Using Artificial Intelligence to predict antimicrobial resistance data disparities across Africa

Sylvester Zibusiso Moyo[1,2*], Alex Katumba[2,3], Charity Serwaa[4], Evance Juma[5], Stephen Obol Opiyo[2,6,7,8]

[1]African Society for Laboratory Medicine, Addis Baba, Ethiopia.
[2]SynBio Africa, Entebbe, Uganda.
[3]Kampala Hospital /Dulwich Health Laboratory, Kampala, Uganda.
[4]Council for Scientific and Industrial Research-Water Research, Ghana.
[5]Nkhotakota District Hospital/ Malawi University of Science and Technology, Malawi.
[6]Patira Data Science, Westerville, Ohio, USA; Kampala, Uganda.
[7]MAGMA Consultants International, Kampala, Uganda.
[8]The Ohio State University, Columbus, Ohio, USA
[*]Team Lead

ABSTRACT

This study investigates the regional variations in antimicrobial resistance (AMR), focusing on *Acinetobacter baumannii* and *Enterococcus faecium*, using data from the ATLAS and the WHO Bacterial Priority Pathogens List 2024 (BPPL 2024). We analyzed the resistance of *A. baumannii* to Meropenem across six regions (Africa, Asia, Europe, Latin America, North America, and Middle East), categorizing them into three distinct groups based on resistance levels. The use of Chow statistics confirmed significant regional differences. Additionally, we examined the resistance patterns of *E. faecium* to Vancomycin, revealing the highest resistance levels in North America and the lowest in Africa. To support research and development, a dashboard was created to identify isolates suitable for "omics" studies, which selected 16 isolates resistant to multiple antibiotics. Furthermore, a comparative analysis of machine learning models demonstrated that XGBoost (AUC = 1) outperformed Random Forest (AUC = 0.71) in predicting *Streptococcus pneumonia* to Penicillin resistance, highlighting its potential in regions with limited laboratory resources. The study's findings provide critical insights for researchers, healthcare professionals, and policymakers in understanding and combating AMR. However, limitations in data availability, particularly in underrepresented regions, suggest the need for more comprehensive surveillance to enhance the global understanding of AMR.

INTRODUCTION

In 2017, the World Health Organization (WHO) created the first Bacterial Priority Pathogens List (BPPL 2017) to direct research and development efforts for new antibiotics (1). The list categorized 13 bacterial phenotypes into three priority tiers: critical, high, and medium. This framework has been instrumental in guiding antimicrobial resistance (AMR) surveillance and prevention initiatives, such as the Global Antimicrobial Resistance Surveillance System (GLASS [2]). In BPPL 2024 (3), the WHO updated the BPPL 2017, removing five pathogens and incorporating new ones. Critical group bacterial pathogens, like *A. baumannii*, pose severe public health threats due to their limited treatment options, high transmissibility, and rising antibiotic resistance. High group pathogens, such as *E. faecium*, are challenging to treat, contribute significantly to disease burdens, and have increasing resistance and high transmissibility, making them critical in certain regions. Medium group pathogens, exemplified

by *S. pneumoniae*, present moderate treatment challenges and have varying resistance trends, with several treatment options in development and potential critical impacts in specific populations and areas. The study used the ATLAS database and an AI model to analyze and predict penicillin resistance trends in *A. baumannii* to Meropenem, *S. pneumoniae* to penicillin and *E. faecium* to Vancomycin. It underscores the need for region-specific analysis due to geographic variability in resistance, aiming to improve AMR surveillance and guide targeted interventions.

OBJECTIVES

Objective 1
Determine the percentage of resistance of the critical pathogen *A. baumannii* to the carbapenem antibiotic Meropenem across all regions represented in the ATLAS dataset. This pathogen has been categorized as critical since BPPL 2017 and continues to be in BPPL 2024. Meropenem is commonly used to treat a wide range of infections.

Objective 2
Analyze resistance patterns of the high-priority pathogen *E. faecium* to Vancomycin, as documented in the BPPL 2024. Develop a dashboard for the region with the highest resistance percentage and identify samples for further research and development.

