# Spatiotemporal modelling of antimicrobial resistance in key gram-negative pathogens: Deciphering global patterns using indirect predictors


## Abstract

**Objectives**

Antimicrobial resistance (AMR) in gram-negative pathogens is a growing global health threat, particularly in low and middle-income countries (LMICs). This study aims to devise an optimized spatiotemporal model for predicting global AMR patterns in key gram-negative pathogens utilizing indirect predictors, thereby augmenting the capacities of regions lacking AMR data.

**Methods**

Building on our prior research that validated the forecasting potential of socio-economic correlations in national AMR trends, we revisited our AMR prediction models for *Escherichia coli, Klebsiella pneumoniae, and Pseudomonas aeruginosa.* We first, retrained our AMR prediction models for these pathogens, extending the socioeconomic World Bank database, originally spanning 1998-2017, to include data up to 2021, enabling us to predict AMR from 2018-2021. Subsequently, we focused on optimizing our models based on prior methods, which required the integration of AMR prevalence data from 2004-2021 for the specified pathogens. We also incorporated spatial correlations and selected key sociodemographic and health-related covariates to improve the model. Using a stacked ensemble model that included multiple regression models we validated our model through a rigorous five-fold cross-validation process.

**Results**

The preliminary results from our ensemble model provide prediction estimates of antimicrobial resistance within each pathogen-drug combination for countries lacking data from 2004-2021. These results establish a starting point for further model refinement, paving the way for improved model performance in future iterations.

**Impact**

Our research, though centered on key gram-negative pathogens, exhibits strong applicability as the framework can be extended to other bacterial pathogens. With the utilization of global datasets and broadly applicable predictors like sociodemographic variables, our methodology demonstrates versatility across diverse geographical regions and healthcare contexts. In conclusion, our project's importance lies in its potential to prioritize AMR surveillance using a data-driven approach, particularly in low and middle-income countries (LMICs), where more than three billion people live in areas with uncertain AMR estimates.


## Introduction

Antimicrobial resistance (AMR) in gram-negative pathogens is a growing global health threat, particularly in low and middle-income countries (LMICs).[1] World Health Organization (WHO) has declared AMR as one of the top ten global health threats.[2] A particular area of concern is resistance in gram-negative pathogens such as *Escherichia coli, Klebsiella pneumoniae, and Pseudomonas aeruginosa.* These pathogens are linked to severe and frequently lethal infections, with treatment options growing alarmingly scarce due to increasing resistance rates.[3] Traditionally, the surveillance of AMR has been laboratory-based, i.e. using data generated from direct clinical and laboratory reports, constituting a critical component of the public health response to this crisis.[4] However, traditional surveillance systems face several challenges. They rely heavily on healthcare infrastructure and the capacity to perform and report microbiological testing, which can be especially lacking in lower-middle-income countries (LMICs).[4]

Additionally, under-reporting and incomplete coverage are further issues compromising the effectiveness of these systems. [4]

Recognising these challenges, a growing interest has been in harnessing indirect predictors to forecast AMR prevalence. Our previous research has shown that the prevalence of AMR can be correlated with various socioeconomic drivers, suggesting the potential of these indirect predictors as valuable tools for AMR prediction. [5] This approach could be particularly beneficial for countries with limited historical AMR data. By harnessing the power of these surrogate markers, robust predictive models can facilitate data-driven prioritisation of surveillance efforts and guide interventions. [5–8] In this context, our project sets out to build upon the foundation of our previous research, which demonstrated the potential of socio-economic correlations to forecast national AMR trends. [5] Our goal is to develop an optimized spatiotemporal model at the national level for estimating AMR in key gram-negative pathogens using indirect predictors.

## Methods

The methodology of our project consists of the following steps:

**Model Evaluation:**

Following the methodology described in Oldenkamp et al,[5] we retrained our AMR prediction models for *Escherichia coli, Klebsiella pneumoniae, and Pseudomonas aeruginosa*, extending the socioeconomic World Bank database we originally used to cover the period 1998-2021 (instead of the original 1998-2017). By performing our beta-binomial regression on the 30 first principal components derived from this extended database, we enlarged the application domain of the models. We then used these newly trained models to predict AMR prevalence for the year-country combinations in the Vivli AMR datasets. *(ATLAS and GEARS)*

**Model Optimization and Development:**

**AMR Dataset:** We first collated and pre-processed the AMR prevalence data from 2004-2021 for the key gram-negative pathogens, leveraging the ATLAS and GEARS datasets from the Vivli repository. Two antibiotic classes were selected for each of these pathogens, representing approximately 78 countries on average, depending on the bug-drug combination. To incorporate spatial correlations, the dataset was georeferenced.

