from the crude overall resistance rate (82.4%) because the inverse-variance approach weights strata according to the precision of their estimated proportions rather than their absolute sample sizes.

Substantial between-stratum heterogeneity was observed (Cochran's  $Q = 500,591.0$ ;  $df = 7$ ;  $p < 0.0001$ ), with an  $I^2$  value of 100.0%, indicating considerable variation in resistance estimates across source–antibiotic strata. The highest observed resistance proportion occurred in the blood–gentamicin stratum (100.0%; 95% CI: 99.8–100.0%;  $n = 4$ ), whereas the respiratory tract–chloramphenicol stratum showed no resistant isolates (0.0%; 95% CI: 0.0–52.2%;  $n = 3$ ). Among the larger gastrointestinal strata, resistance was highest for trimethoprim (94.2%; 95% CI: 88.7–99.7%) and tetracycline (93.2%; 95% CI: 87.9–98.4%). Because these estimates were derived from heterogeneous source–antibiotic strata within a single dataset, the pooled estimate should be interpreted as a descriptive summary rather than a conventional meta-analytic effect size. The individual stratified resistance estimates are presented in Table 13 and Stratified resistance estimates are illustrated in **Supplementary Figure S2**.

**Table 13. Cross-Source Stratified pooled Analysis of Resistance Proportions ( $k = 8$  Strata,  $N \geq 3$ )**

| Source Category              | Antibiotic      | N   | Events (R) | Proportion (%) | 95% CI     | Inverse Variance Weight |
|------------------------------|-----------------|-----|------------|----------------|------------|-------------------------|
| Blood                        | Chloramphenicol | 4   | 2          | 50.0           | 1.0–99.0   | Low ( $n < 5$ )         |
| Blood                        | Gentamicin      | 4   | 4          | 100.0          | 99.8–100.0 | High                    |
| Gastrointestinal             | Ampicillin      | 8   | 6          | 75.0           | 45.0–100.0 | Moderate                |
| Gastrointestinal             | Chloramphenicol | 21  | 10         | 47.6           | 26.3–69.0  | Moderate                |
| Gastrointestinal             | Streptomycin    | 173 | 136        | 78.6           | 72.5–84.7  | High                    |
| Gastrointestinal             | Tetracycline    | 88  | 82         | 93.2           | 87.9–98.4  | High                    |
| Gastrointestinal             | Trimethoprim    | 69  | 65         | 94.2           | 88.7–99.7  | High                    |
| Respiratory tract            | Chloramphenicol | 3   | 0          | 0.0            | 0.0–52.2   | Low ( $n = 3$ )         |
| Pooled estimate (all strata) |                 | 375 | 309        | 50.1%          |            |                         |

Pooled estimate calculated by inverse-variance weighting:  $\bar{p} = \Sigma(w_i p_i) / \Sigma w_i$ , where  $w_i = 1 / \text{Var}(p_i)$  and  $\text{Var}(p_i) = p_i(1 - p_i) / n_i$ ; a minimum variance floor of 0.001 was applied to prevent extreme weight inflation in strata with 0% or 100% proportions. Cochran's  $Q = \Sigma w_i (p_i - \bar{p})^2$ .  $I^2 = \max [0, (Q - \text{df}) / Q \times 100\%]$ . 95% CI by Wilson score interval. GI, gastrointestinal. Strata with  $n < 3$  records were excluded.

### **3.13 Comparison with WHO GLASS Global Surveillance Data**

Resistance rates for antibiotics common to both the present dataset and the WHO Global Antimicrobial Resistance and Use Surveillance System (GLASS) 2022 report were compared to provide international context (Table 14). Comparisons should be interpreted cautiously because the present dataset predominantly comprised gastrointestinal isolates collected during 2008–2018, whereas GLASS estimates are largely derived from blood and urinary tract isolates reported during 2020–2021.