Objective 3
Use machine learning (AI); Random Forest and XGBoost methods to predict the resistance of *S. pneumoniae* to penicillin in Africa using age group, source of samples, In and Outpatient, and gender as predictor variables. Over the past decade, *S. pneumoniae* has shown increasing AMR, complicating treatment strategies. Currently, 16% of *S. pneumoniae* samples in the ATLAS database are resistant to penicillin, the primary antibiotic used for treatment.

METHODS
Bacterial and antibiotics selection were based on information on BPPL 2024, and the commonly used first line antibiotics. The ATLAS dataset was divided into six regions (Africa, Asia, Europe, Latin America, Middle East, and North America) discussed in Opiyo et al. (5), presented in the Vivli AMR Open Data Reuse Data Challenge (6).

Descriptive Statistical Analysis and GIS Dashboard
Descriptive statistics were computed using R version 4.4.0 (7) to provide insights into data distribution. To visualize variations in antibiotic resistance among different pathogens, an interactive dashboard for North America was developed using ArcGIS software (8) discussed in Opiyo et al, (5).

Machine Learning Modeling
Time series was calculated using R version 4.4.0, the machine learning method used was XGBoost. Chow statistics test was used to calculate significant difference among regions. The R scripts are available on GitHub (https://github.com/so13839/SynBio-Team1).

RESULTS

Objective 1
We analyzed the resistance of *A. baumannii* to Meropenem over time for the six regions included in the ATLAS dataset (Figure 1). The distribution of resistance was categorized into

three groups: Africa, Latin America, and the Middle East, where the resistance percentage exceeded 75%; Asia and Europe, where the resistance ranged between 50% and 75%; and North America, where the resistance was below 50%. To assess the differences in resistance across these categories, we employed Chow statistics, as shown in Table 1. The p-values below 0.05 confirmed the three distinct categories. Our results indicate that the resistance of *A. baumannii* to Meropenem varies significantly across regions. By integrating data from the BPPL 2024, we utilized the ATLAS dataset to elucidate this distribution.

Objective 2.

The results reveal that *E. faecium* displays varying resistance patterns to Vancomycin across different regions (Figure 2). In Africa, where data points are limited, the resistance percentage is relatively low. In contrast, regions such as the Middle East and Latin America exhibit higher resistance percentages than Asia and Europe, but lower than North America, which shows the highest resistance levels, exceeding 50%. To address this, we developed a dashboard for North America to identify isolates suitable for research and development (Figure 3). These isolates are valuable for "omics" studies to uncover genes, metabolites, and proteins relevant to antimicrobial resistance (AMR) mitigation. Figure 3A indicates that there are 2,443 Vancomycin-resistant *E. faecium* isolates, with 2,403 of these also being susceptible to Linezolid (Figure 3B). This suggests that the dashboard is effective for selecting appropriate antibiotics for treating *E. faecium*. For the "omics" study, we selected 16 isolates resistant to Linezolid (Figure 4A). Among these, 12 isolates were resistant to Vancomycin, while 4 were susceptible. All 16 isolates are also resistant to Ampicillin, Penicillin, and Levofloxacin. These isolates can be requested from Pfizer for further "omics" research. This selection aligns well with one the objectives of the BPPL 2024, that is research and development. The dashboard can be found at:
https://patira.maps.arcgis.com/apps/dashboards/28c9c3bee2db491c9a6d2d82487fccd0

Objective 3

In our comparative analysis of Random Forest and XGBoost for predicting Penicillin antibiotic resistance using variables from ATLAS, XGBoost outperformed Random Forest, achieving an AUC of 1.00, while Random Forest achieved an AUC of 0.71. The superior predictive accuracy of the XGBoost model underscores its effectiveness in distinguishing resistant cases from susceptible ones. In regions like Africa, where resources for comprehensive laboratory testing and infrastructure are often limited, this machine learning approach offers a crucial solution. By enabling rapid and accurate predictions of antibiotic resistance in new datasets, the XGBoost model can significantly enhance existing diagnostic capabilities. This is especially important in areas with limited access to advanced laboratory facilities, where implementing such computational tools can help improve the monitoring and management of antibiotic resistance more efficiently.