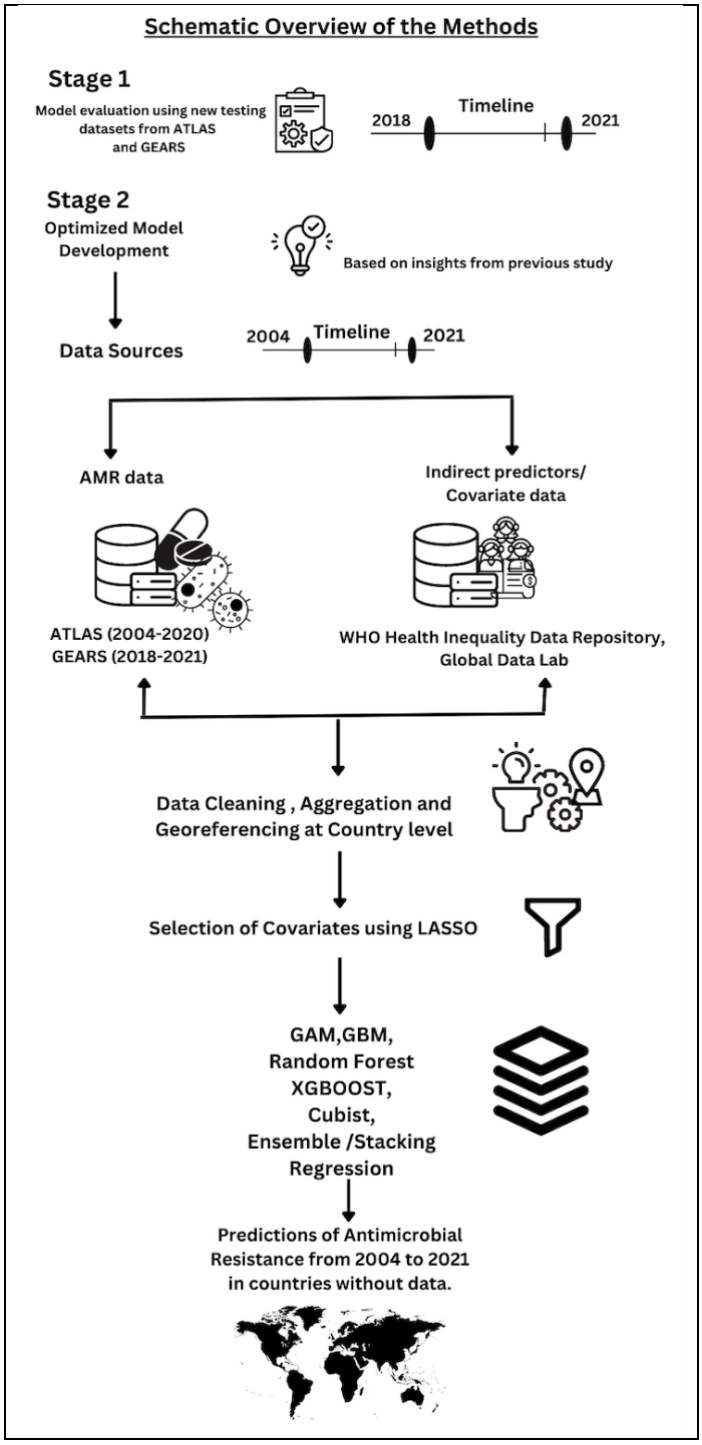
**Selection of covariates:** Based on our previous work and literature review findings, we decided to optimize our existing model by incorporating selected sociodemographic and health-related indirect predictors as covariates. Subsequently, relevant data was procured from WHO Health Inequality Data Repository and the Global Data Lab. [9-10] After the pre-processing phase of preliminary data checks, we applied the least absolute shrinkage and selection operator (LASSO) penalized regression to identify the most influential covariates. Data standardization was also done before applying regression modelling.