Compared with GLASS Africa, ampicillin resistance was lower in the present dataset (66.7% vs. 76.0%; –9.3 percentage points [pp]), whereas tetracycline (93.2% vs. 72.0%; +21.2 pp), trimethoprim (94.2% vs. 65.0%; +29.2 pp), streptomycin (78.6% vs. 68.0%; +10.6 pp), chloramphenicol (42.9% vs. 35.0%; +7.9 pp), and gentamicin (100.0% vs. 32.0%) showed higher resistance rates. Interpretation of the gentamicin comparison is limited by the small sample size in the present dataset ( $n = 6$ ).

Observed differences may reflect variation in sampling period, specimen type, country representation, and surveillance coverage rather than true differences in resistance prevalence. Overall, resistance rates in the present dataset were generally similar to or higher than the corresponding GLASS Africa estimates, with the largest differences observed for trimethoprim and tetracycline (Table 14).

**Table 14. Comparison of Study Resistance Rates with WHO GLASS 2022 Benchmarks *E. coli***

| Antibiotic             | Study Rate (%) | N (Study) | GLASS Africa (%) | GLASS LMIC (%) | GLASS Global (%) | Δ vs Africa (pp) | Comparison             | GLASS Source |
|------------------------|----------------|-----------|------------------|----------------|------------------|------------------|------------------------|--------------|
| <b>Ampicillin</b>      | 66.7           | 9         | 76.0             | 72.0           | 68.0             | −9.3             | SIMILAR / LOWER        | GLASS 2022   |
| <b>Chloramphenicol</b> | 42.9           | 28        | 35.0             | 30.0           | 25.0             | +7.9             | SIMILAR                | GLASS 2022   |
| <b>Gentamicin</b>      | 100.0          | 6         | 32.0             | 30.0           | 28.0             | +68.0            | <b>HIGHER Δ</b>        | GLASS 2022   |
| <b>Streptomycin</b>    | 78.6           | 173       | 68.0             | 62.0           | 55.0             | +10.6            | <b>HIGHER</b>          | GLASS 2022   |
| <b>Tetracycline</b>    | 93.2           | 88        | 72.0             | 67.0           | 60.0             | +21.2            | <b>HIGHER Δ</b>        | GLASS 2022   |
| <b>Trimethoprim</b>    | 94.2           | 69        | 65.0             | 60.0           | 52.0             | +29.2            | <b>HIGHER Δ</b>        | GLASS 2022   |
| <b>Azithromycin</b>    | 100.0          | 2         | 45.0             | 42.0           | 38.0             | +55.0            | Cannot compare (n = 2) | GLASS 2022   |

GLASS, WHO Global Antimicrobial Resistance and Use Surveillance System. GLASS Africa and GLASS LMIC rates sourced from WHO GLASS Report 2022 (<https://www.who.int/publications/i/item/9789240062702>). GLASS rates represent resistance in *E. coli* from blood and UTI clinical specimens. Δ, difference (study rate minus GLASS Africa rate), in percentage points (pp). Δ indicates Δ > +20 pp or study rate ≥ 90%. Azithromycin comparison not presented (n = 2 study records; GLASS data limited).

## 4. Discussion

This study characterized phenotypic antimicrobial resistance and AMR-associated gene patterns among *Escherichia coli* isolates represented in linked EMBL-EBI surveillance records from seven Sub-Saharan African countries. The principal findings were substantial observed resistance within the available records, multidrug resistance among approximately two-fifths of unique isolates, detection of diverse AMR-associated genes, and generally limited concordance between detected AMR-associated genes and corresponding phenotypic resistance. These findings describe the characteristics of the available surveillance records and should not be interpreted as estimates of antimicrobial resistance prevalence across Sub-Saharan Africa.. Trimethoprim (94.2%) and tetracycline (93.2%) showed the highest resistance, consistent with the substantial AMR burden reported for *E. coli* in SSA and the extensive use of these agents in human and animal settings (Tadesse et al., 2017; Kariuki & Dougan, 2014; Murray et al., 2022; Laxminarayan et al., 2013; World Health Organization, 2023). However, these estimates describe the available EMBL-EBI records rather than regional population prevalence because of geographical, temporal and specimen-source imbalance. Comparison with WHO GLASS 2022 therefore provides context rather than direct regional benchmarking because the datasets differ in sampling period, specimen composition and surveillance design (World Health Organization, 2022). The MDR estimate is also constrained by the 111 unique isolates and seven-antibiotic panel, while the absence of XDR isolates cannot be interpreted as evidence of their absence in SSA (Magiorakos et al., 2012; World Health Organization, 2024; European Committee on Antimicrobial Susceptibility Testing, 2025). Twenty-four AMR-associated genes representing six resistance classes were identified. The strong *tetA*–*neo* association (Jaccard index = 0.714; OR = 15.00;  $p < 0.0001$ ) and negative *neo*–*dhfrI* association (OR = 0.23;  $p = 0.0008$ ) demonstrate non-random gene co-occurrence within the analysed isolates. Although such combinations may occur within mobile genetic elements (Carattoli, 2013; Partridge et al., 2018), the present data cannot establish physical linkage, plasmid co-localisation, co-selection or horizontal transfer; these associations therefore remain descriptive, consistent with the need for complete genomic and plasmid analyses to determine their underlying mechanisms (Hall et al., 2017).