The Impact of the Study

The study on antibiotic resistance (AMR) offers significant insights for researchers, healthcare professionals, policymakers, and pharmaceutical companies. By highlighting regional variations in resistance, particularly the high-resistance areas in Africa, Latin America, and the Middle East, the research underscores the need for targeted, region-specific investigations into AMR mechanisms. The development of a dashboard for "omics" studies facilitates the identification of critical genetic and metabolic factors, enabling more focused research and the creation of targeted therapeutic strategies. For healthcare practitioners, the study provides crucial data on resistance patterns, improving treatment decisions and

ensuring effective patient care based on regional resistance profiles. Policymakers benefit from the study's clear categorization of resistance levels, which aids in resource allocation and the implementation of tailored AMR control strategies, particularly in high-resistance regions. Pharmaceutical companies gain from the identification of key isolates for research, supporting the development of new antibiotics. The success of the XGBoost model in predicting resistance patterns highlights the potential of machine learning tools in drug development and AMR research. Additionally, the study enhances global health systems by improving diagnostic capabilities, particularly in resource-limited settings, and supports data-driven interventions to manage AMR more effectively worldwide.

Conclusions and limitations of the study

The study highlights regional differences in antimicrobial resistance (AMR), especially with *A. baumannii's* resistance to Meropenem and *E. faecium's* resistance to Vancomycin. It demonstrates that machine learning models like XGBoost can improve predictive accuracy, providing practical tools for regions with limited lab resources. The development of a dashboard to select suitable isolates for research and treatment is a significant advancement in AMR management. However, the study is constrained by data limitations, particularly in Africa, which may impact the generalizability of the results. This emphasizes the need for better data collection and surveillance to effectively address AMR globally.

Figures and Tables

Table 1. Chow Test Statistics for Time Series Analysis Across Six Regions. Regions with p-values greater than 0.05 do not exhibit significant differences.

	Africa	**Asia**	**Europe**	**Latin America**	**Middle East**	**North America**
Asia	0.043					
Europe	0.036	0.261				
Latin America	0.167	0.042	0.033			
Middle Africa	0.187	0.031	0.042	0.251		
North America	< 0.001	< 0.001	0.038	< 0.001	< 0.001	

Figure 1 illustrates the resistance of *A. baumannii* to Meropenem over time across six regions. The resistance levels are categorized into three groups.

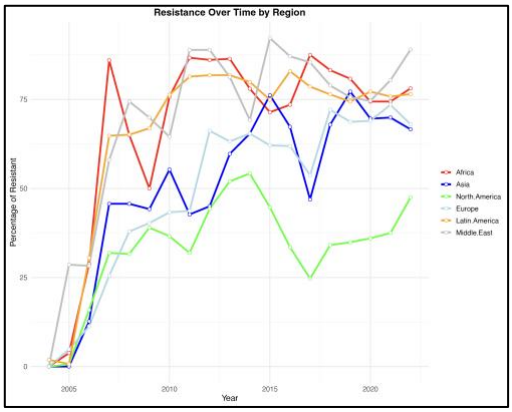


Figure 2 shows the resistance of *E. faecium* to Vancomycin over time across six regions. The resistance levels are categorized into three groups: North America with resistance exceeding 75%.

