Baltic J. Modern Computing, Vol. 12 (2024), No. 1, 30-49
https://doi.org/10.22364/bjmc.2024.12.1.03

Predicting Antimicrobial Resistance Trends Combining Standard Linear Algebra with Machine Learning Algorithms

Filippo CASTIGLIONE[1], Peteris DAUGULIS[2]✉, Emiliano MANCINI[3,4], Rik OLDENKAMP[5], Constance SCHULTSZ[4], Vija VAGALE[2]

[1]Biotechnology Research Center, Technology Innovation Institute, UAE
[2]Daugavpils University, Latvia
[3]Data Science Institute, Hasselt University, Belgium
[4]Department of Global Health, Amsterdam UMC, University of Amsterdam, Netherlands
[5]Department of Environment & Health, Amsterdam Institute for Life and Environment, Vrije Universiteit Amsterdam, Netherlands

0000-0002-1442-3552 , 0000-0003-3866-514X, 0000-0002-5613-234X, 0000-0002-2245-6987, 0000-0003-2280-7844, 0000-0002-5428-6441

filippo.castiglione@tii.ae, peteris.daugulis@du.lv, emiliano.mancini@uhasselt.be, r.oldenkamp@vu.nl, c.schultsz@aighd.org, vija.vagale@du.lv

Abstract. Antimicrobial resistance prediction is a pivotal ongoing research activity that is currently being explored across various levels. In this context, we present the application of two prediction methods that model the antimicrobial resistance of Neisseria gonorrhoeae on the national level as an outcome of socio-economic processes. The methods use two different implementations of the principal component analysis combined with classification algorithms. Using these two methods, we generated forecasts concerning antimicrobial resistance of Neisseria gonorrhoeae, using publicly available databases encompassing over 200 countries from 1998 to 2021. Both approaches exhibit similar mean absolute averages and correlations when comparing available measurements with predictions. Steps of statistical analysis and applications are discussed, including population-weighted central tendencies, geographical correlations, time trends and error reduction possibilities.

Keywords: PCA, principal component regression, antimicrobial resistance, AMR prevalence prediction, Neisseria gonorrhoea, surveillance.

1. Introduction

Antimicrobial resistance (AMR) is an important health issue increasing healthcare costs, operational and research burden for healthcare services, and increasing mortality worldwide. The increasing speed at which bacteria develop resistance to new

✉Corresponding author

antimicrobials combined with the reduced economic return in the development of new antimicrobials leads to a serious threat to global healthcare. AMR is studied and modelled in at least three levels – micro (within-host), meso (between-host) and macro (society) levels. These studies aim to design new antimicrobials, develop treatment guidelines and administration strategies to optimize the introduction of new antimicrobials with the aim of postponing the emergence of resistance as much as possible. This is particularly important considering the slow pace at which new antimicrobials are developed. In this article, we study AMR modelling at a macroscopic level. We follow the justified assumption that AMR is correlated with socioeconomic processes (Blommaert, et al., 2014; Alsan et al., 2015; Collignon et al., 2015; Alvarez-Uria et al., 2016; Collignon et al., 2018) and its quantitative indicators can be predicted using quantitative socio-economic indicators.

Gonorrhoea is the second most prevalent sexually transmissible infection. It is caused by the bacterium Neisseria gonorrhoeae (NG) (Yang and Yan, 2020). The World Health Organization has estimated that approximately 787 million new gonorrhoea cases emerge annually (Unemo et al., 2019). There are no effective gonococcal vaccines, therefore antimicrobial (antibiotic-based) therapy remains the main tool for treating and preventing gonorrhoea infections (Tapsall, 2002; Ison et al., 2013; Cyr et al., 2020;).

The first antibiotic for treating infections caused by gonococcus was introduced in the 1930s. AMR in NG is advancing, leading to significant implications for reproductive, maternal and newborn health (Unemo and Nicholas, 2012; Lewis, 2014; Lewis, 2019; WEB, b).

The lack of quantity and quality of AMR data is one of the major challenges for both national and international AMR surveillance programs, particularly in low and middle-income countries. This challenge becomes even more pronounced for community-acquired pathogens such as NG. The lack of data makes it impossible to effectively estimate the AMR prevalence of drug-resistant gonorrhoea strains (Unemo et al., 2012) in most countries with standard statistical methods. This issue is aggravated by the global increase in the AMR of NG. The possibility of filling the gaps in AMR surveillance data with data-driven predictions would allow the healthcare community to improve planning for the usage of existing drugs, optimize the introduction of new drugs and associated surveillance programs, and define guidelines restricting inappropriate antibiotic usage in areas where the highest possible risk has been estimated.

A range of general-purpose methods used to predict missing information, such as approximation, interpolation, extrapolation and others, have been developed (Meijering, 2002; Mittal, 2016). Machine learning approaches such as dimensionality reductions and representation learning are widely used, see examples in (Bengio et al., 2013; Bzdok et al., 2018). The literature reviews (Sakagianni et al., 2023; Faiza et al., 2023) list main machine learning application areas and methods for AMR. Currently most machine learning applications are designed to serve as clinical decision support tools. They deal with training sets addressing individual patient information (demographics, previous infections and antibiotics treatments etc.) and predict AMR in individual cases. Logistic regression (Hosmer, D.W. and Lemeshow, S., 2000) along with decision tree (Shalev-Shwartz, S., Ben-David, S., 2014) and random forest (Breiman, L., 2001) are mentioned as the most widely used prediction methods (Tang et al., 2022). We note that except of (Oldenkamp et al., 2021) there appears have been no other publications studying the AMR at the society level.

An effective and widely used ingredient of data modelling and analysis is the Principal Component Analysis (PCA) (Hotelling, 1933; Hotelling, 1936). It is used to associate to a discrete set of data points a shifted linear subspace which shows the most significant variables and their linear combinations. PCA is used for dimensionality reduction, linearization of data, filtering out noise and finding the most important linear combinations of data variables (Meglen, 1991; Gorban et al., 2008). It is considered a major unsupervised learning technique in machine learning competing in the data approximation area with linear regression which is a supervised learning technique. PCA is also used in statistical tools developed by statisticians, showcasing a broad range of applications across various fields of science and technology. It serves as a crucial step in acquiring predictions through machine learning methodologies.

