2025 AMR Surveillance Data Challenge

Proposal Title: Modeling Antibiotic Resistance in WHO-Priority Pathogens in Kenya: A Machine Learning and Epidemiological Approach

Date of Submission (dd-mmm-yy): 23 – 07 -2025

Is the team entering for the AMR Student Innovation Award? (Yes/No) **Yes**

Research Team Members details *(put the Lead Applicant 1st in the table)*:

Team Member Name	**Role in the Data Challenge**	**Affiliation**	**Email**	**Country**	**Are they a student? Yes/No**
George Wanjiru	Infectious Disease Epidemiologist	University of Cambridge	ggw24@cam.ac.uk	United Kingdom/Kenya	Yes
Newton Lijoodi	ML Engineer	Moi University	nmlijoodi@gmail.com	Kenya	Yes

Datasets included in the analysis (Tick all those that apply):

- □ GSK – SOAR 201818
- □ GSK – SOAR 201910
- □ Johnson & Johnson – Bedaquiline (DREAM)
- □ Paratek - KEYSTONE
- ✓ Pfizer – ATLAS_Antibiotics
- □ Pfizer – ATLAS_Antifungals
- □ Venatorx – GEARS
- □ PLEA I – Venus Remedies
- □ PLEA II – Venus Remedies
- □ GASAR – Venus Remedies
- □ Surveillance of global clinical isolates of Acinetobacter baumannii-calcoaceticus complex – Innoviva
- □ Shionogi – SIDERO-WT
- □ Other
  Please provide details:____________________________________

Modeling Antibiotic Resistance in WHO-Priority Pathogens in Kenya: A Machine Learning and Epidemiological Approach

Executive Summary

This project leverages the Pfizer ATLAS dataset to analyze antibiotic resistance (ABR) trends in Kenya, with a focus on eight WHO-priority pathogens: *Escherichia coli*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Shigella* spp., *Salmonella* spp., and *Pseudomonas aeruginosa*. The pathogens are responsible for significant clinical and public health burdens in Kenya and across similar low- and middle-income countries (LMICs). We apply machine learning (ML) techniques; including XGBoost and SHAP-based model interpretation to identify predictors of multi-drug resistance and quantify the risk of resistance at the individual and population levels in Kenya and similar LMIC settings. The models incorporate available clinical and demographic variables to uncover resistance patterns across age, sex, patient status (in/out), source of infection and specialty. The analysis is designed to support clinical decision-making by flagging high-risk profiles and informing empirical treatment guidelines using a clinical flagging tool that we developed and the stewardship priority analytic plot which can be incorporated to the flagging tool. These insights can inform national policy and strengthen health systems by integrating predictive tools into clinical and stewardship practice.

1. Objectives

This study aimed to:

1. Map antibiotic resistance (ABR) trends among WHO-priority pathogens in Kenya using the ATLAS dataset (2013–2023).
2. Model multidrug resistance (MDR) using interpretable machine learning algorithms with clinical and demographic predictors and stratify by demographics and clinical variables
3. Develop data-driven clinical and stewardship tools to inform Kenya's AMR action plan.

2. Methods

2.1 Dataset

We utilized the Pfizer ATLAS dataset which contains 10 years of Kenyan ABR surveillance data (2013-2023) and extracted data for 8 WHO priority pathogens (*E.coli, K. pneumoniae, S. aureus, S. pneumoniae, Salmonella* spp., *Shigella* spp., *P. aeruginosa,* and *A. baumannii).* In total, we had a sample size of over 36000 isolates from across Kenya and other key variables such as binary resistance results, year, infection source, inpatient status, species, gender and age.

2.2 Data Processing

We began by processing our data to filter the Kenyan ABR data and address missingness using a reference table for treatment options for various bacteria and multiple imputation for model data. We also filtered various columns that were uninformative such as country since we were only focusing on Kenya and state since these fields lacked variability or completeness.

