London's Global University

Machine learning to predict antibiotic susceptibility of bloodstream infections compared to clinician prescribing

By

NCTG9

A Dissertation submitted in part fulfilment of

the requirements for the degree of

Master of Science in Health Data Science

Institute of Health Informatics
Faculty of Population Health Sciences
University College London

ABSTRACT

Background: Antimicrobial resistance (AMR) poses a major clinical challenge, particularly in Gram-negative bloodstream infections where delays in effective therapy increase mortality. Predictive tools that can guide empiric prescribing are urgently needed.

Methods: UCLH's structured electronic health records were converted into encounter-level datasets that represented the post-identification (post-species) and empiric (pre-species) phases. Piperacillin–tazobactam, ceftriaxone, gentamicin, flucloxacillin, and amoxicillin resistance were predicted using logistic regression, random forest, and XGBoost models. ROC-AUC, PR-AUC, calibration, and SHAP interpretability were used to evaluate performance. By comparing model-guided prescribing with observed practice and benchmarking against the Oxford AMR prediction model, bug–drug mismatch analysis was used to assess clinical utility.

Results: Logistic regression consistently outperformed ensemble methods, achieving modest discrimination overall (ROC-AUC 0.55–0.73). Species information improved prediction for ceftriaxone and flucloxacillin but not for piperacillin–tazobactam, reflecting high local resistance prevalence (53%). Mismatch analysis showed the largest benefit for ceftriaxone (10.2% absolute reduction), smaller gains for piperacillin–tazobactam (4.0%), and paradoxical worsening for gentamicin (−1.9%). Improvements in empiric coverage were comparable to Oxford findings (4–10%).

Conclusion: Predicting antimicrobial resistance (AMR) is challenging for clinician and models alike; however, even basic machine learning models can enhance empirical prescribing relative to existing practices. More patients may be receiving active treatment as a result of machine learning models. Logistic regression offered the most reliable and interpretable performance, supporting its potential integration into decision support. Future work should focus on external validation, integration of molecular data, and deployment within clinical workflows to strengthen antimicrobial stewardship.

Table of Contents

ABSTRACT **1**

1. INTRODUCTION **3**

1.3 Aims and Hypothesis 5

1.4 Objectives 5

2. METHODS **6**

2.1 Dataset 6

2.2 Data Transformation and Feature Engineering 8

2.3 Pre- and Post-Species Datasets 10

2.4 Model Development 10

2.5 Training, Evaluation and Visualisation 11

2.6 Model Interpretability 11

2.7 Bug-drug mismatch analysis 12

3. RESULTS **13**

3.1 Data Preprocessing and Cohort Assembly 13

3.2 Model Performance Across Antibiotics 14

3.3 Model Interpretability 18

3.4 Drug Mismatch Analysis 21

4. DISCUSSION **23**

4.1 Strengths 25

4.2 Limitations and Future work 25

5. CONCLUSION **26**

6. ETHICS **27**

7. ACKNOWLEDGEMENTS **27**

REFERENCES **28**

APPENDIES **33**

1. INTRODUCTION

The World Health Organisation (WHO) has declared antimicrobial resistance (AMR) a global health emergency, making medications for treating infection less effective, which could result in 10 million deaths per year (O'Neill, 2016). In 2019, gram-negative pathogens caused 3.6 million deaths associated with AMR worldwide (Antimicrobial Resistance Collaborators, 2022).

Structural similarities among antibiotics often facilitate cross-resistance, indicating that minor chemical modifications are no longer sufficient to overcome resistance. Notably, multidrug-resistant organisms (MDROs) can resist more than one class of antimicrobial agents, raising concern within the WHO. including extended-spectrum beta-lactamase (ESBL)-producing E. coli, Klebsiella pneumoniae, and carbapenem-resistant strains such as CRKP (Gall, Long and Hall, 2020). These organisms often exhibit resistance to multiple antibiotic classes, limiting available treatment options. Inappropriate or delayed empirical therapy has been associated with significantly higher mortality. In patients with haematological malignancies and K. pneumoniae bloodstream infections, mortality can reach 77.2% alongside increased risk of septic shock. (Nilsson, 2019). Furthermore, overuse of broad-spectrum antibiotics accelerates resistance trends, complicating future treatment decisions (Melzer and Petersen, 2007). These stark differences in outcome highlight the urgent need for tools that can guide timely, effective empiric antibiotic therapy, particularly in resistant Gram-negative infections.

Although alternative therapeutic approaches such as RNA-based antimicrobials and phage therapy are under investigation, they remain in early development and are not yet widely applicable in clinical settings (Biplab Singha, Singh and Soni, 2024) (Nilsson, 2019b). The clinical application in phage therapy is still limited by the narrow host ranges of phage and variability in patient response. Likewise, there is still lack of standardization and large-scale evidence (Nilsson, 2019b). Despite these alternatives, the timely administration of effective empirical antibiotics remains a cornerstone of treatment for bloodstream infections. Delayed or inappropriate therapy has been linked to poor outcomes. Studies show a stark contrast in 30-day mortality between appropriately treated patients (8.8%) and those receiving inappropriate therapy (77.2%) (Ma et al., 2024). The findings reinforce the critical need for prompt and appropriate empirical therapy.

Increased resistance rates are closely associated with antibiotic use, whereas decreases in antibiotic use have been demonstrated to restore bacterial populations' susceptibility, reflecting the fitness costs of resistance. Limiting unnecessary prescriptions is therefore a crucial for lessening the burden of antibiotic resistance, especially during empiric therapy (Bell et al., 2014), (Melnyk, Wong and Kassen, 2014). Empiric antibiotic therapy is the initiation of antibiotic treatment prior to the confirmation of a patient's specific etiologic, source of infection, or the antibiotic resistance

profile of the infecting pathogen (Zumla, 2010). Errors might occur during this stage, for example, ineffective antibiotic prescriptions (i.e., antibiotics that do not eradicate the bacterial pathogen because it is resistant to them) and overly broad antibiotic prescriptions (i.e., antibiotics with lower coverage would be sufficient to treat the infection. This may result in a serious consequence to patients. Ineffective antibiotic treatment will put patients at risk by allowing resistant bacteria to continue infecting them (Paul et al., 2010; Oshima et al., 2016; Ibrahim et al., 2000). Furthermore, frequent use of broad-spectrum antibiotics is likely to raise the prevalence of antibiotic resistance, making antibiotics less efficient. This could possibly case inappropriate empiric therapy (Merli et al., 2015; Carrara et al., 2018).

Recent studies highlight the potential of artificial intelligence (AI) to enhance early diagnosis and antibiotic decision-making. For example, an AI model trained on clinical data outperformed physicians by over 30% in early sepsis diagnosis while reducing false positives (Goh et al., 2021). AI-driven decision support systems can integrate patient-specific variables, local microbiology data, and prescribing guidelines to recommend more antibiotic choices and reduce broad-spectrum overuse driven by diagnostic uncertainty in severe infections (Kollef et al., 2021). Moreover, agents like nitrofurantoin have limited bloodstream infection data, and some commonly prescribed antibiotics (e.g., amoxicillin) are excluded from ICU guidelines due to >80% resistance rates (Vasudeva, Prem Singh Nirwan and Shrivastava, 2016). These findings underscore the need for locally adapted predictive models tailored to hospital-specific data.

Various studies have determined risk factors for antibiotic-resistant infections by looking at patient demographics, comorbidities, past medical history, and other patient characteristics (Wolfe, Cohen and Larson, 2014). Several machine learning (ML) models have also been applied to predict AMR patterns using EHRs, demographics, and microbiology data. Techniques such as decision trees, logistic regression, random forests, and XGBoost have demonstrated AUCs ranging from 0.70 to 0.93 across diverse patient groups and clinical settings (Feretzakis et al., 2021; McGuire et al., 2021; Sousa et al., 2019). These studies illustrate the feasibility of ML for AMR prediction and support its integration into empirical therapy decision tools. Therefore, ML can provide greater accuracy and clinical relevance, provided they are regularly updated with new data.

1.3 Aims and Hypothesis

This project aims to evaluate and enhance a machine learning model for predicting antimicrobial resistance (AMR) in Gram-negative bloodstream infections using structured electronic health record (EHR) data from UCLH. The study compares pre-species (empiric stage) and post-species (post-identification stage) models, evaluates their clinical utility for guiding empiric prescribing, and benchmarks performance against the Oxford AMR prediction model (Eyre et al., 2025). Interpretability tools, including SHAP analyses (feature importance, directionality, and dependence), were applied to enhance transparency and demonstrate that the models captured clinically plausible predictors. Ultimately, the project seeks to reduce inappropriate broad-spectrum antibiotic use, improve patient outcomes, and contribute to AMR mitigation efforts.

Hypothesis: Incorporating species-level microbiological information and interpretable machine learning methods will improve the predictive accuracy and clinical utility of AMR models for Gram-negative bloodstream infections compared with empiric (pre-species) models and existing benchmarks (e.g., the Oxford AMR model).

1.4 Objectives

- To assemble and preprocess structured EHR and microbiology data from UCLH into pre-species and post-species datasets, ensuring clinically appropriate feature engineering.
- To train, evaluate, and interpret machine learning models (logistic regression, random forest, XGBoost) for predicting antimicrobial resistance, using discrimination, calibration, and SHAP-based interpretability.
- To assess clinical utility through bug–drug mismatch analysis, and compare model-guided prescribing with empiric practice, and benchmarking against the Oxford AMR prediction model.
- To discuss limitations, including missing outpatient exposures, temporal drift, and antibiotic-specific variability, and their implications for local model deployment.

2. METHODS

2.1 Dataset

The data was extracted from the electronic health records at UCLH Hospital. The **Camino_Bronze_Source-DID.db** dataset consists of 16 linked tables in duckDBformat, , covering patient demographics, encounters, prescriptions, laboratory and microbiology results, antimicrobial susceptibility testing (AST), and clinical diagnoses.

- **Demographics & encounters**: Patients_DID, Encounters (age, sex, admission/discharge details, ICU/ED flags).
- **Prescribing**: Prescriptions, PrescriptionIndications, AntibioticMedications (drug formulations, subclasses, exposure timing).
- **Microbiology**: EncounterOrders, EncounterSusceptibility, PatientHistoricalOrders, PatientHistoricalOrderResults, PatientHistoricalSusceptibility (index culture results, resistance history, culture timelines).
- **Clinical context**: PatientProblemLists, PatientCancerStaging, PatientAllergies, PatientHistoricalEncounters (comorbidities, cancer, allergies, prior admissions).
- **Unstructured notes**: EncounterNotes (admission and discharge summaries; not used in this version of the model).

The outcome variable was derived from EncounterSusceptibility, where AST results were harmonised into binary labels (resistant vs susceptible).

Five antibiotics were selected as prediction targets piperacillin–tazobactam (PTZ), ceftriaxone, gentamicin, flucloxacillin, and amoxicillin, reflecting empiric prescribing practices at UCLH and their clinical relevance for Gram-negative sepsis management (Table 2 summarises included formulations).