To fill current gaps in the global prevalence map of NG, we used a previously developed statistical model that can predict AMR prevalence based on socio-economic World Bank profiles (Oldenkamp et al., 2021). Oldenkamp and colleagues leveraged robust statistical relationships between countries' socio-economic status and measured AMR in clinical isolates for nine (predominantly) healthcare-associated WHO priority pathogens, to predict AMR prevalence in countries for which AMR data were not available. PCA is used in this method to reduce the dimensionality of the World Bank data set. They did, however, not assess the community-acquired infections and their pathogens. Considering the substantial differences for both infection dynamics and AMR surveillance between strategies for healthcare-associated and community-acquired infections, it is not clear whether such models would be useful to predict AMR prevalence for a community-acquired pathogen such as NG. In an attempt to use relationships between World Bank indicators and AMR in a prediction method and reduce prediction errors we piloted an innovative prediction model (Daugulis et al., 2022). In this model, PCA is used to reduce the dimensionality of the data set encompassing both socio-economic and AMR data – data points have both socio-economic and AMR coordinates. Both methods use normalisation of the socio-economic data to make it dimensionless. We show that the models have reasonable accuracy for NG AMR prediction and provide a wide view of the prevalence of AMR in NG.

The paper is organised as follows. In Section 2 we describe the data and the computational methods. Results, their analysis, possible application and discussion are given in Section 3. Finally, concluding remarks and future work are drawn in Section 4.

2. Methods

2.1. Data Types

Two main types of data were collected. The first type of data consists of socio-economic data collected yearly covering the period 1998-2021. By socio-economic data, we mean numerical socio-economic and demographic indicators profiling individual countries at the national level. The second type of data consists of yearly AMR prevalence of NG for different antimicrobials which can be deduced from antibiograms aggregated at the national level. It is a quantitative measure of the resistant bacteria strains that can be obtained from public national and international surveillance programmes. Below we describe the used data sets.

Socio-economic data – the World Bank database. Socio-economic data was retrieved from the World Bank database (WEB, c). The World Bank is a unique global framework representing more than 200 countries and geographic areas over the period 1998-2021. In total, the database contains more than 14000 indicators covering a wide range of aspects such as population, environment, government finance, national accounts, social policy statistics, development assistance, balance of payment, exchange rates, prices, financial statistics and trade. Our basic assumption is that World Bank panel data represent meaningful quantitative values expressing socio-economic features and processes that can be predictive of AMR. Although some indicators that are known to strongly correlate with AMR, such as antibiotic consumption, are not directly included in the World Bank data set, we assume that they can be expressed as linear functions of a subset of existing socio-economic indicators. While most indicators were available yearly, a subset of them was only available quarterly. In the latter case, we used the sum of the quarterly values as the yearly value. The World Bank observations and indicators with >95% missing data were removed. Finally, the columns (indicators) which are almost constant, were removed. For this purpose, Kvalseth's V_2 is used instead of the coefficient of variation (Kvalseth, 2017). V_2 is computed for each column, and columns with a value less than 0.2 are crossed out. After these steps, the fraction of undefined entries in the World Bank data matrix is reduced to about 70%. The World Bank data was collected using R software on September 8, 2021. Initially, the data matrix contained 14303 indicators and 5064 observations. After cleaning we were left with 7248 indicators and 4542 observations. Since we are interested in AMR predictions at specified time intervals rather than analyzing the data at specific time moments, the focusing of standard panel models is insufficient. We expand the cleaned World Bank panel data matrix by adding time indicators.

Antimicrobial resistance data – WHO-GASP and WHO-GLASS. Our target variable (to be predicted) is the AMR prevalence in NG at the national level. The data was obtained from WHO's AMR surveillance programs "Gonococcal Antimicrobial Surveillance Program" (GASP) (Unemo et al., 2019) and "Global Antimicrobial Resistance and Use Surveillance System" (GLASS) (WEB, e). These databases contain aggregated antibiograms recording numbers of all and resistant isolates of infections in the participating countries. The WHO-GASP data downloaded (9807864 isolates in total) represent AMR prevalence in NG against 5 antibiotics/antimicrobials - Azithromycin, Cephalosporins, Cefixime, Ceftriaxone, Ciprofloxacin, covering a total of 90 countries over the period 2009-2018. An earlier WHO-GASP data version contains AMR prevalence in NG against Azithromycin, Ceftriaxone, Ciprofloxacin and Cefepime, together with data for Cefepime/Ceftriaxone over the period 2009-2016. Our predictions were obtained using the WHO-GASP data. Since 2016, gonococcal AMR surveillance has also been covered as part of the WHO's larger GLASS program (WEB, a). The most recent report includes national AMR prevalence data from 2019. We downloaded the WHO-GLASS data (182275 isolates in total) and used it for independent checking of predictions obtained using the WHO-GASP data.

2.2. The AMR Value and Its Derivation

The AMR value is here defined as the fraction of the recorded resistant infection cases (isolates) over all recorded infection cases. Table 1 shows the notations explaining our computations.

Table 1. Notations describing population groups

Category of people	Notation
The population in a fixed country X	N_P
The number of infected people in X	N_I
The number of tested infected people in X	N_T
The number of resistant tested infected people in X with respect to a fixed antibiotic A	N_R
The total number of resistant infected people in X with respect to A	N_{RI}

Clearly, $N_p>N_I>N_T>N_R$ and $N_{RI}>N_R$. We are ultimately interested in estimating N_{RI} (the number of infected human carriers of resistant bacteria) for each country since it determines the demand for antibiotics and healthcare services. For this purpose, we consider fractions which are shown in Table 2.

Table 2. Fractions used in estimating AMR

Description	Notation	Calculation
The fraction of resistant tested cases (AMR value)	a	N_R/N_T
The fraction of tested infected cases	b	N_T/N_I
The fraction of infected cases	c	N_I/N_P

Since most surveillance programmes record only the numbers of resistant and all isolates, the fraction denoted here by a is the only numerical indicator of AMR available for researchers. We are aware that standards defining infection cases as being resistant and fractions of tested people are different in different countries.