2.3 Modeling MDR and Resistance by Individual/Population Factors

Logistic regression was used as a baseline model due to its interpretability and suitability for binary outcomes. This enabled us to assess the association between MDR and individual-level covariates (e.g., age group, gender, sample source). XGBoost (Extreme Gradient Boosting) was implemented to enhance predictive performance and capture complex interactions among variables. Model validation was performed using a 70/30 train-test split alongside 5-fold cross-validation to assess generalizability and avoid overfitting. To model binary resistance outcomes for individual antibiotics, we trained a separate

XGBoost model using key predictors such as: Age group, Gender, Sample source (e.g., urine, blood) and Clinical specialty (e.g., inpatient, outpatient, ICU).

2.4 Clinical Prediction and Application

To support clinical decision-making, we developed a risk stratification approach based on the predicted MDR probability. Individuals with predicted probabilities ≥ 0.75 were flagged as high-risk for MDR, enabling potential prioritization for targeted stewardship or confirmatory testing. To enhance explainability at the patient level: We used SHAP waterfall plots to visualize individual prediction pathways and risk profiles.

2.5 Stewardship Prioritization

To establish stewardship prioritization, we calculated resistance proportions across the priority antibiotics, classed as per the WHO AWaRe groups (Access, Watch and Reserve). We then flagged high-risk combinations (Watch – resistance ≥50%, Reserve – resistance ≥30%) and developed Figure 2 to visualize this.

3. Results

Our analyses revealed distinct patterns of ABR across WHO-priority pathogens in Kenya from 2013 to 2023. These findings, in detail, are as summarized below:

3.1 Summary Statistics

E. coli and *K. pneumoniae* were the most commonly isolated pathogens (each >8000 isolates). Resistance rates (Figures 3 and 4) across the years were however, higher among *A. baumannii* (75-100%) , *E. coli* (87-100%) and *K. pneumoniae* (91-100%). Inpatient isolates also had higher overall resistance and MDR rates compared to outpatient samples. Lower levels were also recorded in *S. pneumoniae* (57-64%%) and *S. aureus* (0-50%) and *P. aeruginosa* (25-40%). Most 100% resistance was reported mainly post the breaking point (mid-2018) for all three species that reported 100% resistance at a point. The data also provides key summaries based on key demographics and clinical factors included and this allowed us to have a precursor to the ML models (In R code – GitHub link).

3.2 MDR Modelling

The machine learning models demonstrated robust performance in predicting MDR status across bacterial species. The logistic regression model achieved an average AUC of 0.79, while the XGBoost model improved predictive accuracy to an AUC of 0.86. Key predictors consistently associated with MDR included: Bacterial species, Inpatient status, Age group, Infection source and Sample year. The XGBoost model showed strong internal validation metrics: AUC: 0.86, Sensitivity: 82%, Specificity: 77% and F1 Score: 0.79.

3.3 SHAP Analysis

Interpretability using SHAP (Shapley Additive Explanations) identified the most influential features contributing to individual MDR predictions. High-risk predictions were strongly driven primarily by: Inpatient status (hospitalization), Older age and Infection source. This aligned with the global variance importance features but individualised the features to individual patient level bringing in more nuances such as older age being a higher risk compared to younger to middle ages (19-64).

The SHAP analysis also uncovered nonlinear interactions and individual-level variation in MDR risk factors. Based on these insights, a clinical decision-support tool was developed to: Flag high-risk patients and visualize personalized risk explanations using SHAP waterfall plots. A sample top 10 high-risk output by isolate ID for individual patients is shown in Table 1. It could also be integrated into health information systems where possible and include recommended next steps.

3.3 Resistance Trends and Stewardship Implications (Figure 2)

We observed that; Access group antibiotics showed a wide range of resistance. Ampicillin had the highest resistance overall at 95.4%, followed by Penicillin at 59.8%, indicating widespread resistance to these commonly used first-line agents. Watch antibiotics demonstrated variable resistance. Critically high resistance was observed in: Erythromycin (53.7%) and Ceftazidime (50.5%). Other Watch drugs showed moderate resistance: Cefepime (46.5%), Clindamycin (41.5%), Levofloxacin (35.6%), Piperacillin-tazobactam (32.5%) and amoxycillin clavulanate (26.4%). Meanwhile, Amikacin (20.8%), Meropenem (13.6%), and Ceftriaxone (4.9%) demonstrated comparatively lower resistance. There was no reported resistance for *linezolid* which is the only reserve antibiotic reported, reaffirming its status as a critical last-resort antibiotic.