Table name	Description
AntibioticMedications	Master list of antibiotic medications, generic names, formulations, subclasses, and identifiers.
EncounterNotes	Free-text clinical notes associated with patient encounters (e.g., admission notes, discharge summaries).
EncounterOrders	Orders placed during an encounter, including laboratory tests, microbiology samples, and imaging.
EncounterSusceptibility	Antimicrobial susceptibility test (AST) results linked to specific encounters; reports S/I/R classification.
Encounters	Core table of patient hospital encounters, including dates, locations, and encounter-level metadata.
PatientAllergies	Documented allergies to drugs (including antibiotics) and allergy types/severity.
PatientCancerStaging	Cancer diagnosis and staging information for relevant patients.
PatientHistoricalEncounters	Previous hospital admissions or visits before the index culture.
PatientHistoricalOrderResults	Historical laboratory or microbiology order results predating the current infection.
PatientHistoricalOrders	Historical orders (e.g., microbiology/lab tests) linked to prior encounters.
PatientHistoricalPrescriptions	Records of prior prescriptions (antibiotics and non-antibiotics) linked to patients.
PatientHistoricalSusceptibility	Historical AST results linked to prior infections.
PatientProblemLists	Documented chronic and acute medical problems during care.
Patients_DID	Master patient table (de-identified), including demographics such as age, sex, and unique identifiers.
PrescriptionIndications	Clinical indication recorded for each prescription (e.g., sepsis, pneumonia).
Prescriptions	Prescription-level detail of medications issued during encounters (dose, route, frequency, start/stop dates).

Table 1. Overview of the 16 interlinked tables in the Camino_Bronze_Source-DID.db dataset. These tables capture demographics and encounters, prescriptions and indications, microbiology and antimicrobial susceptibility results, and clinical context.

Antibiotic	Variants / Formulations found	Records
Piperacillin–tazobactam	Piperacillin 4g + tazobactam 500mg, powder for infusion	7
	Piperacillin 2g + tazobactam 250mg, powder for infusion	5
	Piperacillin with tazobactam, intravenous (unspecified strength)	3
Ceftriaxone	“cefTRIAXone” (plain entry)	11
	Ceftriaxone 1g, powder for injection	9
	Ceftriaxone 2g, powder for injection	8

Gentamicin	"gentamicin" (plain entry)	22
	Gentamicin 80mg/2ml injection	11
	Gentamicin 20mg/2ml injection	8
Flucloxacillin	Flucloxacillin 500mg capsules	36
	Flucloxacillin 250mg capsules	35
	Flucloxacillin 125mg/5ml oral solution	18
Amoxicillin	Amoxicillin 500mg capsules	44
	Amoxicillin 250mg capsules	43
	Amoxicillin 250mg/5ml oral suspension	30

Table 2 summarizes the formulations of the antibiotics identified in the prescribing records. This was group with clinical prescribing practice for modelling.

2.2 Data Transformation and Feature Engineering

The structure pipeline is generated to transform raw electronic record to clinical feature importance while preventing data leakage. Starting from microbiology results that defined the index encounter and resistance outcome, we integrated historical resistance profiles, demographic and encounter information, prior antibiotic prescriptions, and diagnosis-based comorbidities. Each feature group was anchored to the encounter timeline to ensure that only information available at the prediction timepoint (pre-species vs. post-species) was used. This process produced a comprehensive encounter-level dataset in which every row represented a unique antibiotic–encounter pair with an associated binary resistance label. This stage produces processed antibiotic data, which will then ready for the model training.

2.2.1 Prescription Features

Firstly, antibiotic history was grouped at pharmacological subclass level. Multiple generic formulations which were seen within the same subclass were collapsed into a single exposure feature to align with stewardship practice. This was joined with Prescription records, resulting in

generic name , subclass, and medical recorded indication. Datetime were normalised time since last exposure to study time exposure to last 30/90/365 day prior to the index prescription. Prescription time was used as the reference to set look-back windows and align with microbiology results.

2.2.2 Encounter Features

Prescription and AntibioticMedication were linked to get the admission and discharge times for target encounter. Invalid records (nulls, discharge before admission, stays >100 days) and overlapping encounters (<6hr gap) were excluded. Healthcare-utilisation covariates were derived by counting historical encounters in the previous 30 days, 1 year, and 2 years before encounter start. Microbiology timelines (blood culture draw, species ID, AST results) were aligned to the encounter to determine whether antibiotics were administered before or after these events.

2.2.3 Diagnosis Features

Diagnoses/problem-list entries were extracted. For each index encounter, features included counts of prior diagnoses within 30 days, 1 year, and 2 years, as well as time since last relevant diagnose. ICD-10 codes were mapped to comorbidity groups, including a cancer flag. Duplicate patient–code entries were collapsed to the earliest record.

2.2.4 Microbiology and Sensitivity Features

Antimicrobial susceptibility testing (AST) results were harmonised into a three-class system (Susceptible, Intermediate, Resistant). At the encounter level, outcomes were summarised using a predefined hierarchy (R > I > S) to derive a binary resistance label (resistant = 1, susceptible = 0; intermediates excluded) and associated timestamps. Only the most recent result for each encounter–antibiotic pair was retained, and entries with missing values were excluded.

Historical resistance profiles were derived by linking PatientHistoricalSusceptibility to PatientHistoricalOrders. Indicators were created to capture resistance to the same antibiotic within 30 days, 1 year, and 2 years, anchored to the index encounter. These variables provided context for the likelihood of resistance based on previous microbiology findings.

Finally, EncounterOrders were joined to OrderResults to derive encounter-level flags for blood, urine, and other cultures, including order times, abnormal results, and whether testing occurred.

2.3 Pre- and Post-Species Datasets

Two datasets were created to reflect different stages of clinical information. The pre-species dataset (empiric stage) represents the time of blood culture sampling before organism identification. It included demographics (age, sex), encounter-level details (admission/discharge times, ICU/ED flags, prior healthcare utilisation), laboratory results (e.g., neutrophil counts), prior antibiotic exposures (summarised at 30-day, 1-year, and 2-year look-back windows), and diagnosis-based comorbidity indicators (ICD-10 groupings, cancer staging). Organism-related variables were excluded to prevent information leakage.

The post-species dataset (post-identification stage, typically ~24 hours later) contained the same features with the addition of organism-level predictors, reflecting information available after microbiology identification but prior to susceptibility testing.

Both datasets were processed into harmonised feature matrices, with non-predictive identifiers removed and missing values imputed. This structure enabled direct comparison of model performance at the empiric versus post-identification stages.

2.4 Model Development

2.4.1 Data Partitioning

Machine learning classifiers were trained to predict antimicrobial resistance for each antibiotic. Data were partitioned into temporally ordered training, validation, and test sets with proportion of 60:20:20 to preserve class balance. Suitable hyperparameters were chose, and validation data were used to monitor performance and tune thresholds. while the held-out test set provided an unbiased estimate of generalisation. This is to ensure that each model was trained on quality-controlled data.

2.4.2 Algorithms

Logistic Regression was used as an interpretable baseline, with features z-score standardised and balanced class weights applied, Random Forest captured non-linear interactions using depth restriction and class weighting, serving as a robust ensemble benchmark with interpretable feature importance. Additionally, XGBoost was optimised for class imbalance), trained with early stopping, and tuned for precision–recall (PR-AUC).

Deep learning was not pursued due to modest sample size, structured features, and the importance of interpretability. Collectively, these models provided complementary coverage of linear, bagging, and boosting approaches, ensuring clinically meaningful comparisons.

2.5 Training, Evaluation and Visualisation

Thresholds for binary classification were selected on the validation set, with the primary objective of achieving ≥90% sensitivity to prioritise early detection of resistance.

Model were evaluated on held-out test set using both numerical performance metrics and graphical visualization. For each antibiotic-timepoint dataset, predictions were generated as both class labels and probabilities.

Overall discrimination between resistant and susceptible outcomes was quantified using area under the receiver operating characteristic curve. Because of the imbalanced nature of the outcome as resistance less frequent than susceptibility, the area under precision-recall curve (PR-AUC) was the primary performance measure. Additionally, sensitivity (recall), specificity, precision (positive predictive value), and F1 score were also recorded. This provided clinically interpretable measures of the trade-off between false negatives (undertreatment risk) and false positives (unnecessary broad-spectrum coverage).

Receiver operating characteristic (ROC) curves and AUC were plotted for each model, and Precision recall (PR) curve were generated with baseline precision. histograms of predicted probability distributions were created for resistant and susceptible groups to demonstrate class separation. Feature importance plots were generated for tree-based and linear models, ranking predictors according to their relative contributions.

2.6 Model Interpretability

This was assessed using SHAP (Shapley Additive Explanations) to quantify the contribution of individual features to model predictions. These were presented through:

- **Bar plots** (mean absolute SHAP values) to rank overall feature importance,
- **Directionality plots** to illustrate whether features increased or decreased resistance risk
- **Dependence plots** to show how feature effects varied across value ranges and interacted with other predictors.

These analyses identified clinically plausible drivers of resistance, including recent antibiotic exposure, prior resistance history, and ICU admission. To evaluate the added value of microbiological information, pre- vs post-species models were compared, highlighting both the stability of clinical predictors and the emergence of *species identity* as an additional determinant once culture results were incorporated.

2.7 Bug-drug mismatch analysis

To assess the clinical impact, bug–drug mismatch analysis was performed to compare observed empiric prescribing with both susceptibility results and model-guided prescribing strategies. This analysis was implemented for piperacillin–tazobactam, ceftriaxone, and gentamicin, which represented key agents in sepsis management pathways in the UK. This study represent complementary empiric strategies, including monotherapy (ceftriaxone), broad β-lactam (PTZ), and adjunctive aminoglycoside (gentamicin).

Piperacillin–tazobactam (PTZ) combination injection is a first-line agent for severe sepsis, valued for broad coverage of β-lactamase–producing organisms and frequent use in polymicrobial infections (Perry and Markham, 1999). Ceftriaxone with a long half-life, allows once-daily dosing, making it suitable for empiric therapy and outpatient parenteral treatment (Richards et al., 1984). It offers broad Gram-negative activity, including against multidrug-resistant Enterobacteriaceae (Zanichelli et al., 2023). Furthermore, Gentamicin, an aminoglycoside, is commonly given as a single adjunct dose in UK sepsis. The ability against aerobic gram-negative bacteria making gentamicin a good option to treat several common infections (Phillips et al., 1977).

For each encounter, empiric antibiotic prescriptions were identified and linked to the antimicrobial susceptibility results (S/I/R classification). The mismatch rate was defined as the proportion of encounters in which the prescribed antibiotic was inactive against the causative pathogen. This reflects the risk of undertreatment under existing prescribing practice.To simulate model-guided prescribing, predicted resistance probabilities were applied at varying thresholds (10–30%). At each threshold, the model was assumed to recommend avoiding the antibiotic if the predicted probability of resistance exceeded the threshold, and mismatch rates were recalculated.

3. RESULTS

3.1 Data Preprocessing and Cohort Assembly

Antibiotic	Encounters	Patients	Features	Encounters with SIR	Resistant Cases	Resistance Rate
Piperacillin–tazobactam	18,307	12,359	51	730	387	53.00%
Ceftriaxone	8,175	6,941	51	443	144	32.50%
Gentamicin	12,644	10,106	51	3,726	616	16.50%
Flucloxacillin	6,222	5,417	51	1,636	249	15.20%

Table 3 shows the final modelling cohorts for each antibiotic after preprocessing. The number of encounters with valid susceptibility results varied by agent.

UCLH EHR contains 16 relational tables covering prescriptions, encounters, diagnoses, microbiology results, and laboratory data. Key sources included Prescriptions (n = 526,510 records), EncounterOrders (n = 10,016,694), EncounterSusceptibility (n = 281,295), and PatientHistoricalSusceptibility (n = 503,384). Antibiotics were identified through the AntibioticMedications mapping table, which allowed matching of generic formulations to pharmacological subclasses as seen in table 1,2 (e.g., PTZ, ceftriaxone, gentamicin, flucloxacillin, amoxicillin).