Assuming that the isolates tested are a representative sample of the total infected population we put $N_R/N_T = N_{RI}/N_I=a$ (the fraction of resistant tested infection cases is close to the fraction of all resistant infection cases). Therefore, the fraction a represents the fraction of resistant infected people in each country. Additionally, it follows that $N_{RI} = aN_I = acN_P$. We note that we have no information about the fraction c. This consideration significantly limits the possibility of estimating N_{RI}. We are left with analyzing and predicting the fractions.

2.3. Prediction Methods

The beta-binomial principal component regression (BBPCR) method. A framework for AMR prediction and surveillance based on a modified beta-binomial principal component regression was designed and implemented (Oldenkamp et al., 2021). It takes

as input a training set based on paired socio-economic data and available AMR measurements, and it outputs AMR predictions and the corresponding confidence intervals. We apply this method in the case of NG. Our training set combined socio-economic data from the World Bank with AMR measurements on NG strains from WHO-GASP.

To reduce the dimensionality of the World Bank data, PCA was performed using the *nipals* function of the R package *pcaMethods* (Stacklies et al., 2007). The number of principal components equal to 30 was chosen to explain at least 90% of the data variance. Thus 30 nontrivial orthonormal linear combinations of 7248 World Bank indicators were chosen to express the essential variability of the World Bank data. Each observation, a vector of length 7248, was substituted by the list of projections of this vector onto these linear combinations (PC scores), a vector of length 30.

In the principal component (PC) regression model, the rows of the training set matrix X correspond to country-year-antimicrobial triples, while the columns are the PC *scores* of the socioeconomic data points. In addition to these scores, the training set also included variables related to time (normalized year and squared year). The inclusion of time variables was motivated by the dependence of socio-economic indicators on time, and by the goal to predict AMR changes over time. A single model combining all antibiotics is used, all AMR measurements are arranged into a single vector.

The vector of AMR values is transformed by the logit link function. The regression equation essentially is $y = X\beta + \alpha + \varepsilon$, wahere y is the vector of transformed AMR values.

A 5-times repeated 5-fold country-groupwise cross-validation was performed as an optimal error rate estimate procedure. Model quality was measured by the mean coefficient of determination over the 5 folds. The model parameters are computed using *vgml* function (Yee, 2010).

After training, model predictions were transformed back to AMR values by applying the inverse logit function, using the *predictvglm* function. Prediction quality for this model was estimated using the mean absolute prediction error and the mean coefficient of determination.

Minimal PCA-distance method (MPCD). The BBPCR method applies PCA analysis in the space containing only independent (e.g. socio-economic) variables. Such a feature may obscure relationships between the independent variables and AMR (Artigue, Smith, 2019). In an attempt to apply PCA innovatively, we piloted an alternative method, the minimal PCA-distance method (Daugulis et al., 2022). Similar to the BBPCR, the MPCD method is based on dimensionality reduction through PCA. But in this method, PCA is applied to reduce the dimensionality of data points in the n+1 dimensional space including n socio-economic indicators plus time, and one extra dimension corresponding to the AMR values. Regression is not used in this method, and PCA is used in combination with a metric (Euclidean or other) in the ambient real linear space of variables. The idea behind this method is to consider the distance between two hyperplanes - the hyperplane spanned by the first principal components and the hyperplane of candidate points. The point on the candidate hyperplane in the minimal distance to the principal component hyperplane produces the AMR prediction. The steps of the method can be interpreted in terms of machine learning - the addition of the extra AMR dimension and performing PCA in this space can be interpreted as a feature of supervised learning.

For each antibiotic, a separate PCA-hyperplane is constructed using the AMR measurements for that antibiotic.

In this method, the structure of the World Bank socio-economic matrix is used to impute its missing values. We impute the missing values via interpolation by cubic splines for data grouped by countries and indicators. Missing values for a specific indicator and specific country are imputed as values of the cubic spline determined by existing values for the same indicator-country pair. This approach is justified by the assumption that the values of a given socio-economic indicator vary smoothly over time for a given country. In comparison, the BBPCR method uses *nipals* function which performs PCA without taking into account the training set structure.

Detection of outliers and influential observations. In both methods *Cook's distance* (Cook, 1979), with modifications implied by the methods, is used to identify and remove from training sets the outlying data or data with the largest influence on the predictions. In the BBPCR method, the total Cook's distance of each country's data is computed iteratively and the country with the maximal Cook's distance value is removed. In the MPCD method, outliers are identified using our variation of Cook's distance idea based on (Kim, 2017; Zimek and Schubert, 2017;). The measure of the influence of a given point p is the difference between sums of projection squares of all data points onto PC planes with and without p (Daugulis et al., 2022). To save computation time, outliers are removed in one step - the top 5% of points having high influence measures are removed before finalizing the PC hyperplane for prediction in the MPCD method.

2.4. Analysis and Application Methods

Population-weighted central tendencies. Central tendencies (means, medians, modes) of AMR value predictions for the world summarize information over the whole data set and are useful in drawing justified conclusions. For instance, AMR values that are higher than the average or median value may imply an urgent need to introduce a new antibiotic.

Because the numbers of human carriers of resistant bacteria are important for estimating demand for antibiotics and health-care policies, it would be too naïve to consider unweighted averaging. Instead, we should consider population-weighted averaging. As in section 1, we use the following notation: N_I is the number of infected people; N_{RI} is the number of resistant infected people; N_T is the number of tested people; N_R is the number of resistant people; $a=N_R/N_T$ $b=N_T/N_I$, $c=N_I/N_P$. We see that the number of carriers of resistant strains is proportional to a and N_P (that is, $N_{RI} = caN_P$). Although the coefficient c is unknown, it is possible to compute and interpret population-weighted averages of the quantity numbers of human carriers of resistant bacteria (N_{RI}). Suppose we have m countries with total populations N_1, N_2,..., N_m and AMR values a_1, a_2,..., a_m. Then the total number of resistant human carriers is $c\Sigma a_k N_k$ and the fraction of resistant (infected) human carriers is $c\Sigma a_k N_k/\Sigma N_{Ik}=c\Sigma a_k N_k/c\Sigma N_k=\Sigma a_k N_k/\Sigma N_k$, which is the population-weighted average of a.