**Modelling Framework**: The model-building phase utilized a stacked ensemble model, which included a generalized additive model (GAM), Gradient Boosting Machine (GBM), Random Forest(RF), Cubist and XGBOOST regression models. The prediction models were validated using a rigorous five-fold cross-validation process before stacking. The model validity was assessed by calculating the root mean square error (RMSE) and coefficient of determination (R-squared) for each pathogen distribution model. The modelling framework was then used to provide estimates of antimicrobial resistance in each pathogen–drug combination for countries without AMR data from 2004-2021.

**Ethical Considerations and Open Science Commitment:**

Adhering to ethical research principles, our study utilizes only fully anonymized metadata, thereby ensuring confidentiality.

***Note: This study is presently ongoing, with only initial modelling outcomes available at this point. Further enhancements and adaptations to these models will be pursued, utilizing insights gained from these preliminary findings. We commit to Open Science and will be by publishing our results in an open-access journal at the end of our study.***




# Results

**Model Evaluation – Model 1**

The results showed that model performance drastically decreased when applying to this truly external, new dataset from the Vivli repository. The Mean Squared Prediction Errors (MSPEs) for these models, when based on 5-fold five times repeated cross-validation on the original training dataset only, were 0.94, 1.60 and 0.50 for E. coli, K. pneumoniae and P. aeruginosa, respectively. These increased to 2.02, 2.58 and 1.70, respectively, when model performance was assessed against new data over the period 2018-2021. While this drop in model performance might partly be explained by the fact that we applied these models beyond the period covered by the training datasets, it is also a clear indication that our approach might be improved by advanced techniques and additional indirect predictors.

**Preliminary Results (Optimized Model 2)**

The ensemble model provided prediction estimates of antimicrobial resistance in each pathogen–drug combination for countries without AMR data from 2004-2021.The ensemble model used to predict pathogen distributions exhibited varying degrees of success for different pathogen and antibiotic combinations, as indicated by the root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination ($R^2$) metrics. The predictive model for *Escherichia coli* showed moderate performance against third-generation cephalosporins and fluoroquinolones, with

R-squared values of 0.44 and 0.51, respectively. The predictions for *Klebsiella pneumoniae* showed a stronger performance, particularly against carbapenems, with an R-squared of 0.68, indicating a robust model. However, models *for Pseudomonas aeruginosa* reported less robust performances, especially against aminoglycosides, showing the lowest R-squared of 0.27 among all relationships. These results signify that while the model performs well for certain pathogen-antibiotic combinations, improvements are required for better performance across other different combinations.

**Model Validation Metrics (In-sample Predictions)**

| Pathogen | Antibiotic | RMSE | MAE | $R^2$ |
|---|---|---|---|---|
| *Escherichia coli* | Third-generation cephalosporins | 13.38271 | 9.625442 | 0.4386906 |
| *Escherichia coli* | Fluoroquinolones | 14.29736 | 10.47818 | 0.5141896 |
| *Klebsiella pneumoniae* | Third-generation cephalosporins | 15.8576 | 11.47168 | 0.5943364 |
| *Klebsiella pneumoniae* | Carbapenems | 7.689341 | 4.724921 | 0.6805688 |
| *Pseudomonas aeruginosa* | Third-generation cephalosporins | 13.87353 | 9.640338 | 0.3951398 |
| *Pseudomonas aeruginosa* | Aminoglycosides | 12.12427 | 7.353999 | 0.2672316 |

***Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared ($R^2$)***

Our preliminary results suggest a potential for enhancing our model's performance. Our next steps would include:

- Incorporating more diverse covariates from sources like *Global Health Data Exchange* (GHDX), the *World Bank* and other sources to enrich our dataset and potentially boost performance.
- Fine-tuning the hyperparameters of the covariates before employing stacking techniques. By optimizing these parameters, we hope to ensure the best possible performance of each individual model in the ensemble.
- Lastly, we are exploring the application of spatiotemporal Gaussian regression.