Each dataset was stratified into pre-species and post-species. feature set of 43 predictors was retained for both feature sets. The number of encounters contributing valid susceptibility results varied substantially, from 443 (ceftriaxone) to 3,726 (gentamicin). Resistance prevalence was highest for PTZ (53.0%), followed by ceftriaxone (32.5%), with lower rates for gentamicin (16.5%) and flucloxacillin (15.2%). These rates are consistent with institutional surveillance data, underscoring the clinical relevance of the study cohort.

3.2 Model Performance Across Antibiotics

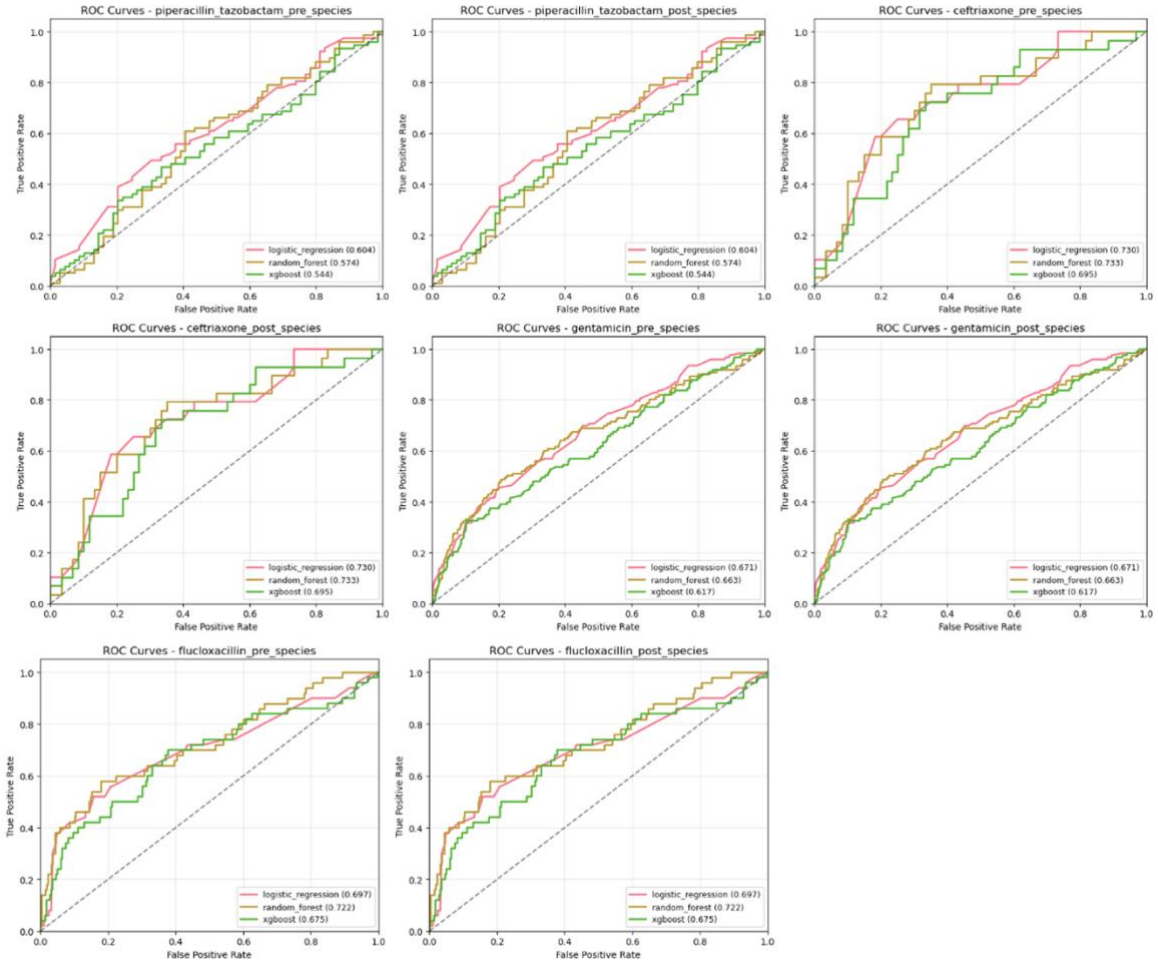


Figure 1. Receiver Operating Characteristic (ROC) curves for predicting resistance to target antibiotics in pre-species (empiric stage) and post-species (identification stage) datasets.

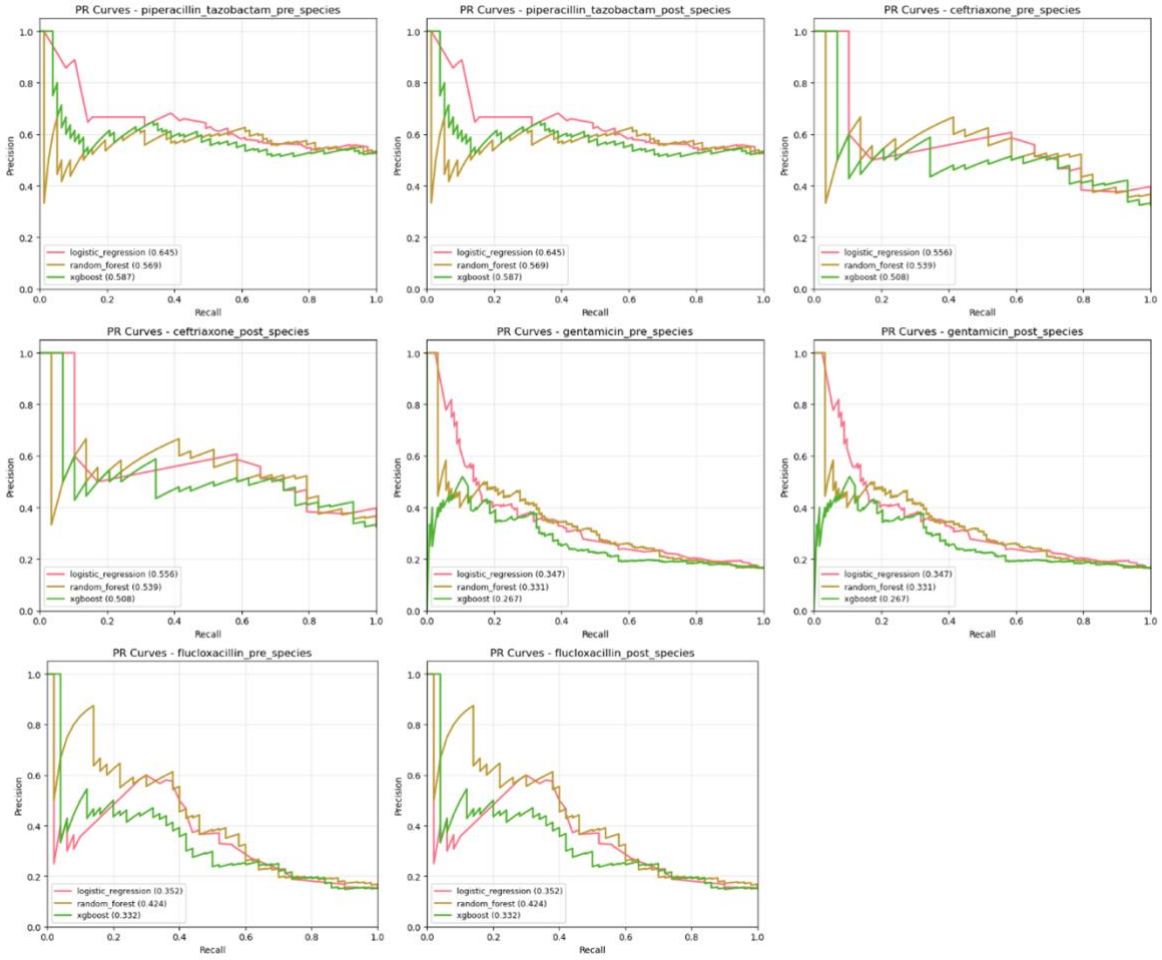


Figure 2. Precision–Recall (PR) curves for predicting resistance to target antibiotics in pre-species (empiric stage) and post-species (identification stage) datasets.

Across all antibiotics and timepoints, logistic regression consistently outperformed the ensemble methods, achieving the highest ROC-AUC and PR-AUC values in most settings (Figures 1,2). For PTZ pre-species, logistic regression achieved a ROC-AUC of 0.604 and PR-AUC of 0.636, compared with 0.574/0.576 for random forest and 0.544/0.592 for XGBoost. Similar patterns were observed for gentamicin and flucloxacillin, where logistic regression maintained superior or comparable discrimination relative to non-linear models. Ceftriaxone yielded the best overall results, with ROC-AUC values up to 0.733 (random forest, pre-species) and 0.730 (logistic regression, post-species), indicating moderate discriminative performance.

Precision–recall curves confirmed these findings: logistic regression maintained higher precision at equivalent levels of recall, particularly for PTZ and gentamicin, while random forest and XGBoost tended to plateau earlier.

Model discrimination was modest overall, with ROC-AUC values typically in the range of 0.55–0.65, indicating limited ability to separate resistant from susceptible outcomes (Figure 3). Across

antibiotics and timepoints, logistic regression consistently outperformed or matched the ensemble models, achieving the highest ROC-AUC and PR-AUC in most settings.

Incorporating species-level features improved discrimination, particularly for ceftriaxone and flucloxacillin, highlighting the value of organism information. By contrast, PTZ remained challenging to predict, likely reflecting the high resistance prevalence in our cohort (53% vs. ~7% in Oxford), which constrained predictive performance.