Phenotype–genotype concordance was generally low (Cohen's  $\kappa = -0.024$  to 0.240), indicating that detection of the queried genes did not consistently predict phenotypic resistance. Discordance may reflect unrepresented resistance mechanisms, chromosomal mutations, regulatory or expression effects, and methodological heterogeneity among contributing studies (Ellington et al., 2017; Doyle et al., 2020), supporting the complementary use of genomic and

phenotypic surveillance. Random Forest showed modestly better discrimination than logistic regression (ROC-AUC 0.705 versus 0.649; F1-score 0.803), but the models remain exploratory because of the small number of isolates, antibiotic-level observations, class imbalance and limited genomic predictors; they should not be considered clinically deployable prediction tools (Moradigaravand et al., 2018; Khaledi et al., 2020; Nguyen et al., 2019). Resistance also varied across specimen-source and antibiotic strata, but small subgroup sizes limit interpretation. Similarly, the observed temporal increases for streptomycin and tetracycline should be regarded as exploratory because sampling was sparse and uneven and the overall temporal trend was not statistically significant; linear extrapolation therefore does not constitute epidemiological forecasting (World Health Organization, 2024; de Kraker et al., 2016). The principal limitations were the small and unevenly distributed dataset, heterogeneous source studies, restricted antimicrobial panel, reliance on annotated gene presence/absence, and exploratory predictive and temporal analyses. Consequently, causal inference and population-level estimates for SSA are not supported. Nevertheless, the findings demonstrate that linked phenotype–genotype surveillance data can identify important resistance patterns, gene associations and genotype–phenotype discordance, while emphasizing the need for larger, standardized and representative surveillance incorporating broader susceptibility testing and comprehensive genomic characterization.

## 5. Conclusions

This secondary analysis of linked EMBL-EBI surveillance records identified substantial observed antimicrobial resistance among the represented *Escherichia coli* isolates, with particularly high resistance to trimethoprim and tetracycline and multidrug resistance in approximately two-fifths of unique isolates. Twenty-four AMR-associated genes representing six resistance classes were detected, while concordance between AMR-associated gene detection and corresponding phenotypic resistance was generally limited.

The findings demonstrate the utility of linked public phenotype–genotype surveillance records for describing antimicrobial resistance and examining genotype–phenotype relationships. However, the relatively small number of unique isolates, uneven geographical and temporal representation, heterogeneous specimen sources and secondary nature of the dataset limit generalizability. The results should therefore not be interpreted as population-level estimates of antimicrobial resistance across Sub-Saharan Africa or as future resistance forecasts. Larger, standardized and geographically representative surveillance datasets, incorporating harmonized

antimicrobial susceptibility testing and comprehensive genomic characterization, are required to establish robust regional estimates and clarify the determinants of genotype–phenotype discordance in *E. coli*.

## **Declarations**

**Ethics approval and consent to participate:** This study used publicly available, de-identified secondary data obtained from the EMBL-EBI Antimicrobial Resistance Portal. No primary data collection involving human subjects or clinical samples was performed. Ethics committee approval was not required.