Geographical correlation analysis. The dynamics of community-acquired infections must have a part implied by human mobility – travel, migration and contacts. One can conjecture that the AMR of community-acquired infections for neighbouring countries are related. We can assume that the AMR of a given community-acquired

pathogen in a country should be close to the (population-weighted) mean AMR of its neighbours or, more generally, the mean with respect to an appropriate geography-based distance matrix. We perform the Moran I analysis, see (Oldenkamp et al., 2021; Getis, 2010). The AMR value a for each country is compared with the average (lagged) AMR values a_{lag} of its neighbouring countries. For each country-year-antimicrobial combination, we have the point (a,a_{lag}) and take the union of such points over the world defining a set S. We consider the linear regression line L (the Moran I line) $y=kx+b$ for S. If (a,a_{lag}) is below L then it is below $(a,ka+b)$ (a is higher to a_{lag} than expected) and to the right of (a_0,a_{lag}), where a_0 is such that $a_{lag}=ka+b$ (the expected lagged value for a is lower than a_{lag}). Outlying countries having their points (a,a_{lag}) below L should be considered as countries at risk – *Moran outliers.* In these countries, the AMR is progressing faster than what could be explained by contact and travel dynamics.

The rate of change of the AMR. The rate of AMR value change (overloading terminology we call it *the AMR trend*) is another numerical indicator that can be used to assess the AMR and determine countries where the introduction of a new antibiotic could be prioritized. Countries with a relatively high AMR trend can be considered to be at risk and require attention. Countries with relatively low AMR trends should be studied further.

If AMR values for a country in years Y and $Y-d$ are a and a', respectively, then the AMR trend is estimated as $r=(a-a')/d$. Using the above notation, $(rN_P)c$ is equal to the change in the number of resistant human carriers of the pathogen. Again, the absolute number of the rate of change of human numbers is not possible to estimate without knowing the coefficient c (the fraction of infected people), we can consider ranking lists and population-weighted averages. The AMR trend is measured in %/year.

Identifying capabilities to improve predictions by adding more AMR data. In addition to AMR predictions, useful information can be obtained using the UI-generating feature of the beta-binomial principal component regression method. We remind the reader that prediction errors (UI) in the beta-binomial principal component regression method are computed using the beta-binomial distribution and that they depend on the fraction of resistant isolates and the total number of isolates.

We are interested in obtaining predictions with as small local and global UI as possible. Therefore, we can ask the following question posed and implemented in (Oldenkamp et al., 2021). In which countries should we increase the number of tested infected people so that the total prediction error according to the beta-binomial principal component regression method in a suitable sense is as small as possible? This can be done by running a loop over countries with an artificially increased number of tested cases (isolates) for the country under iteration, computing the whole prediction process for that iteration. We increase the number of tested people by 30 since it is the minimal number sufficient for the inclusion of AMR data.

Countries corresponding to higher total prediction reduction are considered to have a high prediction precision impact. Error reduction can be considered both for the same country (direct effect) or for other countries (indirect effect). Requesting additional data from such countries can be considered an optimal step for increasing the precision of AMR prediction which can be further used for various purposes.

3. Results and Discussion

3.1. National Predictions of AMR

Using the beta-binomial principal component regression method we computed AMR predictions using the WHO-GASP version for the period 2009-2018 available in 2022, for 202 out of a total of 235 world countries and areas during the period 1998-2021, as available from the World Bank data. Two countries (Australia and Canada) were removed from the training set because Cook's analysis identified them as outliers. Using the minimal PCA-distance method, AMR predictions were computed using the WHO-GASP data available in 2021 for the years 1998-2021 for 204 countries. The number of countries having AMR measurements in GASP ranges from 65/235=27% (cefixime) to 87/235=37% (ceftriaxone). By applying our methods, we improved the coverage to 202/235=86% of all the world countries and areas specified by the United Nations (WEB, d). We note that AMR predictions are made for all Low and Low-Middle Income Countries (LMIC), where AMR estimates are currently lacking. As previously mentioned, the criteria for bacterial resistance and the proportions of tested individuals can vary across different countries.

After obtaining AMR predictions and their 95% confidence intervals (for the BBPCR method), we compared measured and predicted AMR values for those years and countries where measurements exist, using the Mean Absolute Error (MAE) and correlations, see Table 3. There are missing entries in the table since computations for the two methods used different GASP versions which had different sets of antimicrobials.

The total MAE for all antibiotics combined for the BBPCR and MPCD methods is 8.8% and 6.9%, respectively. The correlation between the total measurement and prediction vectors for the BBPCR and MPCR methods is 0.862 and 0.877, respectively. We note that in this sense the difference between the two methods is insignificant and both methods provide comparable levels of accuracy. The cross-validated q^2 (predictive correlation coefficient) for the BBPCR method after 5-fold cross-validation is 0.725.

Table 3. MAE and correlations comparing measured and predicted NG AMR values for 202 and 204 countries

Antimicrobial	MAE (%, BBPCR)	MAE (%, MPCD)	Corr. (BBPCR)	Corr. (MPCD)
Azithromycin	8.94	5.36	0.04	0.44
Ceftriaxone	3.41	3.95	0.43	0.18
Ciprofloxacin	19.2	15.73	0.55	0.67
Cefixime	4.92	NA	0.08	NA
Cefepime	NA	2.91	NA	0.64

Correlations between the BBPCR predictions for different antibiotics are close to 1 (ranging from 0.78 to 0.99). This is consistent with the fact that in the BBPCR method PC scores depend only on socio-economic indicators and time and do not depend on AMR values. For most antibiotic pairs they are higher than correlations between measurement vectors.

Correlations between the MPCD method predictions for various antibiotics vary from -0.11 to 0.51. Correlations between prediction vectors are lower than correlations between measurement vectors. In contrast with the BBPCR method, no clear features relating measurement and prediction correlations for antibiotic pairs are visible which is against expectations regarding a prediction method. Thus we have found only one potential benefit of the MPCD method – the closeness of its MAE and correlations comparing AMR measurements and predictions to that of the BBPCR method.