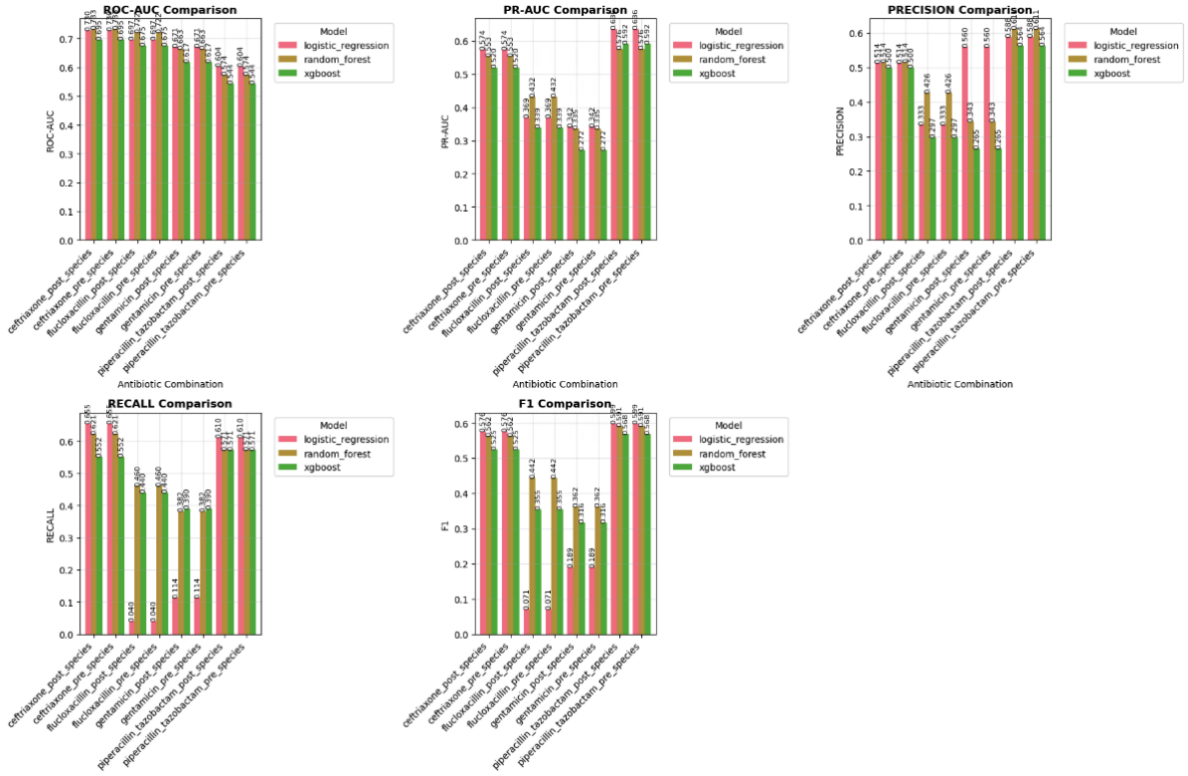
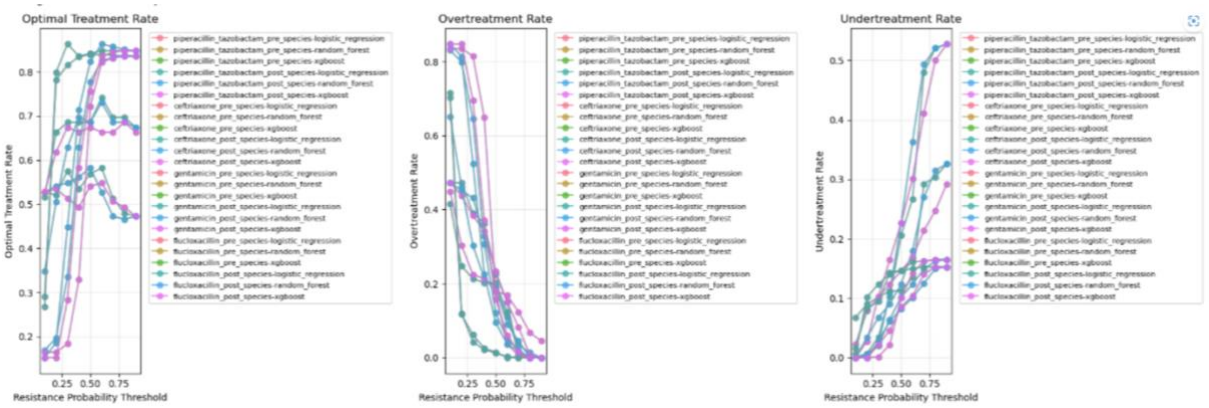


Figure 3. Comparative performance of logistic regression, random forest, and XGBoost models across antibiotics and datasets (pre-species vs post-species).

Additionally, confusion matrices (Appendix A1) showed models were more accurate in predicting susceptible than resistant cases, reflecting the challenge of predicting rare resistance events. Logistic regression yielded a more balanced classification, reducing false negatives compared to ensemble methods, though undertreatment risk remained substantial.



Feature 4 Clinical decision analysis of model-guided prescribing across resistance probability thresholds by measuring optimal treatment (% active antibiotic), overtreatment (% unnecessarily broad antibiotic), and undertreatment (% inactive antibiotic, highest risk)

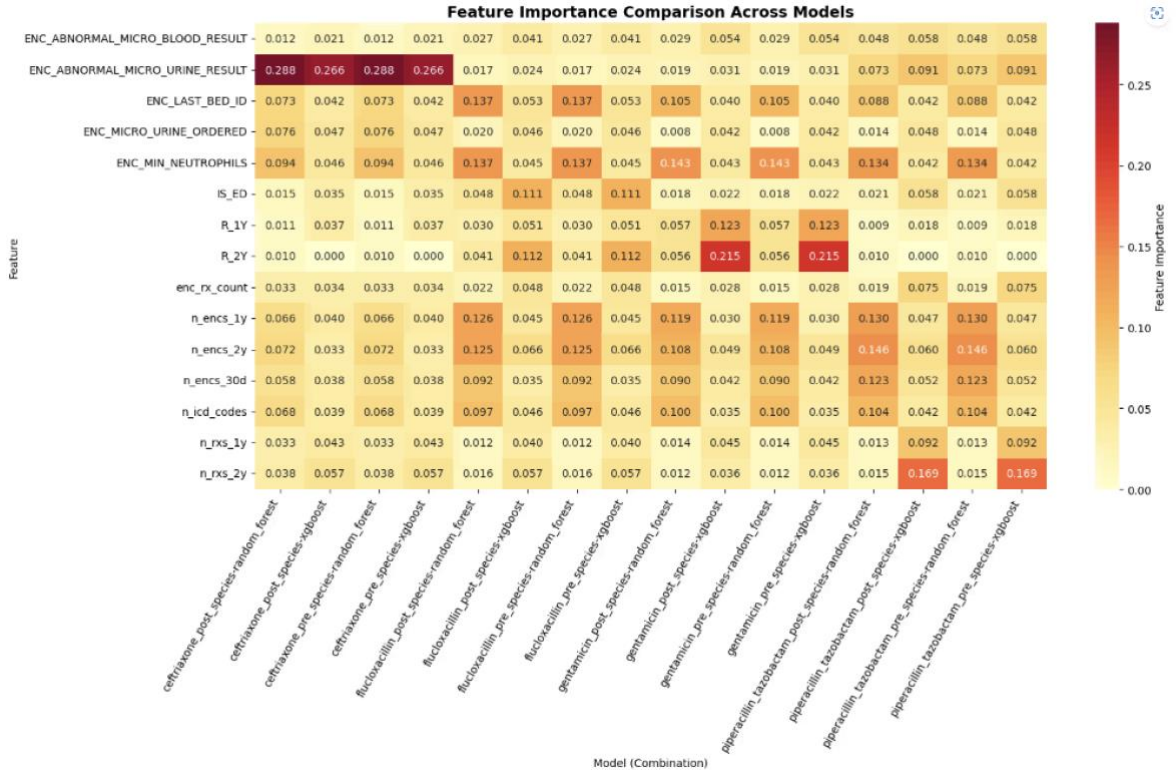


Figure 5. Feature importance comparison across models.

From Clinical decision analysis shown in Figure 4. As resistance probability thresholds increased, overtreatment rates decreased across all antibiotics, while undertreatment rates rose, reflecting the inherent trade-off in model-guided prescribing. Logistic regression achieved the most favourable balance, with higher optimal treatment rates across thresholds compared to random forest and XGBoost. For PTZ, logistic regression reduced overtreatment at thresholds ≥0.25 without a marked rise in undertreatment, while ensemble models either overtreat or undertreat more aggressively at equivalent thresholds. Similar patterns were observed for gentamicin and flucloxacillin, with logistic

regression sustaining higher optimal treatment rates. These findings suggest that logistic regression models could potentially support more efficient empiric prescribing by minimising overtreatment while controlling undertreatment risk.

Additionally, Feature importance analysis (Figure 5) demonstrated that recent abnormal microbiology results (blood and urine cultures), neutrophil count, and prior resistance history were among the strongest predictors across models.

3.3 Model Interpretability

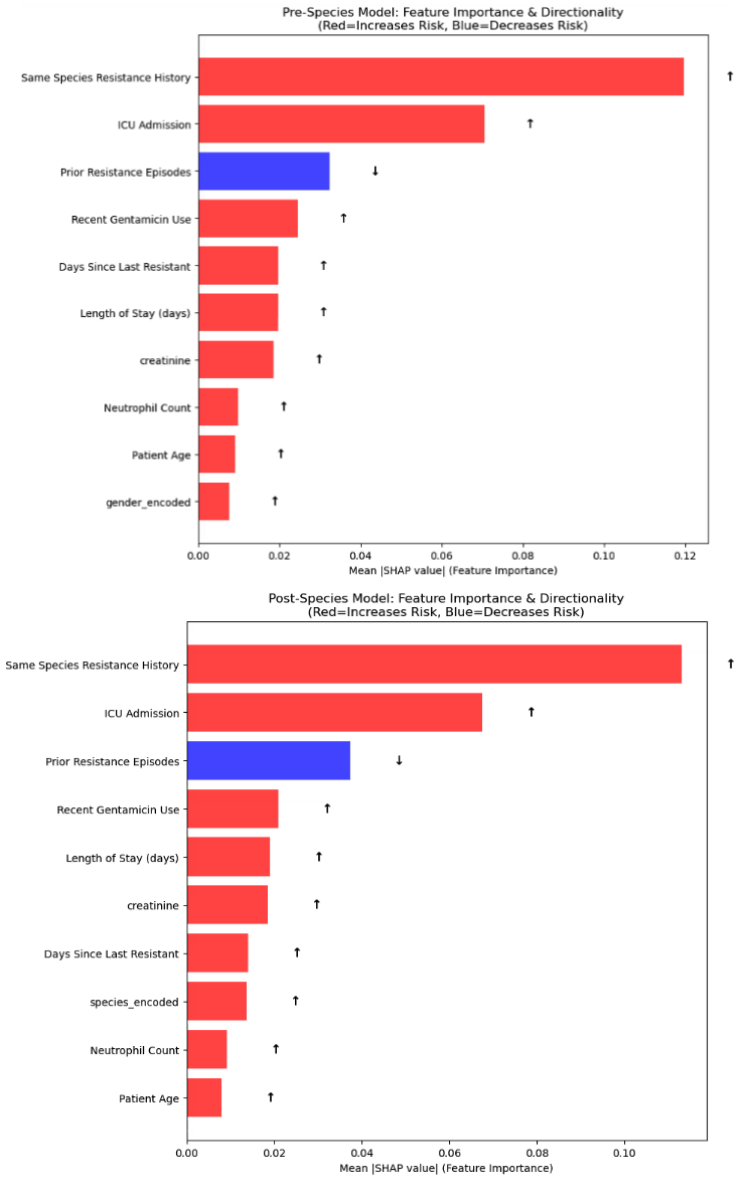


Figure 8. SHAP feature importance and directionality (Pre- and Post-species models). Bar plots showing the top predictors of resistance in pre-species and post-species models. Same-species resistance history and ICU admission were the strongest contributors to resistance risk in both models. Recent antibiotic use, prior resistance episodes, and hospital stay length also featured prominently. Directionality indicated that most predictors increased resistance risk (red), while prior resistance episodes had a small inverse effect (blue).

.