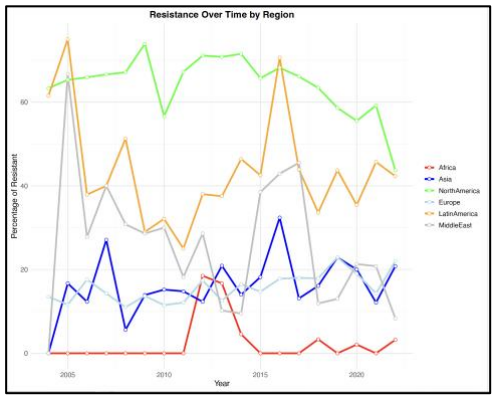
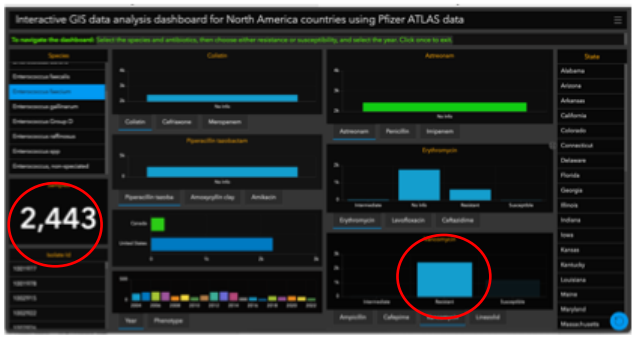


Figure 3 presents a dashboard for selecting antibiotics to treat *E. faecium*. In Figure 3A, it is shown that there are 2,443 isolates of Vancomycin-resistant *E. faecium* (red circle), with 2,403 of these isolates also being susceptible to Linezolid (green circle), as illustrated in Figure 3B.

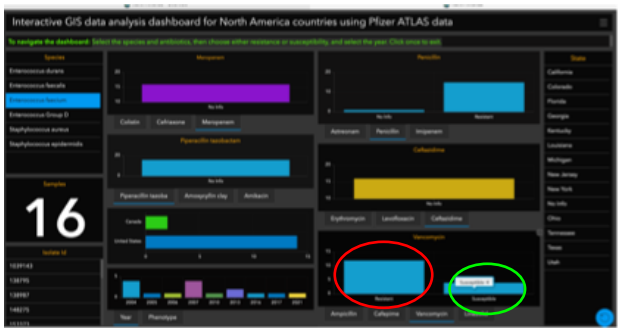
3A



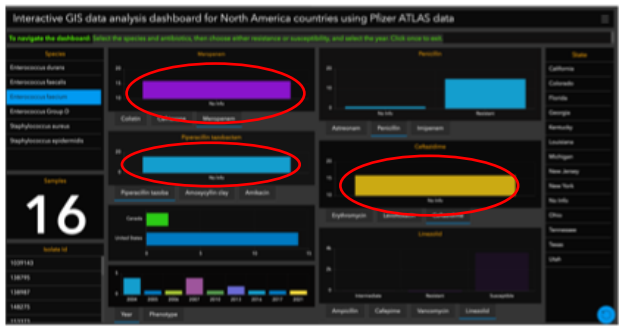
3B

Figure 4 displays a dashboard used for selecting isolates resistant to Linezolid for "omics" studies. In Figure 4A, 12 isolates are marked as resistant to Vancomycin (indicated by red circle), while 4 are susceptible (indicated by green circle). Additionally, Figure 4B shows that all 16 isolates are resistant to Ampicillin, Penicillin, and Levofloxacin (indicated by red circles).

4A



4B



References

1. https://www.who.int/news/item/17-05-2024-who-updates-list-of-drug-resistant-bacteria-most-threatening-to-human-health
2. https://www.who.int/initiatives/glass
3. https://www.who.int/publications/i/item/9789240093461
4. https://amr.vivli.org
5. Opiyo SO, Nalunkuma R, Nanyonga SM et al. Empowering Global AMR Research Community: Interactive GIS dashboards for AMR data analysis and informed decision-making [version 1; peer review: 1 approved with reservations]. Wellcome Open Res 2024, 9:234
6. Li R, D'Arcy CV, Hermsen ED et al. Creative usage of Vivli's Antimicrobial Resistance Surveillance Open Data through the Re-Use Data Challenge [version 1; peer review: not peer reviewed]. Wellcome Open Res 2024, 9:235
7. R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL https://www.R-project.org/.
8. R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL https://www.R-project.org/.