3.2. Prediction Analysis from the Antibiotic Introduction Point of View

Basic statistical analysis. Before doing statistics we can visualize the computed AMR predictions using coloured maps. Figure 1 shows the world map where each country is coloured according to the predicted AMR to Azithromycin using the beta-binomial principal component regression method.

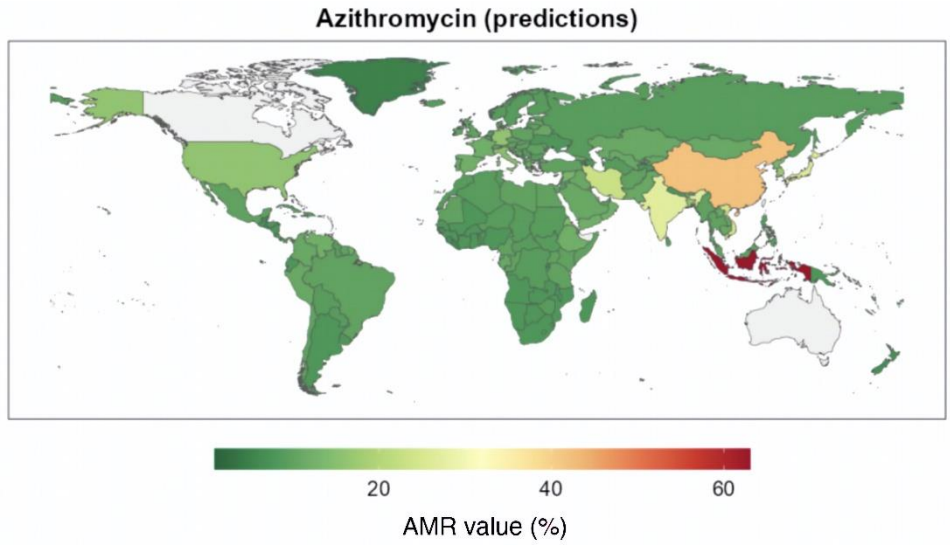

Figure 1. Azithromycin NG AMR prediction map (%, BBPCR method, GASP 2009-2018, predictions for the last year with the WB data for each country, 2020-2021.

The basic prediction analysis involves ranking countries or regions according to their predicted AMR value. Countries with the largest predicted AMR value are priority candidates for the introduction of new antimicrobials. The reason is that new antimicrobials in these countries would decrease the number of human carriers of pathogens and, therefore, slow down the spread of existing resistant pathogens. Additionally, if a population is highly resistant to existing antibiotics, a new antibiotic would result in higher therapeutic success. On the contrary, countries with low predicted

AMR would not benefit from the introduction of new antibiotics, and bacteria strains resistant to the new antibiotic would start to develop earlier. Below we give two lists with the top ten countries sorted by predicted AMR values. For all antimicrobials, the countries with maximal AMR values are the same because the regression is made by combining the AMR measurements in a single vector (a single model).

We show countries with maximal AMR value predictions for Azithromycin and Ceftriaxone in Table 4. For two countries (Iran and Bangladesh) AMR measurements are not available, our predictions give its first estimates.

Table 4. Countries with maximal predicted NG AMR values (%, BBPCR method, GASP, 2009-2018, predictions for the last year with the WB data for each country, 2020)

	Azithromycin			Ceftriaxone		
Rank	Country	AMR (%)	95% UI	Country	AMR (%)	95% UI
1.	Indonesia	63	44-79	Indonesia	26	13-43
2.	China	42	29-56	China	13	8-20
3.	India	27	16-41	India	7	4-12
4.	Japan	26	19-35	Japan	7	4-10
5.	Vietnam	24	16-35	Vietnam	6	4-10
6.	Iran	23	1-90	Iran	5	0-63
7.	S.Korea	17	12-23	S.Korea	4	3-6
8.	Bangladesh	16	11-23	Bangladesh	4	2-6
9.	Germany	16	10-24	Germany	4	2-6
10.	USA	16	9-27	USA	4	2-7

In Table 5 we compare the BBPCR-predicted values for these countries with the GASP data for the last available year. We note that there are countries with the error larger than the MAE. Many of the measurements are based on small numbers of isolates.

Table 5. Comparison of predicted and GASP NG AMR data for countries with maximal predicted AMR values (%, BBPCR method, GASP, 2009-2018, predictions for the last year with the WB data for each country, latest GASP data).

	Azithromycin			Ceftriaxone		
Rank	Country	AMR (%)	GASP (%)	Country	AMR (%)	GASP (%)
1.	Indonesia	63	20	Indonesia	26	80
2.	China	42	14.5	China	13	11.7
3.	India	27	9.4	India	7	3.1
4.	Japan	26	26	Japan	7	19
5.	Vietnam	24	0.5	Vietnam	6	2.3
6.	Iran	23	NA	Iran	5	NA
7.	S.Korea	17	0	S.Korea	4	8.3
8.	Bangladesh	16	NA	Bangladesh	4	NA
9.	Germany	16	4	Germany	4	0.1
10.	USA	16	0.5	USA	4	0.05

Table 6 shows countries with maximal AMR values obtained by the minimal PCA-distance method (Daugulis et al., 2022). There are no WHO-GASP measurements for Bolivia, Cameroon, Congo DR, Guinea, Guyana, Iran, Kuwait, Kosovo, Lesotho, Mozambique, Papua New Guinea, Sao Tome and Principe, Tuvalu, Zambia. It can be noted that maximal AMR values are close to 100% and differ markedly from the measurements. Although the mean square error of both methods (BBPCR method and MPCD method) are close, the MPCD method seems to distort predictions for countries having outlying data with respect to the PCA hyperplane. This may follow from the fact that regression is not used in the MPCD method. The fact that there were predicted AMR values that were close to 100%, together with the properties of correlations mentioned above, led us to conclude that the MPCD method generates too many outlying predictions for a working prediction method. In applications, we focused on the BBPCR method.