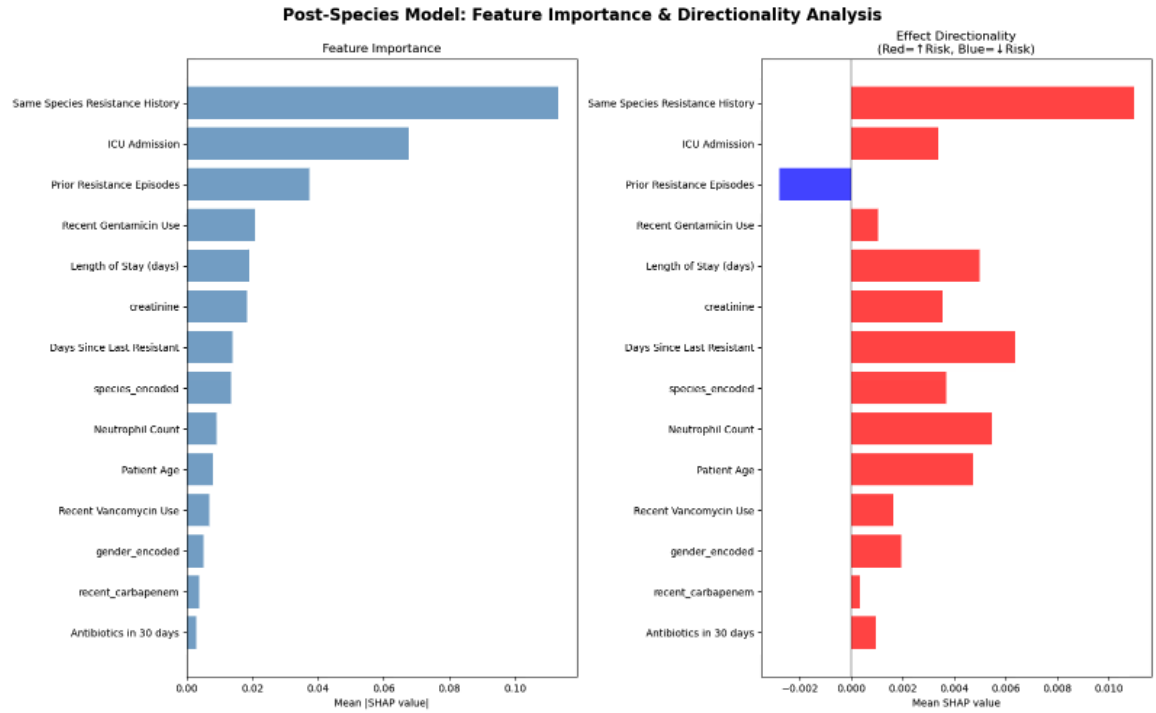


Figure 9. SHAP feature importance and directionality analysis (post-species model). Expanded interpretability analysis of the post-species model. The left panel ranks predictors by overall importance, while the right panel shows their directional effects. Same-species resistance history, ICU admission, and prior resistance episodes dominated predictions, with most features positively associated with resistance. The inclusion of species identity among the top features highlights the added value of microbiological information.

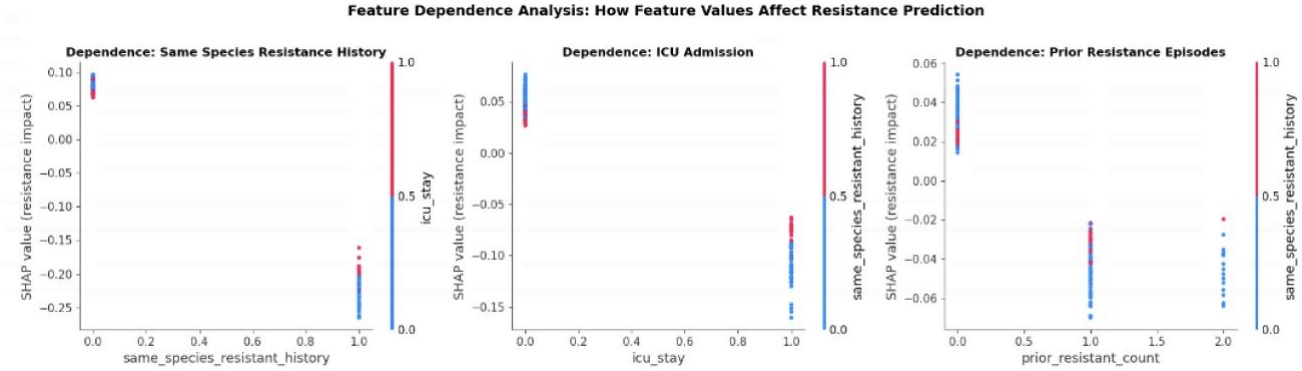


Figure 10. SHAP dependence plots for key predictors. Color indicates interaction effects with other features while steep slopes indicate strong feature effects at certain value ranges. Dependence plots illustrating non-linear relationships for same-species resistance history, ICU admission, and prior resistance episodes. Non-linear relationships show how feature effect change with different values. Prior resistant isolates in the same species increased predicted resistance. ICU admission had a clear threshold effect, with higher predicted risk for ICU patients. The first prior resistance episode exerted the strongest effect, with diminishing incremental impact from subsequent episodes.

According to the figure 8, same-species resistance history consistently emerged as the strongest predictor (SHAP importance 0.12 in the pre-species model, 0.11 in the post-species model, Appendix A3). Other highly ranked predictors included ICU admission (0.07 in both models), prior resistance episodes, recent gentamicin use, and length of hospital stay. Microbiological data shifted importance onto species identity, which ranked among the top 10 predictors in the post-species model but not pre-species. These predictors are biologically plausible: prior resistance reflects selective pressure, ICU admission indicates nosocomial risk, and longer hospital stay increases exposure. Directionality analysis (figure 9) indicated that in the post-species model, the most influential predictors of resistance were prior resistance history to the same species, ICU admission, and previous resistance episodes. Most features were associated with increased resistance risk, with exception of prior resistance episodes, which represented mixed directionality.

Additionally, dependence plots (figure 10) were generated to further explore non-linear relationships for key predictors. There is a steep threshold effect in same-species resistance history: once a resistant isolate had been identified, predicted risk rose sharply. ICU admission demonstrates binary effect with higher predicted resistance for ICU patients compared to non-ICU encounters. Previous resistance episodes showed diminishing returns, with the first resistant episode having the biggest impact and later episodes adding less predictive value.

3.4 Drug Mismatch Analysis

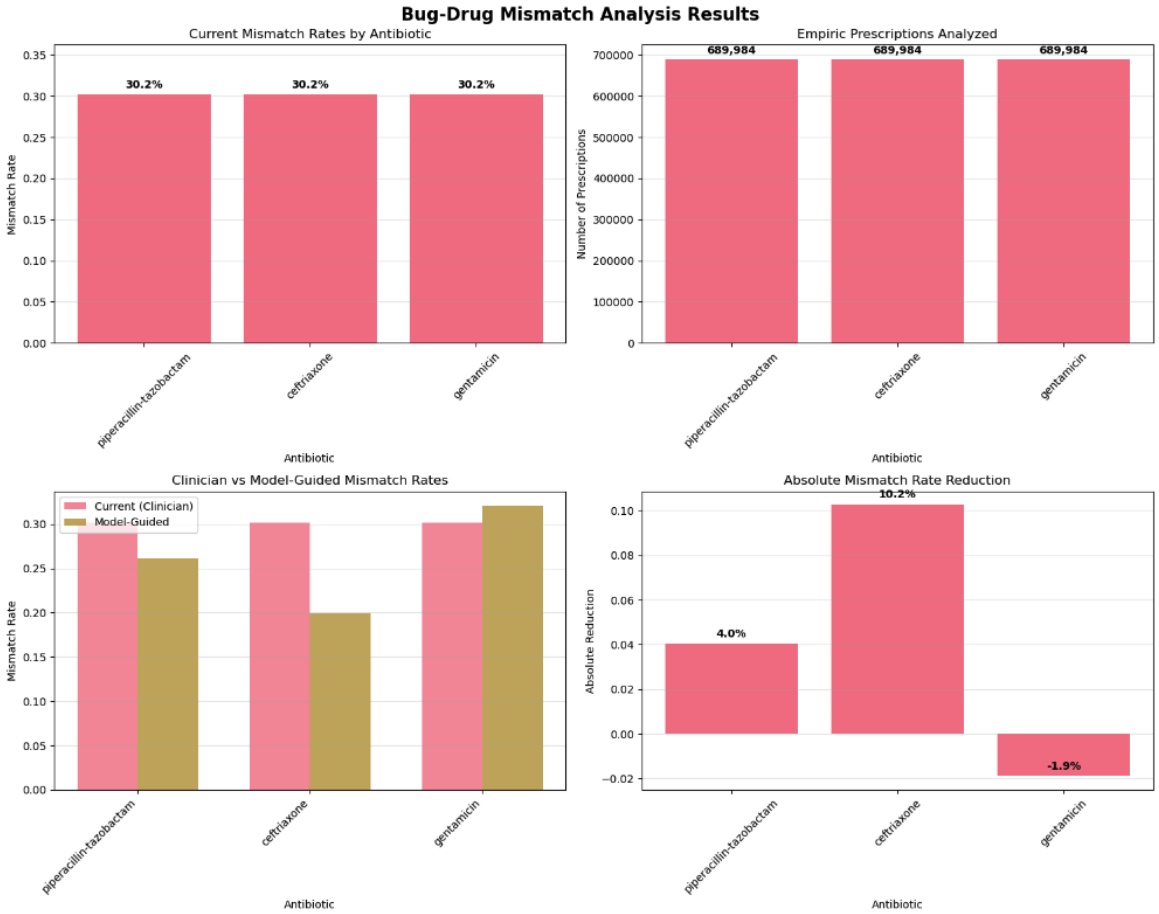


Figure 11. Summary comparison of mismatch rates across antibiotics.

Antibiotic	Threshold	Clinician mismatch rate	Model mismatch rate	Absolute reduction	Relative reduction	Cases avoided
Piperacillin–tazobactam	0.20	30.2%	26.2%	4.0%	13.4%	209,764
Ceftriaxone	0.20	30.2%	19.9%	10.2%	33.9%	312,100
Gentamicin	0.20	30.2%	32.1%	–1.9%	–6.3%	–128,456

Table 4. Represent sensitivity of mismatch reduction at different thresholds.

To evaluate the clinical utility of the models, bug–drug mismatch analysis was performed to compare observed empiric prescribing to simulated model-guided prescribing (Figure 11). Across 2,069,952 empiric prescriptions, baseline mismatch rate was 30.2%, representing the proportion of encounters where the empiric antibiotic was inactive against the isolated pathogen. This reflects a substantial risk of undertreatment under current empiric pathways.

Simulation of model-guided prescribing at a 20% resistance probability threshold demonstrated heterogeneous effects across antibiotics. For ceftriaxone, the model achieved the largest improvement, reducing mismatch by 10.2% in absolute terms (relative reduction 33.9%), equating to over 300,000 encounters where undertreatment could have been avoided. PTZ showed a smaller but still meaningful reduction of 4.0% (13.4% relative reduction, ~209,000 cases avoided). In contrast, model-guided prescribing for gentamicin paradoxically increased mismatch rates (−1.9%), suggesting that false reassurance from model predictions may have worsened undertreatment risk.

The predictive models may reduce undertreatment risk for broad-spectrum agents such as ceftriaxone and piperacillin–tazobactam, but highlight the potential for harm when applied indiscriminately, as illustrated by gentamicin. Overall, the model improved empiric antibiotic matching by an average of 4.1%, with the biggest gains for ceftriaxone (10.2%), supporting targeted implementation of AMR prediction models in high-mismatch settings.

4. DISCUSSION

Machine learning approaches have shown moderate accuracy in predicting antimicrobial resistance in bloodstream infections. In this study, supervised models were applied to structured clinical data; however, supplementary information such as antimicrobial susceptibility test results remains essential to improve forecasting accuracy (Doyle et al., 2020). Supervised learning has also been used to identify genetic traits linked to antibiotic sensitivity in Escherichia coli (Shaik et al., 2022), highlighting the potential value of integrating genomic and phenotypic data with clinical features to refine predictions.

The study illustrates that ML trained on UCLH electronic health records achieve modest discrimination in predicting antimicrobial resistance for Gram-negative bloodstream infections. Across all antibiotics and timepoints, performance was limited, with ROC-AUC values ranging from 0.55 to 0.73. There was an improvement from pre- to post-species, which highlight the importance of incorporating microbiological information (Lewin-Epstein et al., 2020).