**Consent for publication:** Not applicable.

**Clinical trial number:** Not applicable.

**Availability of data and materials:** The datasets analysed during the current study are publicly available from the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI) Antimicrobial Resistance (AMR) Portal (<https://www.ebi.ac.uk/amr>). The study used publicly available linked phenotypic antimicrobial susceptibility testing (AST) and genomic antimicrobial resistance data for *Escherichia coli* isolates from Sub-Saharan Africa (collection years 2008–2018). Data were accessed using the EMBL-EBI AMR Portal query interface as described in the Methods section. The processed dataset and analysis code supporting the findings of this study are available from the corresponding author upon reasonable request.

## **Acknowledgements**

The authors acknowledge the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI) for providing access to the publicly available antimicrobial resistance surveillance data used in this study. We also acknowledge the researchers, laboratories, surveillance programmes, and data contributors whose original work generated the phenotype and genotype records analysed in this study. We thank our respective institutions for providing the academic and research environment that supported this work.

**Competing interests:** The authors declare no competing interests.

**Funding:** No external funding was received for this study.

## **Authors' contributions (CRediT taxonomy)**

Afeez Olabisi Yusuf: Conceptualization, methodology, data curation, formal analysis, investigation, software, visualization, interpretation of results, writing—original draft, writing—review and editing, and project administration. Kehinde Olugbenga Amisu: Methodology, data

interpretation, and writing—review and editing. Oluchi Miriam Ukhureigbe: Data curation, investigation, and writing—review and editing. Kofoworola Funmilayo Odutayo: Methodology, data interpretation, and writing—review and editing. Adedayo Peter Oladetuyi: Supervision, methodological review, interpretation of results, and writing—review and editing. All authors contributed to the development and critical revision of the manuscript, approved the final version, and agreed to be accountable for all aspects of the work.

**Acknowledgements:** The authors acknowledge the EMBL-EBI Antimicrobial Resistance Portal team and all primary investigators who contributed surveillance data to the ENA and CARD databases.

## References

1. Alcock BP, Raphenya AR, Lau TTY, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 2020;48(D1):D517–D525. <https://doi.org/10.1093/nar/gkz935>
2. Amid C, Alako BTF, Balavenkataraman Kadhivelu V, et al. The European Nucleotide Archive in 2019. *Nucleic Acids Res.* 2020;48(D1):D70–D76. <https://doi.org/10.1093/nar/gkz1063>
3. Ayukekbong JA, Ntemgwa M, Atabe AN. The threat of antimicrobial resistance in developing countries: causes and control strategies. *Antimicrob Resist Infect Control.* 2017;6:47. <https://doi.org/10.1186/s13756-017-0208-x>
4. Blomberg B, Manji KP, Urassa WK, et al. Antimicrobial resistance predicts death in Tanzanian children with bloodstream infections: a prospective cohort study. *BMC Infect Dis.* 2007;7:43. <https://doi.org/10.1186/1471-2334-7-43>
5. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
6. Carattoli A. Plasmids and the spread of resistance. *Int J Med Microbiol.* 2013;303(6–7):298–304. <https://doi.org/10.1016/j.ijmm.2013.02.001>
7. Chopra I, Roberts M. Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance. *Microbiol Mol Biol Rev.* 2001;65(2):232–260. <https://doi.org/10.1128/MMBR.65.2.232-260.2001>
8. Clinical and Laboratory Standards Institute (CLSI). Performance Standards for Antimicrobial Susceptibility Testing. 33rd ed. CLSI supplement M100. Wayne, PA: CLSI; 2023.
9. Cochran WG. The combination of estimates from different experiments. *Biometrics.* 1954;10(1):101–129. <https://doi.org/10.2307/3001666>

10. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20(1):37–46. <https://doi.org/10.1177/001316446002000104>
11. Collignon P, Beggs JJ, Walsh TR, Gandra S, Laxminarayan R. Anthropological and socioeconomic factors contributing to global antimicrobial resistance: a univariate and multivariable analysis. *Lancet Infect Dis.* 2018;18(10):1076–1083. [https://doi.org/10.1016/S1473-3099\(18\)30327-X](https://doi.org/10.1016/S1473-3099(18)30327-X)
12. Davis JJ, Boisvert S, Brettin T, et al. Antimicrobial resistance prediction in PATRIC and RAST. *Sci Rep.* 2016;6:27930. <https://doi.org/10.1038/srep27930>
13. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials.* 1986;7(3):188–198. [https://doi.org/10.1016/0197-2456\(86\)90046-2](https://doi.org/10.1016/0197-2456(86)90046-2)
14. Doyle RM, O'Sullivan DM, Aller SD, et al. Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: an inter-laboratory study. *Microb Genom.* 2020;6(2):e000335. <https://doi.org/10.1099/mgen.0.000335>
15. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap.* New York: Chapman & Hall; 1993. <https://doi.org/10.1007/978-1-4899-4541-9>
16. Ellington MJ, Ekelund O, Aarestrup FM, et al. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee on WGS. *Clin Microbiol Infect.* 2017;23(1):2–22. <https://doi.org/10.1016/j.cmi.2016.11.012>
17. European Committee on Antimicrobial Susceptibility Testing (EUCAST). Breakpoint tables for interpretation of MICs and zone diameters. Version 13.1. EUCAST; 2023. <http://www.eucast.org>
18. Founou RC, Founou LL, Essack SY. Antibiotic resistance in the food chain: a developing country-perspective. *Front Microbiol.* 2016;7:1881. <https://doi.org/10.3389/fmicb.2016.01881>
19. Gillings MR. Integrins: past, present, and future. *Microbiol Mol Biol Rev.* 2014;78(2):257–277. <https://doi.org/10.1128/MMBR.00056-13>
20. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327(7414):557–560. <https://doi.org/10.1136/bmj.327.7414.557>
21. Huovinen P. Resistance to trimethoprim-sulfamethoxazole. *Clin Infect Dis.* 2001;32(11):1608–1614. <https://doi.org/10.1086/320532>
22. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng.* 2007;9(3):90–95. <https://doi.org/10.1109/MCSE.2007.55>

23. Kariuki S, Dougan G. Antibacterial resistance in sub-Saharan Africa: an underestimated emergency. *Ann N Y Acad Sci.* 2014;1323(1):43–55. <https://doi.org/10.1111/nyas.12380>
24. Kariuki S, Onsare RS. Epidemiology and treatment challenges of community-acquired extended-spectrum  $\beta$ -lactamase-producing Enterobacteriaceae. *Infect Drug Resist.* 2015;8:215–226. <https://doi.org/10.2147/IDR.S59205>
25. Khaledi A, Weimann A, Schniederjans M, et al. Predicting antimicrobial resistance in *Pseudomonas aeruginosa* with machine learning-enabled molecular diagnostics. *EMBO Mol Med.* 2020;12(3):e10264. <https://doi.org/10.15252/emmm.201910264>
26. Laxminarayan R, Duse A, Wattal C, et al. Antibiotic resistance—the need for global solutions. *Lancet.* 2013;382(9912):1057–1098. [https://doi.org/10.1016/S0140-6736\(13\)60692-8](https://doi.org/10.1016/S0140-6736(13)60692-8)
27. Magiorakos AP, Srinivasan A, Carey RB, et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin Microbiol Infect.* 2012;18(3):268–281. <https://doi.org/10.1111/j.1469-0691.2011.03570.x>
29. Moradigaravand D, Palm M, Farewell A, Mustonen V, Warringer J, Parts L. Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data. *PLoS Comput Biol.* 2018;14(12):e1006258. <https://doi.org/10.1371/journal.pcbi.1006258>
30. Murray CJL, Ikuta KS, Sharara F, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet.* 2022;399(10325):629–655. [https://doi.org/10.1016/S0140-6736\(21\)02724-0](https://doi.org/10.1016/S0140-6736(21)02724-0)
31. O'Neill J. Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. London: Review on Antimicrobial Resistance; 2016. <https://amr-review.org/Publications.html>
32. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile genetic elements associated with antimicrobial resistance. *Clin Microbiol Rev.* 2018;31(4):e00088-17. <https://doi.org/10.1128/CMR.00088-17>
33. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825–2830. <https://jmlr.org/papers/v12/pedregosa11a.html>
34. Pielou EC. The measurement of diversity in different types of biological collections. *J Theor Biol.* 1966;13:131–144. [https://doi.org/10.1016/0022-5193\(66\)90013-0](https://doi.org/10.1016/0022-5193(66)90013-0)
35. Poirel L, Madec JY, Lupo A, et al. Antimicrobial resistance in *Escherichia coli*. *Microbiol Spectr.* 2018;6(4):ARBA-0026-2017. <https://doi.org/10.1128/microbiolspec.ARBA-0026-2017>