Table 6. Countries with maximal predicted Azithromycin and Ceftriaxone NG AMR values (%, MPCD method, GASP 2009-2018, predictions for the last year with the WB data for each country)

Rank	Country	Azithromycin AMR (%)	GASP (%)	Country	Ceftriaxone AMR (%)	GASP (%)
1.	Guinea	100	NA	Indonesia	100	80
2.	Lesotho	100	NA	Iran	100	NA
3.	Madagascar	100	0	Kuwait	100	NA
4.	Uganda	100	0	Papua New Guinea	100	NA
5.	Zambia	100	NA	Vietnam	100	6
6.	Peru	99	9	Bolivia	99	NA
7.	Kosovo	97	NA	Mozambique	93	NA
8.	Congo,D.R.	91	NA	Tuvalu	92	NA
9.	STP	91	NA	Kosovo	92	NA
10.	Cameroon	82	NA	Guyana	90	NA

Population-weighted central tendencies. Table 7 shows the global population-weighted central tendencies for *Neisseria gonorrhoea* AMR predictions using GASP data for various antimicrobials (GASP 2009-2018, predictions for the last year with the WB data for each country). Mean AMR values show the global progress of AMR and the effectiveness of specific antimicrobials. Coefficients of variation for all antibiotics are below 1, this may mean that the AMR value is low-variance according to our model and additional indicators would increase the variance, i.e. the sensitivity of the prediction method. We note that Ciprofloxacin has the highest population-weighted mean. This may imply a need to develop and introduce new antimicrobials substituting Ciprofloxacin.

Apart from taking population-weighted averages over the world, we can consider various ways to partition the world into subsets, i.e. consider income groups or continents. Table 8 shows the population-weighted central tendencies for LMIC. We note that the difference between the world and LMIC means is not significant.

Table 7. Population-weighted central tendencies for NG AMR predictions over all countries (%, BBPCR method, GASP 2009-2018, predictions for the last year with the WB data for each country)

Antimicrobial	Weighted mean (%)	Weighted st.dev. (%)	Weighted median (%)	Weighted mode (%)	Coefficient of variation
Azithromycin	22.9	14.8	15.5	42	0.65
Cefixime	15.1	11.2	9.5	17	0.74
Ceftriaxone	6.5	5.7	6.5	7	0.88
Ciprofloxacin	75.7	12.9	74.5	92	0.17

Table 8. Population-weighted central tendencies for NG AMR predictions over LMIC (%, BBPCR method, GASP 2009-2018, predictions for the last year with the WB data for each country)

Antimicrobial	Weighted mean (%)	Weighted st.dev. (%)	Weighted median (%)	Weighted mode (%)	Coefficient of variation
Azithromycin	22.3	14.7	25.5	27	0.66
Cefixime	14.7	11.6	16	17	0.79
Ceftriaxone	6.4	6.3	6.5	7	0.98
Ciprofloxacin	75.8	12.9	84	85	0.17

Geographical correlation analysis. Figure 2 shows the S-set in the case of Azithromycin. Among Moran outliers with respect to the model estimates, there are Lesotho, Lithuania, Moldova, and Switzerland.

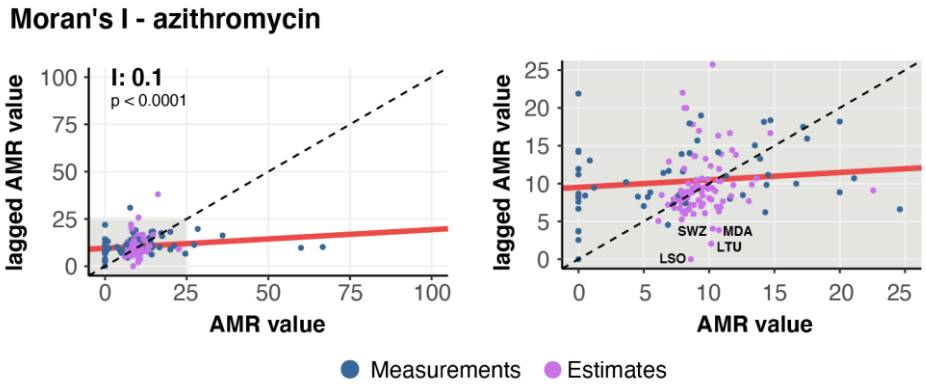

Figure 2. The Moran set for Azithromycin in the full range of AMR values from 0% to 100% (left panel). A closer inspection of AMR values and lagged values in the range 0%-25% (right panel). In this range of AMR values, we identified several outlier countries (LSO = Lesotho; LTU = Lithuania; MDA = Moldova; SWZ = Switzerland).

Temporal tendency observation. Our national AMR predictions are computed as time series for several consecutive years, i.e. for the years having the World Bank data. It allows us to estimate the rate of change of the predicted AMR value.

To demonstrate AMR value time series in one picture we consider groups of countries which have the same income level – income groups. Figure 3 shows these time graphs for azithromycin and ceftriaxone. The time graphs are obtained using GASP data, dots correspond to GLASS data. We can notice that population-weighted income averages appear to be converging, the maximal AMR value difference changes from 10% in 2009 to 3% in 2018. This may be a manifestation of worldwide globalization processes.

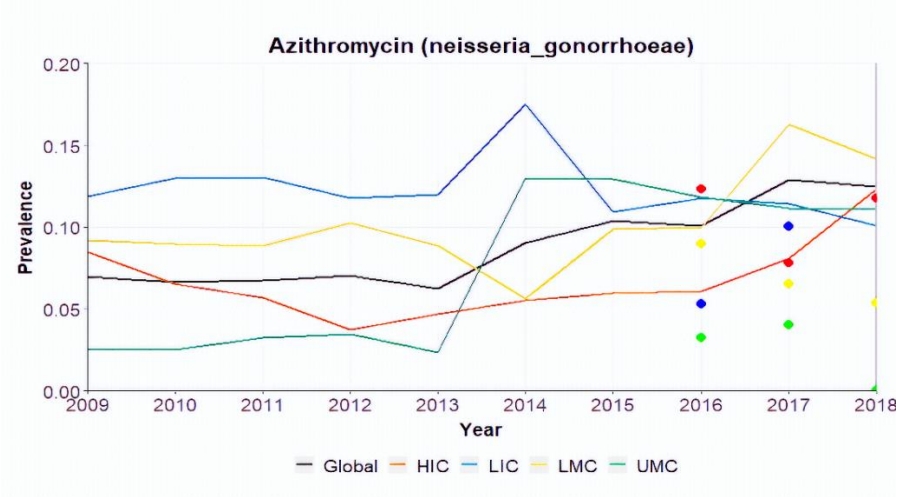

Figure 3. Population-weighted country-group mean prediction time graphs (Azithromycin, BBPCR method, 2009-2018), compared with GLASS (dots).