Logistic regression consistently outperformed non-linear ensemble methods, offering superior discrimination, while random forest and XGBoost provided weaker and less stable performance. Incorporation of species-level features improved discrimination, particularly ceftriaxone and flucloxacillin, yet PTZ remained difficult to predict. This likely reflects some prevalence of resistance at UCLH (53%), which constrained model separability.

When compared with external studies, including the Oxford AMR model (Eyre et al., 2025), the UCLH models achieved lower discrimination. The Oxford group reported ROC-AUCs of 0.68–0.74 at baseline and 0.80–0.88 with species identification, compared to 0.55–0.73 in this study. Other studies using EHR-based prediction also have reported higher AUROCs (0.70–0.93) (Sakagianni et al., 2023). A major factor likely explaining this disparity is the markedly different prevalence of resistance, particularly for PTZ. At UCLH, PTZ resistance was observed in 53% of isolates, compared with ~7% in Oxford. This demonstrates how model separability decreases as the ratio of resistant to susceptible cases narrows.

Whereas Oxford and other studies reported that boosting methods generally outperformed simpler models (Lewin-Epstein et al., 2021), logistic regression consistently yielded the best results at UCLH. This divergence may reflect differences in cohort size, feature structure, and data sparsity, suggesting that simpler models can be more robust in certain local contexts (Nusinovici et al., 2020). The richer feature set and larger cohort available at Oxford likely enabled ensemble methods to generalise better, whereas at UCLH, logistic regression proved more reliable.

Despite differences in baseline prevalence and performance, the bug–drug mismatch analyses showed comparable improvements across sites. The Oxford model increased the proportion of patients receiving an active antibiotic by 5–9% depending on the agent (Eyre et al., 2025). At UCLH, mismatch rates were reduced by 10.2% for ceftriaxone and 4.0% for piperacillin–tazobactam, while predictions for gentamicin paradoxically increased mismatch by 1.9%. Overall, this translated to an improvement of ~4–10% across antibiotics, aligning with the range reported by Oxford. The larger gains observed for ceftriaxone are clinically important given its frequent empiric use in Gram-negative infections, whereas the poorer performance for gentamicin may reflect class imbalance and the narrower clinical role of aminoglycosides at UCLH (Alsowaida et al., 2022) (Chaves and Tadi, 2023). These findings indicate that predictive models can reduce mismatch for certain antibiotics but highlight the importance of tailoring model-guided prescribing strategies to individual agents.

Our dependence plots further extend previous findings by demonstrating threshold-type effects, whereby the presence of prior resistance in the same species sharply increases predicted risk. The lower performance observed in the UCLH cohort, together with the smaller gains from incorporating species features, underscores the significant influence of local epidemiology on predictive accuracy. This finding echoes Tang et al.'s study, which emphasised prevalence and setting-specific context as critical determinants of model performance.

Interpretation of clinical decision analysis

The clinical decision analysis highlights the practical implications of model selection for empiric prescribing. Logistic regression demonstrated a more favourable trade-off between overtreatment and undertreatment, particularly for piperacillin–tazobactam. This finding is clinically important: overtreatment contributes to unnecessary antibiotic exposure and resistance selection, while undertreatment risks patient harm through inadequate empiric coverage (Paul et al., 2010). The ability of logistic regression to sustain higher optimal treatment rates at moderate thresholds suggests that simple models may offer more reliable guidance where resistance prevalence is high and data are heterogeneous (Ardila, González-Arroyave and Tobón, 2025).

Nevertheless, random forest and XGBoost models produced less stable treatment recommendations, with greater fluctuations in overtreatment and undertreatment across thresholds. This instability limits their clinical interpretability and may reduce clinician confidence in model-guided prescribing. Feature importance analyses further identified abnormal microbiology results, neutrophil counts, and prior resistance history as key predictors, aligning with established risk factors and reinforcing the biological plausibility of the models.

Study	Setting & Country	Data Sources	Algorithms	Reported Performance
Sousa et al., 2019	Hospital admissions, Spain	Demographic + clinical data, blood cultures, AST	Decision Tree	AUC ≈ 0.76 for broad Gram-negative resistance
Feretzakis et al., 2020	Medical wards, Greece	Demographics, culture/AST data, Gram stain, specimen type	Multinomial Logistic Regression	AUC ≈ 0.76 across multiple bacterial species
Feretzakis et al., 2020	Intensive Care Unit, Greece	Demographics, culture/AST data, Gram stain, specimen type	LR, RF, k-NN, J48, MLP	AUC ≈ 0.73, highlighting feasibility of ML across varied models
McGuire et al., 2021	Hospital admissions, USA	Demographics, labs, vitals, billing codes, culture & AST (67 features)	XGBoost	AUC ≈ 0.85 for carbapenem resistance
Eyre et al., 2025 (Oxford)	Hospital bloodstream infections, UK	Linked EHR and microbiology, pre- and post-species	Gradient boosting ensembles	ROC-AUC 0.68–0.74 pre-species; 0.80–0.88 post-species

Table 5. Selected studies applying machine learning to predict antimicrobial resistance from clinical and microbiology data. Studies were chosen to illustrate a range of settings, data sources, and modelling approaches. Adapted from Sakagianni et al. (2023)

4.1 Strengths

The integration of multiple data sources enabled the construction of clinically meaningful predictors that reflect patient history, antimicrobial exposures, and laboratory findings, providing a robust foundation for model development. Stratification into pre- and post-species datasets, with 43 retained predictors, ensured alignment with real clinical decision stages and enabled evaluation of the added value of microbiological information. Different modelling approaches were compared systematically, and performance was complemented by an emphasis on interpretability. Confusion matrices and SHAP analyses demonstrated that key predictors, including microbiology results, neutrophil counts, and prior resistance history, aligned with established clinical knowledge, lending biological plausibility, and supporting clinician trust.

4.2 Limitations and Future work

This study has several limitations that should be acknowledged. First, as with other AMR prediction research (Lewin-Epstein et al., 2020; Eyre et al., 2025), the models developed here are predictive only. While SHAP analyses highlighted biologically plausible features such as prior resistance history, ICU admission, and recent antibiotic exposure, these associations cannot be interpreted causally. Establishing mechanistic links would require integration with experimental or molecular data.

Second, the dataset was restricted to hospital prescribing and microbiology records from UCLH, without linkage to community prescribing, outpatient pharmacy data, or contextual factors such as allergy history and clinician decision-making. The absence of these exposures, known to influence resistance dynamics (Roemhild et al., 2015; Kim et al., 2022), may have reduced model accuracy and contributed to paradoxical findings such as gentamicin, where low baseline resistance prevalence (16.5%) and its adjunctive role in sepsis management limited predictive utility. Sample sizes were also modest for some antibiotics (e.g., ceftriaxone n=443), which may have constrained discrimination.

Future work should prioritize external validation across independent hospitals, ideally within decision support systems for clinicians. Prospective evaluation in real-time clinical workflows would clarify the impact of model-guided prescribing on patient outcomes and stewardship. Incorporating molecular biomarkers, such as rapid species identification could complement EHR-based models and improve empiric-stage predictions.

CONCLUSION

Predicting antimicrobial resistance in bloodstream infections remains a major challenge for clinicians, where delays or inappropriate empiric therapy can have serious consequences. This study showed that, although machine learning models trained on routine hospital data achieved only modest accuracy, they nonetheless demonstrated potential to improve prescribing decisions compared with current practice. Even simple, interpretable models such as logistic regression could reduce undertreatment, while also highlighting situations where prediction is less reliable. These findings suggest that machine learning can complement microbiological testing by supporting earlier and more targeted empiric prescribing. The improvements observed, though modest, are clinically meaningful in the context of antimicrobial stewardship, where reducing unnecessary broad-spectrum use while ensuring adequate coverage is essential.

ETHICS

This project uses retrospective, anonymized electronic health record (EHR) data from UCLH, accessed securely via the SAFEHR platform. Data use complies with GDPR, UCL data protection policies, and the UCLH Data Access Process for Research (DAP-R). As the dataset contains no identifiable information, it is not classified as special category data, and formal NHS REC approval is not required. Analysis will be conducted within the SAFEHR environment, and findings will be reported to prevent re-identification. Data will be used solely for academic research on antimicrobial resistance prediction and clinical decision support.

ACKNOWLEDGEMENTS

I would like to thank to my supervisors, Dr Steve Harris for support and guidance throughout the project.

Special thanks also go to UCLH for providing access to the electronic health record and microbiology data, and to Professor David Eyre and the Oxford research group for their published work, which served as a valuable benchmark. Finally, I acknowledge the MSc Health Data Science programme at UCL for the training and resources that enabled this study.

REFERENCES

Alsowaida, Y.S., Benitez, G., Bin Saleh, K., Almangour, T.A., Shehadeh, F. and Mylonakis, E. (2022). Effectiveness and Safety of Ceftriaxone Compared to Standard of Care for Treatment of Bloodstream Infections Due to Methicillin-Susceptible *Staphylococcus aureus*: A Systematic Review and Meta-Analysis. *Antibiotics*, 11(3), p.375. doi:https://doi.org/10.3390/antibiotics11030375.

Antimicrobial Resistance Collaborators (2022). Global Burden of Bacterial Antimicrobial Resistance in 2019: a Systematic Analysis. *The Lancet*, 399(10325), pp.629–655. doi:https://doi.org/10.1016/S0140-6736(21)02724-0.

Ardila, C.M., González-Arroyave, D. and Tobón, S. (2025). Machine learning for predicting antimicrobial resistance in critical and high-priority pathogens: A systematic review considering antimicrobial susceptibility tests in real-world healthcare settings. *PLoS ONE*, 20(2), p.e0319460. doi:https://doi.org/10.1371/journal.pone.0319460.

Arumugham, V.B., Gujarathi, R. and Cascella, M. (2023). Third-Generation Cephalosporins. [online] NIH. Available at: https://www.ncbi.nlm.nih.gov/books/NBK549881/.

Barnett, K., Mercer, S.W., Norbury, M., Watt, G., Wyke, S. and Guthrie, B. (2012). Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*, 380(9836), pp.37–43. doi:https://doi.org/10.1016/S0140-6736(12)60240-2.

Bell, B.G., Schellevis, F., Stobberingh, E., Goossens, H. and Pringle, M. (2014). A systematic review and meta-analysis of the effects of antibiotic consumption on antibiotic resistance. *BMC Infectious Diseases*, 14(1). doi:https://doi.org/10.1186/1471-2334-14-13.

Biplab Singha, Singh, V. and Soni, V. (2024). Alternative therapeutics to control antimicrobial resistance: a general perspective. *Frontiers in Drug Discovery*, 4. doi:https://doi.org/10.3389/fddsv.2024.1385460.

Carrara, E., Pfeffer, I., Zusman, O., Leibovici, L. and Paul, M. (2018). Determinants of inappropriate empirical antibiotic treatment: systematic review and meta-analysis. *International Journal of Antimicrobial Agents*, 51(4), pp.548–553. doi:https://doi.org/10.1016/j.ijantimicag.2017.12.013.

Chaves, B.J. and Tadi, P. (2023). Gentamicin. [online] PubMed. Available at: https://www.ncbi.nlm.nih.gov/books/NBK557550/.

Doyle, R.M. *et al.* (2020). Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: an inter-laboratory study. *Microbial Genomics*, 6(2). doi:https://doi.org/10.1099/mgen.0.000335.