## Added Value of the Study

### Innovation

The innovative aspect of our project stems from our fundamental components: the use of indirect predictors like socio-demographic indicators, health-related indicators and experimentation with different modelling approaches. Traditional surveillance of antimicrobial resistance (AMR) relies heavily on direct microbiological data, which is often limited or absent, particularly in low-and-middle-income countries (LMICs). Our project uses indirect predictors to predict AMR prevalence, offering a creative solution to circumvent data availability issues. By combining various sources of indirect predictors, we're able to provide a multi-faceted view of the determinants of AMR.

### Generalizability

The proposed research exhibits strong generalizability. Although we focus on key gram-negative pathogens *(Escherichia coli, Klebsiella pneumoniae, Pseudomonas aeruginosa),* the methodological framework can be extended to other pathogens. This is possible because our models rely on indirect predictors, which can be applicable to other bacterial pathogens too. The methodology is also versatile in terms of geographical application. By harnessing global datasets and widely applicable predictors such as sociodemographic and environmental variables, we have ensured our models can be applied across different regions.

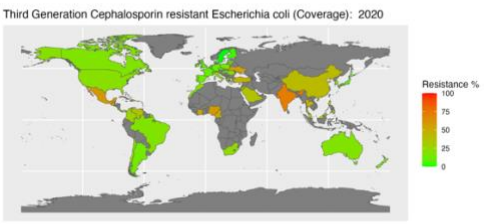
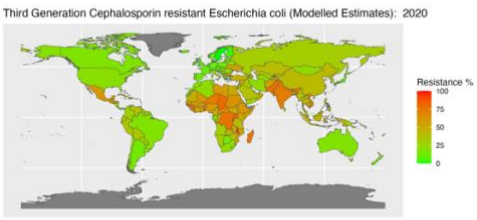
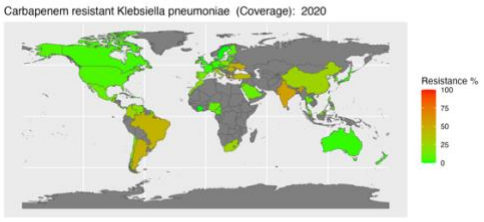
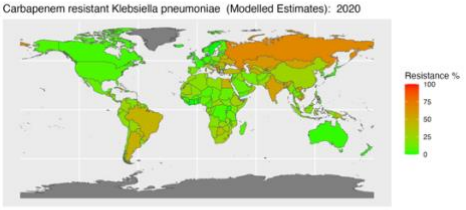
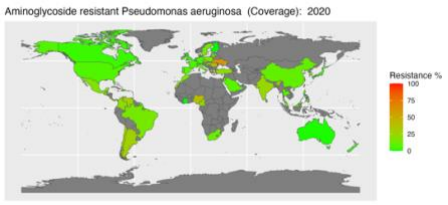
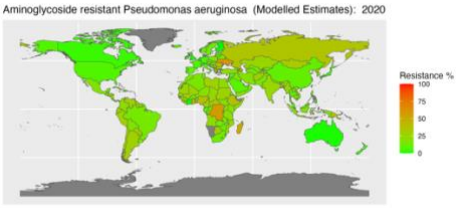
### Impact

The impact of our project, if successfully implemented, would be far-reaching. By providing actionable data, our models can facilitate data-driven decision-making in public health, leading to

more effective surveillance and interventions. A significant portion of the global population lives in areas where AMR estimates are uncertain. Our project offers a practical solution to this problem.

## Conclusion

In conclusion, our project's importance lies in its potential to prioritize AMR surveillance, particularly in LMICs, where more than three billion people live in areas with uncertain AMR estimates. By synergizing indirect predictors and using improved modelling techniques, we enable a more targeted and data-driven approach to prioritize surveillance. As we move forward, we plan to refine our models and explore their application to other priority pathogens, expanding their scope and impact.

| AMR Coverage | AMR Modelled Estimates |
|---|---|
|  |  |
|  |  |
|  |  |

***Note: Model predictions and coverage for other pathogen-drug combinations spanning 2004-2021 can be accessed at the following GitHub repository: https://github.com/spk125/Vivli_DataReuse_Challenge2023.git. In addition, animated time series of the modelled AMR estimates for the same period and the R code are also accessible within the same repository.***