AMR value time series allows us to estimate the rate of change of the predicted AMR value. We compute the AMR trend (the rate of change) for each country and for population-weighted central tendencies such as mean and variance, taking $d=5$ and Y equal to the last year of prediction. In Figure 4 we visualize the AMR trend for azithromycin using the coloured map.

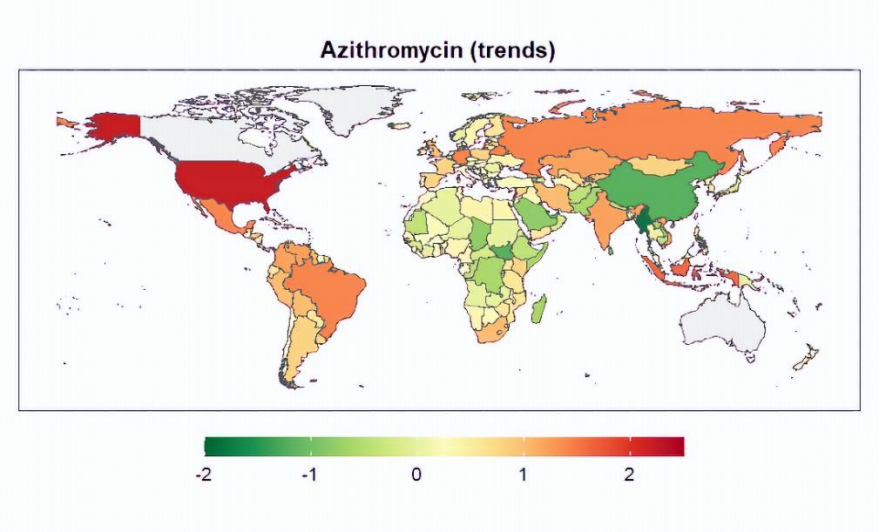

Figure 4. Azithromycin trends map (%/year, BBPCR method, GASP 2009-2018, starting from predictions for the last year with the WB data for each country during the period).

The ranking of countries can be repeated using the AMR trend values. Countries with the largest AMR trend value are countries where the AMR is progressing and a response will be required soon. Table 9 shows countries with top trend values.

Table 9. Countries with highest Azithromycin NG AMR trends (%/year, BBPCR method, GASP 2009-2018, trends starting from predictions for the last year with the WB data for each country)

World	%/year	95% UI, %/year	LMIC	%/year	95% UI, %/year
United States	2.2	0.2 - 4.8	Indonesia	1.6	0 -8.6
Indonesia	1.6	0 - 8.6	Micronesia, India, Vietnam	1.2	0.2 - 2.1
Belarus, Brazil, Germany, St.Lucia, Mexico, Russia	1.4	0.2 - 2.8	Bolivia, Iran, Kyrgyzstan, Samoa	1	0 - 17

Population-weighted central tendencies of the AMR trends for various antimicrobials give us statistical information about the rate of change of human carriers of resistant bacteria strains. Table 10 shows global population-weighted central tendencies for trends. We notice again that Ciprofloxacin has the highest populated-weighted mean.

Table 10. Global population-weighted central tendencies for NG AMR prediction trends (%, BBPCR method, GASP 2009-2018, predictions for the last year with the WB data for each country)

Antimicrobial	Weighted mean (%/year)	Weighted stand.dev. (%/year)	Weighted mode (%/year)	Coefficient of variation
Azithromycin	0.38	0.84	1.2	2.21
Cefixime	0.24	0.62	0.8	2.58
Ceftriaxone	0.13	0.34	0.2	2.62
Ciprofloxacin	0.97	1.79	1	1.84

Identifying capabilities to improve predictions by adding more AMR measurement data. Figure 5 shows the error reduction capacity of various countries for AMR predictions in the case of Azithromycin. Each vertical bar coloured in red and blue corresponds to one country. The red bar corresponds to the direct effect - the population-weighted fraction of error reduction for country X caused by increasing the number of tested persons in X by 30. The blue bar corresponds to the indirect effect - the population-weighted fraction of error reduction for countries other than X caused by increasing the number of tested persons in X by 30. Figure 5 also shows clusters of counties based on the PCA scores (projections) of their World Bank data points. We can identify countries causing maximal error reductions globally (Iran) or in various clusters (Austria, Myanmar, South Sudan, Sudan, USA).

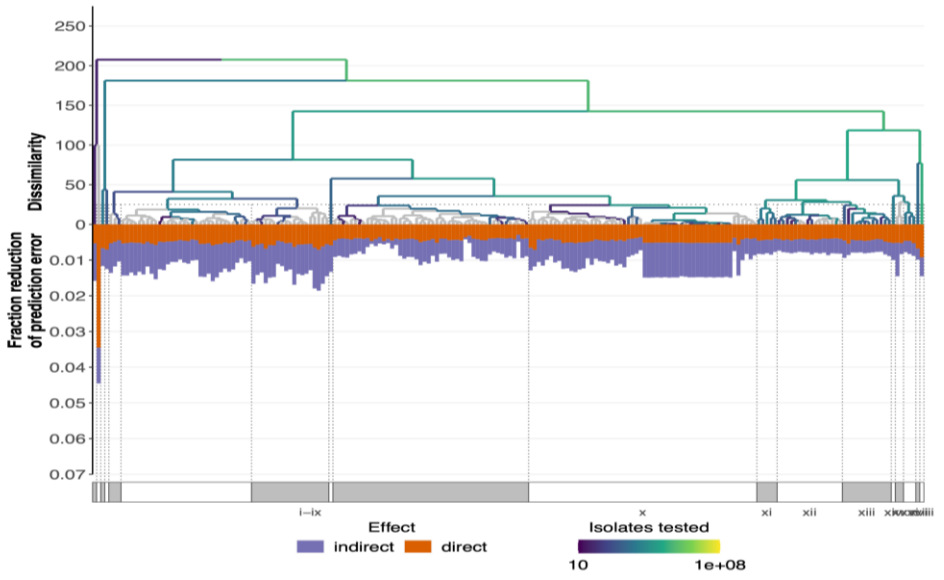

Figure 5. Distribution of error reduction for prediction of the AMR value of Azithromycin.