Feretzakis, G., Sakagianni, A., Loupelis, E., Kalles, D., Skarmoutsou, N., Martsoukou, M., Christopoulos, C., Lada, M., Petropoulou, S., Velentza, A., Michelidou, S., Chatzikyriakou, R. and Dimitrellos, E. (2021). Machine Learning for Antibiotic Resistance Prediction: A Prototype Using Off-the-Shelf Techniques and Entry-Level Data to Guide Empiric

Antimicrobial Therapy. *Healthcare Informatics Research*, 27(3), pp.214–221. doi:https://doi.org/10.4258/hir.2021.27.3.214.

Gall, E., Long, A. and Hall, K.K. (2020). Making healthcare safer III: A critical analysis of existing and emerging patient safety practices. [online] Agency for Healthcare Research and Quality (US). Available at: https://www.ncbi.nlm.nih.gov/books/NBK555533/.

Garcia-Vidal, C. *et al.* (2021). Machine Learning to Assess the Risk of Multidrug-Resistant Gram-Negative Bacilli Infections in Febrile Neutropenic Hematological Patients. *Infectious Diseases and Therapy*, 10(2), pp.971–983. doi:https://doi.org/10.1007/s40121-021-00438-2.

Goh, K.H. *et al.* (2021). Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature Communications*, 12(1), p.711. doi:https://doi.org/10.1038/s41467-021-20910-4.

Goodman, K.E., Lessler, J., Harris, A.D., Milstone, A.M. and Tamma, P.D. (2019). A methodological comparison of risk scores versus decision trees for predicting drug-resistant infections: A case study using ESBL bacteremia. *Infection Control & Hospital Epidemiology*, 40(4), pp.400–407. doi:https://doi.org/10.1017/ice.2019.17.

Habib, A.R. and Gross, C.P. (2023). FDA Regulations of AI-Driven Clinical Decision Support Devices Fall Short. *JAMA Internal Medicine*. doi:https://doi.org/10.1001/jamainternmed.2023.5006.

Ibrahim, E.H., Sherman, G., Ward, S., Fraser, V.J. and Kollef, M.H. (2000). The Influence of Inadequate Antimicrobial Treatment of Bloodstream Infections on Patient Outcomes in the ICU Setting. *Chest*, 118(1), pp.146–155. doi:https://doi.org/10.1378/chest.118.1.146.

Kang, C.-I. *et al.* (2004). Risk Factors for and Clinical Outcomes of Bloodstream Infections Caused by ESBL-Producing *Klebsiella pneumoniae*. *Infection Control & Hospital Epidemiology*, 25(10), pp.860–867. doi:https://doi.org/10.1086/502310.

Khan, N. *et al.* (2024). Guaranteeing Correctness in Black-Box Machine Learning: A Fusion of Explainable AI and Formal Methods for Healthcare Decision-Making. *IEEE Access*, 12, pp.90299–90316. doi:https://doi.org/10.1109/access.2024.3420415.

Kim, J.I. *et al.* (2022). Machine Learning for Antimicrobial Resistance Prediction: Current Practice, Limitations, and Clinical Perspective. *Clinical Microbiology Reviews*, 35(3). doi:https://doi.org/10.1128/cmr.00179-21.

Kollef, M.H. *et al.* (2021). Timing of antibiotic therapy in the ICU. *Critical Care*, 25(1). doi:https://doi.org/10.1186/s13054-021-03787-z.

Lewin-Epstein, O. *et al.* (2020). Predicting Antibiotic Resistance in Hospitalized Patients by Applying Machine Learning to Electronic Medical Records. *Clinical Infectious Diseases*, 72(11), pp.e848–e855. doi:https://doi.org/10.1093/cid/ciaa1576.

Ma, Z. *et al.* (2024). High mortality associated with inappropriate initial antibiotic therapy in hematological malignancies with *Klebsiella pneumoniae* bloodstream infections. *Scientific Reports*, 14(1). doi:https://doi.org/10.1038/s41598-024-63864-5.

Martínez-Agüero, S., Mora-Jiménez, I., Álvarez-Rodríguez, J., Marques, A.G. and Soguero-Ruiz, C. (2022). Interpretable clinical time-series modeling for early prediction of antimicrobial multidrug resistance. *Future Generation Computer Systems*, 133, pp.68–83. doi:https://doi.org/10.1016/j.future.2022.02.021.

Martínez-Agüero, S., Mora-Jiménez, I., Álvarez-Rodríguez, J. and Soguero-Ruiz, C. (2019). Machine Learning Techniques to Identify Antimicrobial Resistance in the ICU. *Entropy*, 21(6), p.603. doi:https://doi.org/10.3390/e21060603.

McGuire, R.J. *et al.* (2021). A Pragmatic Machine Learning Model To Predict Carbapenem Resistance. *Antimicrobial Agents and Chemotherapy*, 65(7). doi:https://doi.org/10.1128/aac.00063-21.

Melnyk, A.H., Wong, A. and Kassen, R. (2014). The fitness costs of antibiotic resistance mutations. *Evolutionary Applications*, 8(3), pp.273–283. doi:https://doi.org/10.1111/eva.12196.

Melzer, M. and Petersen, I. (2007). Mortality following bacteraemic infection caused by ESBL-producing *E. coli* compared to non-ESBL producing *E. coli*. *Journal of Infection*, 55(3), pp.254–259. doi:https://doi.org/10.1016/j.jinf.2007.04.007.

Merli, M. *et al.* (2015). The Spread of MDR Infections Is Leading to an Increase in Empirical Antibiotic Treatment Failure in Cirrhosis: A Prospective Survey. *PLOS ONE*, 10(5), p.e0127448. doi:https://doi.org/10.1371/journal.pone.0127448.

Nilsson, A.S. (2019). Pharmacological limitations of phage therapy. *Upsala Journal of Medical Sciences*, 124(4), pp.218–227. doi:https://doi.org/10.1080/03009734.2019.1688433.

Nusinovici, S. *et al.* (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, 122, pp.56–69. doi:https://doi.org/10.1016/j.jclinepi.2020.03.002.

O'Neill, J. (2016). Tackling drug-resistant infections globally: Final report and recommendations. *Archives of Pharmacy Practice*, 7(3). doi:https://doi.org/10.4103/2045-080x.186181.

Oshima, T. *et al.* (2016). Empiric Antibiotic Therapy for Severe Sepsis and Septic Shock. *Surgical Infections*, 17(2), pp.210–216. doi:https://doi.org/10.1089/sur.2014.096.

Paul, M. *et al.* (2010). Systematic Review and Meta-Analysis of the Efficacy of Appropriate Empiric Antibiotic Therapy for Sepsis. *Antimicrobial Agents and Chemotherapy*, 54(11), pp.4851–4863. doi:https://doi.org/10.1128/AAC.00627-10.

Perry, C.M. and Markham, A. (1999). Piperacillin/tazobactam: an updated review of its use in the treatment of bacterial infections. *Drugs*, 57(5), pp.805–843. doi:https://doi.org/10.2165/00003495-199957050-00017.

Phillips, I., Eykyn, S., King, B.A., Jenkins, C., Warren, C.A. and Shannon, K.P. (1977). The in vitro antibacterial activity of nine aminoglycosides and spectinomycin. *The Journal of Antimicrobial Chemotherapy*, 3(5), pp.403–410. doi:https://doi.org/10.1093/jac/3.5.403.

Richards, D.M., Heel, R.C., Brogden, R.N., Speight, T.M. and Avery, G.S. (1984). Ceftriaxone: A Review of its Antibacterial Activity, Pharmacological Properties and Therapeutic Use. *Drugs*, 27(6), pp.469–527. doi:https://doi.org/10.2165/00003495-198427060-00001.

Roemhild, R., Barbosa, C., Beardmore, R.E., Jansen, G. and Schulenburg, H. (2015). Temporal variation in antibiotic environments slows down resistance evolution in *Pseudomonas aeruginosa*. *Evolutionary Applications*, 8(10), pp.945–955. doi:https://doi.org/10.1111/eva.12330.

Sakagianni, A. *et al.* (2023). Using Machine Learning to Predict Antimicrobial Resistance—A Literature Review. *Antibiotics*, 12(3), p.452. doi:https://doi.org/10.3390/antibiotics12030452.

Shaik, S. *et al.* (2022). Genome Informatics and ML-Based Identification of AMR-Encoding Features and Virulence in *E. coli* Lineages. *mBio*, 13(1). doi:https://doi.org/10.1128/mbio.03796-21.

Shang, J.S., Lin, Y.E. and Goetz, A.M. (2000). Diagnosis of MRSA with neural networks and logistic regression. *Health Care Management Science*, 3(4), pp.287–297. doi:https://doi.org/10.1023/a:1019018129822.

Sousa, A. *et al.* (2018). Validation of a clinical decision tree to predict if a patient has a bacteraemia due to a β-lactamase producing organism. *Infectious Diseases*, 51(1), pp.32–37. doi:https://doi.org/10.1080/23744235.2018.1508883.

Tang, R. *et al.* (2022). Machine Learning in Predicting Antimicrobial Resistance: a Systematic Review and Meta-Analysis. *International Journal of Antimicrobial Agents*, p.106684. doi:https://doi.org/10.1016/j.ijantimicag.2022.106684.

Vasudeva, N., Nirwan, P.S. and Shrivastava, P. (2016). Bloodstream infections and antimicrobial sensitivity patterns in a tertiary care hospital of India. *Therapeutic Advances in Infectious Disease*, 3(5), pp.119–127. doi:https://doi.org/10.1177/2049936116666983.

Wolfe, C.M., Cohen, B. and Larson, E. (2014). Prevalence and risk factors for antibiotic-resistant community-associated bloodstream infections. *Journal of Infection and Public Health*, 7(3), pp.224–232. doi:https://doi.org/10.1016/j.jiph.2014.01.001.

Xu, J. *et al.* (2021). Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. *Briefings in Bioinformatics*, 22(5). doi:https://doi.org/10.1093/bib/bbab083.