4. Impact of the Work

This project contributes to both AMR science and policy through the following ways. To our knowledge, it delivers the first application of interpretable machine learning-driven mapping of MDR patterns across eight WHO-priority pathogens in Kenya, offering granular insights by region, age, gender and infection source. These findings can directly support data-informed prescribing, particularly empirical therapy when combined with the stewardship priority data, by alerting clinicians to high-resistance zones and individual risk factors.

Limitations

While these analyses were strong and robust and provided key insights on ABR and MDR in Kenya, there were several limitations. Firstly, while we initially aimed to cover a 20-year period, the data was limited to 10 years (2013-2023). While 10 years provides sufficient temporal resolution, longer timelines could reveal cyclical patterns. The mid-2018 breakpoint changes also require in-depth analysis to identify potential changes that could have influenced resistance thresholds. Additionally, there was no regional data and this limited the possibility of a geospatial analysis. Further, Kenya has no official drug cost list and the cost varies by region and type of service (private or public), thus limiting a cost analysis. There may also be potential biases introduced by the data collection methods including data from urban hospitals or those with reporting systems and also microbiological processes robustness. Finally, only internal validation for the models could be done due to the non-availability of comparable independent data.

Key Figures and Tables

Figure 1: Clinical MDR Risk Distribution

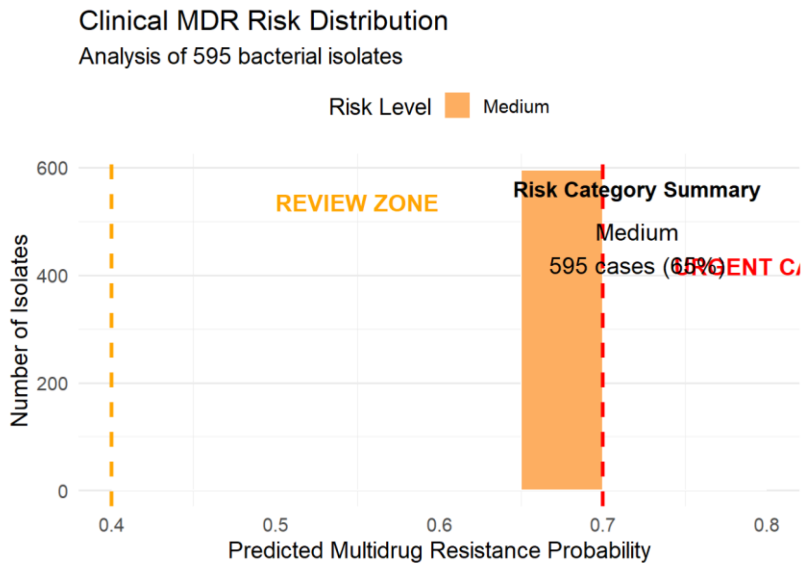

Figure 2: Stewardship Priority Map (By WHO AWaRe Classification)