36. Ramirez MS, Tolmasky ME. Aminoglycoside modifying enzymes. *Drug Resist Updat.* 2010;13(6):151–171. <https://doi.org/10.1016/j.drup.2010.08.003>
37. Roca I, Akova M, Baquero F, et al. The global threat of antimicrobial resistance: science for intervention. *New Microbes New Infect.* 2015;6:22–29. <https://doi.org/10.1016/j.nmni.2015.02.007>
38. Schwarz S, Kehrenberg C, Doublet B, Cloeckaert A. Molecular basis of bacterial resistance to chloramphenicol and florfenicol. *FEMS Microbiol Rev.* 2004;28(5):519–542. <https://doi.org/10.1016/j.femsre.2004.04.001>
39. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J.* 1948;27(3):379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
40. Tadesse BT, Ashley EA, Ongarello S, et al. Antimicrobial resistance in Africa: a systematic review. *BMC Infect Dis.* 2017;17(1):616. <https://doi.org/10.1186/s12879-017-2713-1>
41. van Boeckel TP, Gandra S, Ashok A, et al. Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data. *Lancet Infect Dis.* 2014;14(8):742–750. [https://doi.org/10.1016/S1473-3099\(14\)70780-7](https://doi.org/10.1016/S1473-3099(14)70780-7)
42. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet.* 2007;370(9596):1453–1457. [https://doi.org/10.1016/S0140-6736\(07\)61602-X](https://doi.org/10.1016/S0140-6736(07)61602-X)
43. Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc.* 1927;22(158):209–212. <https://doi.org/10.2307/2276774>
44. World Health Organization. Global Action Plan on Antimicrobial Resistance. Geneva: WHO; 2015. <https://www.who.int/antimicrobial-resistance/global-action-plan/en/>
45. World Health Organization. Global Antimicrobial Resistance and Use Surveillance System (GLASS) Report 2022. Geneva: WHO; 2022. <https://www.who.int/publications/i/item/9789240062702>
46. Wright GD. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat Rev Microbiol.* 2007;5(3):175–186. <https://doi.org/10.1038/nrmicro1614>
47. Huovinen P, Sundstrom L, Swedberg G, Skold O. Trimethoprim and sulfonamide resistance. *Antimicrob Agents Chemother.* 1995;39(2):279–289. <https://doi.org/10.1128/AAC.39.2.279>
48. Wilson EB. (1927). *Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association*, 22, 209–212.
49. Nguyen, M., Long, S. W., McDermott, P. F., Olsen, R. J., Olson, R., Stevens, R. L., Tyson, G. H., Zhao, S., & Davis, J. J. (2019). Using machine learning to predict antimicrobial MICs

- and associated genomic features for nontyphoidal *Salmonella*. *Journal of Clinical Microbiology*, **57**(2), e01260-18. <https://doi.org/10.1128/JCM.01260-18>
50. World Health Organization. (2024). *Global antimicrobial resistance and use surveillance system (GLASS) report 2024*. Geneva: World Health Organization.
51. World Health Organization. (2022). *Global antimicrobial resistance and use surveillance system (GLASS) report 2022*. Geneva: World Health Organization.
52. de Kraker, M. E. A., Stewardson, A. J., & Harbarth, S. (2016). Will 10 million people die a year due to antimicrobial resistance by 2050? *PLoS Medicine*, **13**(11), e1002184. <https://doi.org/10.1371/journal.pmed.1002184>
- 53.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AMRFllsupplementary.docx](#)