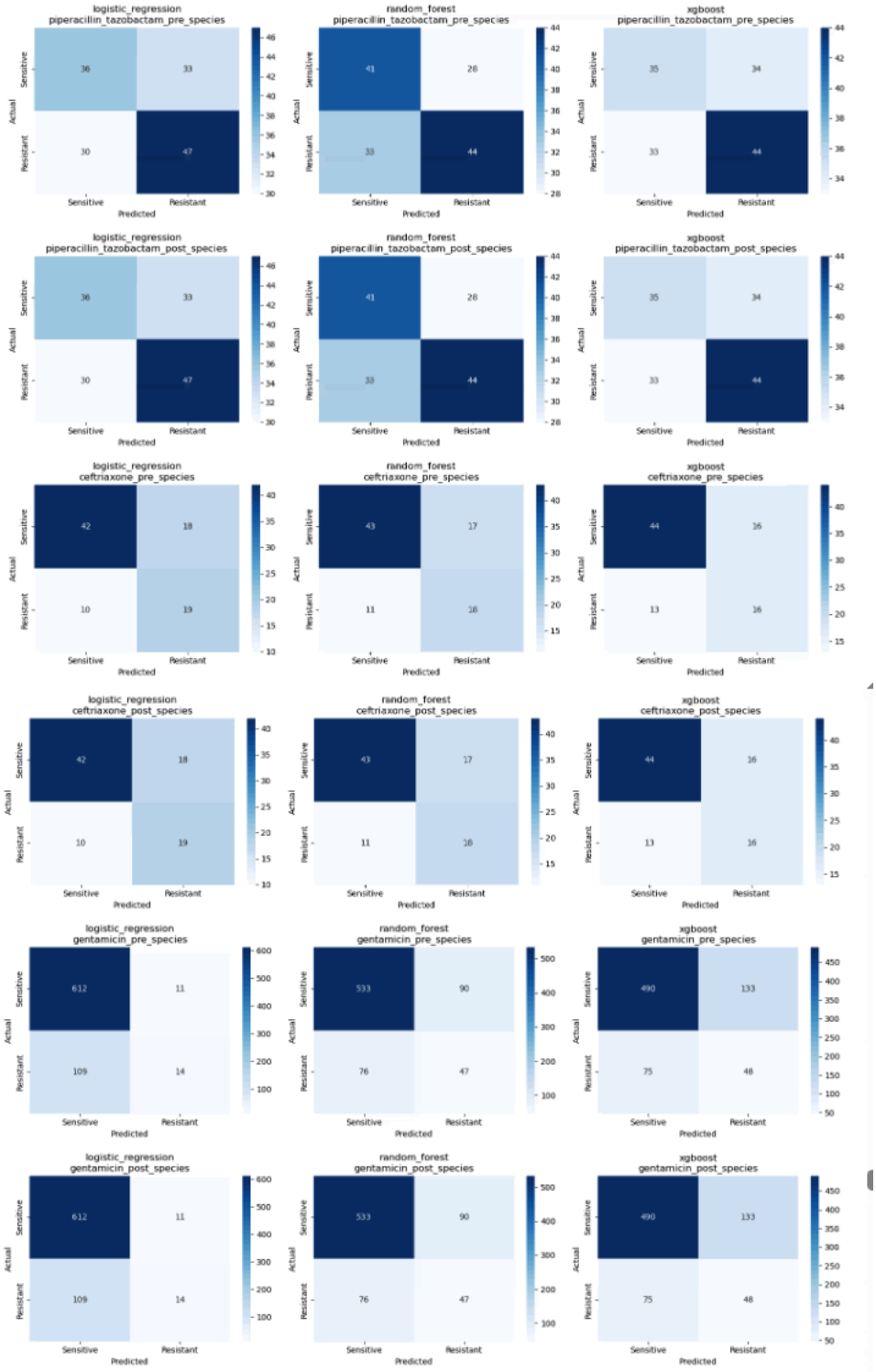
Yelin, I. *et al.* (2019). Personal clinical history predicts antibiotic resistance of urinary tract infections. *Nature Medicine*, 25(7), pp.1143–1152. doi:https://doi.org/10.1038/s41591-019-0503-6.

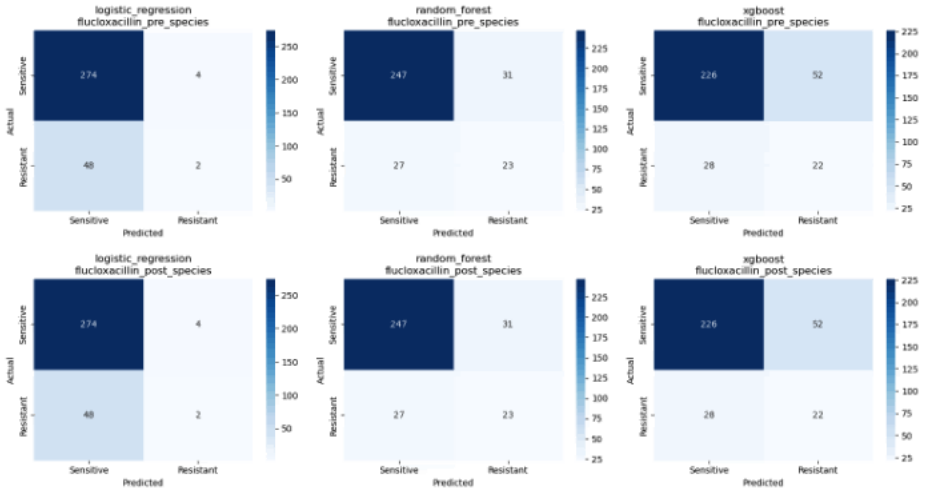
Zanichelli, V. *et al.* (2023). The WHO AWaRe antibiotic book and prevention of antimicrobial resistance. *Bulletin of the*

WHO, 101(4), pp.290–296. doi:https://doi.org/10.2471/blt.22.288614.

Zumla, A. (2010). *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*. *The Lancet Infectious Diseases*, 10(5), pp.303–304. doi:https://doi.org/10.1016/s1473-3099(10)70089-x.

APPENDIES


logistic_regression
piperacillin_tazobactam_pre_species
36
33
30
47
random_forest
piperacillin_tazobactam_pre_species
41
28
33
44
xgboost
piperacillin_tazobactam_pre_species
35
34
33
44
logistic_regression
piperacillin_tazobactam_post_species
36
33
30
47
random_forest
piperacillin_tazobactam_post_species
41
28
33
44
xgboost
piperacillin_tazobactam_post_species
35
34
33
44
logistic_regression
ceftriaxone_pre_species
42
18
10
19
random_forest
ceftriaxone_pre_species
43
17
11
18
xgboost
ceftriaxone_pre_species
44
16
13
16
logistic_regression
ceftriaxone_post_species
42
18
10
19
random_forest
ceftriaxone_post_species
43
17
11
18
xgboost
ceftriaxone_post_species
44
16
13
16
logistic_regression
gentamicin_pre_species
612
11
109
14
random_forest
gentamicin_pre_species
533
90
76
47
xgboost
gentamicin_pre_species
490
133
75
48
logistic_regression
gentamicin_post_species
612
11
109
14
random_forest
gentamicin_post_species
533
90
76
47
xgboost
gentamicin_post_species
490
133
75
48
Actual
Sensitive
Resistant
Predicted
Sensitive
Resistant



Appendix A1. Confusion matrices for antibiotic-specific models.

	antibiotic	empiric_prescriptions	with_outcomes	total_prescriptions	resistant_cases	mismatch_rate	model_mismatch_rate	absolute_reduction	relative_reduction
1	piperacillin-tazobactam	10574	689984	689984	208315	0.30191279797792414	0.2615575371975352	0.040355260780388924	13.366528696587318
2	ceftriaxone	10574	689984	689984	208315	0.30191279797792414	0.19942098633443067	0.10249181164349347	33.94748825817833
3	gentamicin	10574	689984	689984	208315	0.30191279797792414	0.32087438525837125	-0.018961587280447112	-6.280484764953085

ceftriaxone

	threshold	clinician_mismatch_rate	model_mismatch_rate	absolute_reduction	relative_reduction	cases_avoided	proportion_avoided
1	0.1	0.30191279797792414	0.16683656621057014	0.135076231767354	44.74014770888605	550686	0.7981141591689083
2	0.15	0.30191279797792414	0.3157975108803543	-0.013884712902430152	-4.598914983208298	349462	0.5064784110935906
3	0.2	0.30191279797792414	0.26746404232464643	0.03444875565327771	11.410167400653416	241262	0.34966318059549206
4	0.25	0.30191279797792414	0.4219277102631585	-0.12001491228523437	-39.751515367695625	274683	0.39810053566459513
5	0.3	0.30191279797792414	0.30516133377708043	0.0032485357991562935	-1.0759847945875505	185192	0.2684004266765606

gentamicin

	threshold	clinician_mismatch_rate	model_mismatch_rate	absolute_reduction	relative_reduction	cases_avoided	proportion_avoided
1	0.1	0.30191279797792414	0.29161930829718635	0.010293489680737788	3.4094247576382797	560651	0.8125565230498099
2	0.15	0.30191279797792414	0.1574048694450072	0.14450792853291694	47.86412815248837	438251	0.6351611051850478
3	0.2	0.30191279797792414	0.30893890442353805	-0.007026106445613911	-2.3271972876511384	303393	0.43971019617846213
4	0.25	0.30191279797792414	0.271399059908386	0.030513738069538165	10.106805101971638	155568	0.22546009776458584
5	0.3	0.30191279797792414	0.27799345490787725	0.023919343070046895	7.9225999130371	106035	0.15367747657916705

piperacillin

	threshold	clinician_mismatch_rate	model_mismatch_rate	absolute_reduction	relative_reduction	cases_avoided	proportion_avoided
1	0.1	0.30191279797792414	0.38433269105618434	-0.0824198930782602	-27.299237935679276	611617	0.8864220039884982
2	0.15	0.30191279797792414	0.16606922147241995	0.1358435765055042	44.99430875912623	397907	0.5766901841202114
3	0.2	0.30191279797792414	0.29517301240067234	0.0067397855772518	2.232361669459475	236057	0.3421195274093312
4	0.25	0.30191279797792414	0.34932938980429534	-0.0474165918263712	-15.705393128064186	261941	0.37963344077543826
5	0.3	0.30191279797792414	0.2843970356086545	0.01751576236926966	5.801596516140536	183843	0.26644530887672757

Appendix A2. Bug–drug mismatch analysis.
Observed mismatch rates compared with model-predicted mismatch rates at thresholds of 10–30% for ceftriaxone, gentamicin, and piperacillin–tazobactam. Results are shown as absolute and relative reductions in undertreatment risk, with corresponding case counts.

```
Pre-Species Model:
==================================================
SHAP values shape: (200, 13, 2)
X_sample shape: (200, 13)
Detected 3D SHAP values, taking first class for analysis...

🏆 Top 10 Most Important Features:
 1. Same Species Resistance History | Importance: 0.1197 | ↑ Increases resistance risk
 2. ICU Admission                   | Importance: 0.0706 | ↑ Increases resistance risk
 3. Prior Resistance Episodes       | Importance: 0.0324 | ↓ Decreases resistance risk
 4. Recent Gentamicin Use           | Importance: 0.0245 | ↑ Increases resistance risk
 5. Days Since Last Resistant       | Importance: 0.0197 | ↑ Increases resistance risk
 6. Length of Stay (days)           | Importance: 0.0196 | ↑ Increases resistance risk
 7. creatinine                      | Importance: 0.0186 | ↑ Increases resistance risk
 8. Neutrophil Count                | Importance: 0.0099 | ↑ Increases resistance risk
 9. Patient Age                     | Importance: 0.0091 | ↑ Increases resistance risk
10. gender_encoded                  | Importance: 0.0077 | ↑ Increases resistance risk
```

```
Post-Species Model:
==================================================
SHAP values shape: (200, 14, 2)
X_sample shape: (200, 14)
Detected 3D SHAP values, taking first class for analysis...

🏆 Top 10 Most Important Features:
 1. Same Species Resistance History | Importance: 0.1131 | ↑ Increases resistance risk
 2. ICU Admission                   | Importance: 0.0676 | ↑ Increases resistance risk
 3. Prior Resistance Episodes       | Importance: 0.0373 | ↓ Decreases resistance risk
 4. Recent Gentamicin Use           | Importance: 0.0209 | ↑ Increases resistance risk
 5. Length of Stay (days)           | Importance: 0.0190 | ↑ Increases resistance risk
 6. creatinine                      | Importance: 0.0185 | ↑ Increases resistance risk
 7. Days Since Last Resistant       | Importance: 0.0141 | ↑ Increases resistance risk
 8. species_encoded                 | Importance: 0.0137 | ↑ Increases resistance risk
 9. Neutrophil Count                | Importance: 0.0092 | ↑ Increases resistance risk
10. Patient Age                     | Importance: 0.0080 | ↑ Increases resistance risk
```

Appendix A3. Supplementary feature ranking outputs.
Top 10 predictors from SHAP analysis of pre-species and post-species models. Features such as prior same-species resistance, ICU admission, and recent antibiotic use consistently ranked as the most influential, supporting biological plausibility and alignment with established clinical risk factors.