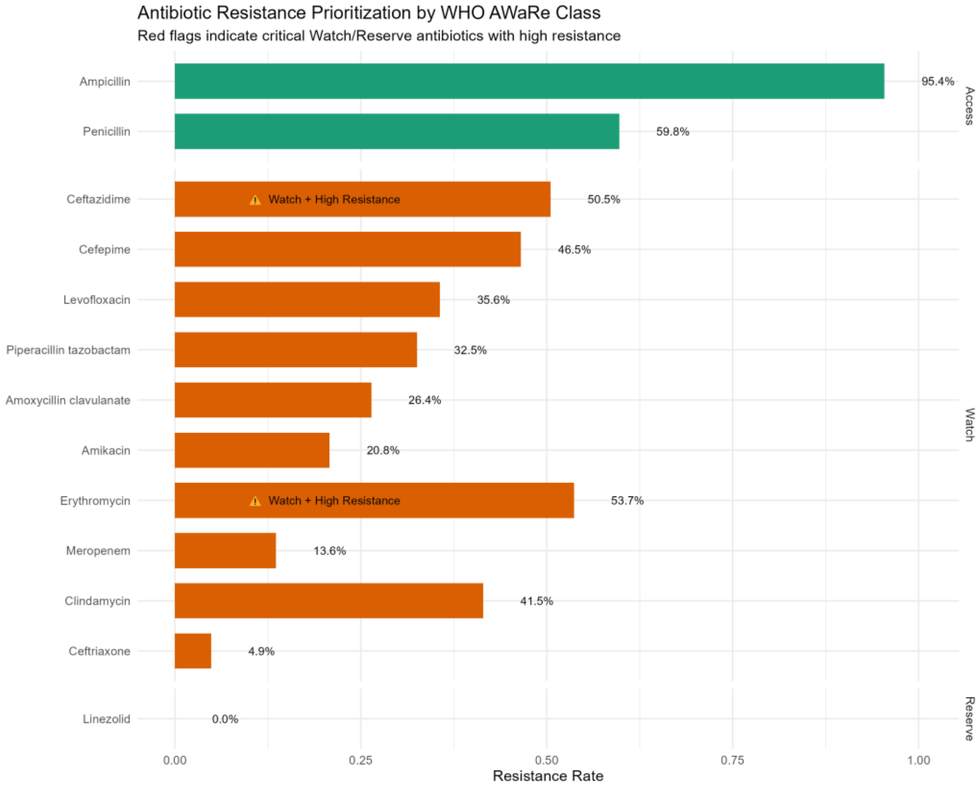

Table 1: Clinical Flagging Tool for top 10 High-Risk Isolates Output

Isolate Id	MDR Probability	Risk Score	Risk Level	Infection Source	Patient Type	Ward

Isolate Id	MDR Probability	Risk Score	Risk Level	Infection Source	Patient Type	Ward
39429	1.000	100	High	Genitourinary: Other	Outpatient	Medicine General
1106363	1.000	100	High	Colon	Inpatient	Medicine General
1147342	1.000	100	High	Cellulitis	Inpatient	Medicine General
1106674	0.918	92	High	Respiratory: Other	Inpatient	General Unspecified ICU
1046192	0.857	86	High	Respiratory: Other	None Given	None Given
1106397	0.857	86	High	Respiratory: Other	None Given	None Given
1147323	0.852	85	High	Urine	None Given	None Given
39471	0.848	85	High	Respiratory: Other	Outpatient	Surgery ICU
1058988	0.848	85	High	Respiratory: Other	Inpatient	Surgery ICU
1147334	0.839	84	High	Urine	None Given	Other

References

1. World Health Organization. WHO Publishes List of Bacteria for Which New Antibiotics Are Urgently Needed [Internet]. World Health Organization. 2017. Available from: https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed

2. Moja L, Zanichelli V, Mertz D, Gandra S, Cappello B, Cooke GS, et al. WHO's Essential Medicines and AWaRe: Recommendations on first- and second-choice Antibiotics for Empiric Treatment of Clinical Infections. Clinical Microbiology and Infection [Internet]. 2024 Feb 9; Available from: https://www.sciencedirect.com/science/article/pii/S1198743X24000594

3. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software [Internet]. 2011;45(3). Available from: https://www.jstatsoft.org/article/view/v045i03

4. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions [Internet]. arXiv.org. 2017. Available from: https://arxiv.org/abs/1705.07874v2

5. Chen T, Guestrin C. XGBoost: a Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. 2016 Aug 13;1(1):785–94.

6. Republic of Kenya. National Policy on Prevention and Containment of Antimicrobial Resistance [Internet]. React Group Africa. React Group Africa; 2017 [cited 2025 Jul 23]. Available from: https://www.reactgroup.org/wp-content/uploads/2021/08/Kenya-AMR-NAP-2017.pdf