4. Conclusion

We have presented a description and implementation of two different methodologies for the prediction of AMR of Neisseria gonorrhoeae, a community-acquired pathogen, which are based on quantitative socio-economic indicators at the national level and AMR measurements for a subset of countries. These appear to be the first published macro (society) level AMR predictions for a community-acquired pathogen. Our work has extended the approach started in (Oldenkamp et al., 2021). In the minimal PCA-distance method, PCA is innovatively used in the space containing both independent (e.g. socio-economic) and AMR components. This feature addresses a problem mentioned in the literature (Artigue and Smith, 2019). However, this method generates more outliers compared to the beta-binomial principal component regression method. Both methods continue the current trend in machine learning to use PCA as a dimension-reduction tool for unsupervised learning. The observed MAE values and prediction errors may be related to the quality of AMR monitoring in various countries.

We describe several approaches to prioritize countries for the purpose of introducing of a new antibiotic. Our results can also be used by pharmaceutical companies to develop guidelines and strategies to prolong the efficacy of new antibiotics released in the market.

For better model training the introduction of additional variables, such as antibiotic consumption and treatment practices on the national level, in training sets can be considered. Future work also needs to be done to develop methods for AMR prediction on the sub-national level in the absence of adequate socio-economic data and AMR measurements at that level.

Acknowledgements

Help and consultations were provided by Antonio Cappuccio. The authors acknowledge partial funding from the following national funding agencies participating in the project MAGIcIAN JPI-AMR (https://www.magician-amr.eu/): Latvian Council of Science (LZP, Latvia) and the Italian Ministry of Education and Research (MIUR, Italy).

References

Alsan, M., Schoemaker, L., Eggleston, K., Kammili, N., Kolli, P., Bhattacharya, J. (2015). Out-of-pocket health expenditures and antimicrobial resistance in low-income and middle-income countries: an economic analysis. *The Lancet infectious diseases*, 15(10), 1203-1210. http://dx.doi.org/10.1016/S1473-3099(15)00149-8

Alvarez-Uria, G., Gandra, S., Laxminarayan, R. (2016). Poverty and prevalence of antimicrobial resistance in invasive isolates. *International Journal of Infectious Diseases*, 52, 59-61. http://dx.doi.org/10.1016/j.ijid.2016.09.026

Artigue, H., Smith, G. (2019). The principal problem with principal components regression. Cogent Math. Stat. 6, 1622190.

Bengio, Y., Courville A., Vincent, P. (2013). Representation Learning: A Review and New Perspectives, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(8), 1798-1828. http://dx.doi.org/10.1109/TPAMI.2013.50.

Blommaert, A., Marais, C., Hens, N., Coenen, S., Muller, A., Goossens, H. and Beutels, P. (2014). Determinants of between-country differences in ambulatory antibiotic use and antibiotic resistance in Europe: A longitudinal observational study. Determinants of between-country differences in ambulatory antibiotic use and antibiotic resistance in Europe: a longitudinal observational study. *Journal of Antimicrobial Chemotherapy*, **69**(2), 535-547. http://dx.doi.org/10.1093/jac/dkt377.

Breiman, L. (2001). Random Forests, Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324.

Bzdok, D., Altman, N., Krzywinski, M. (2018). Statistics versus Machine Learning, *Nature Methods*, **15**(4), 233-234. https://doi.org/10.1038/nmeth.4642.

Collignon, P., Athukorala, P.C., Senanayake, S. and Khan, F. (2015). Antimicrobial resistance: the major contribution of poor governance and corruption to this growing problem. *PloS one*, **10**(3), e0116746. http://dx.doi.org/10.1371/journal.pone.0116746

Collignon, P., Beggs, J.J., Walsh, T.R., Gandra, S., Laxminarayan, R. (2018). Anthropological and socioeconomic factors contributing to global antimicrobial resistance: a univariate and multivariable analysis. *The Lancet Planetary Health*, **2**(9), e398-e405 http://dx.doi.org/10.1016/S2542-5196(18)30186-4

Cook, R. D. (1979). Influential Observations in Linear Regression, *Journal of the American Statistical Association*. American Statistical Association. 74 (365): 169-174. http://dx.doi.org/10.1080/01621459.1979.10481634

Cyr, S. S., Barbee, L., Workowski, K. A., Bachmann, L. H., Pham, C., Schlanger, K., Torrone, E., Weinstock, H., Kersh, E. N., Thorpe, P. (2020). Update to CDC's treatment guidelines for gonococcal infection, 2020, *Morbidity and Mortality Weekly Report*, **69**(50), 1911–1916.

Daugulis P., Vagale V., Mancini E., Castiglione F. (2022). A PCA-based data prediction method, *Baltic J. Modern Computing*, **10**(1), 1-16. http://dx.doi.org/10.22364/bjmc.2022.10.1.01.

Farhat, F., Athar, M. T., Ahmad, S., Madsen, D. Ø., Sohail, S. S. (2023). Antimicrobial resistance and machine learning: past, present, and future, Frontiers in Microbiology, 14. http://dx.doi.org/10.3389/fmicb.2023.1179312

Feretzakis, G., Sakagianni, A., Loupelis, E., Kalles, D., Skarmoutsou, N., Martsoukou, M., Christopoulos C, Lada M, Petropoulou S, Velentza A, Michelidou S, Chatzikyriakou R, 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), 214-221. http://dx.doi.org/10.4258/hir.2021.27.3.214

Getis, A. (2010). The Analysis of Spatial Association by Use of Distance Statistics, *Geographical Analysis*, **24**(3), 189–206. http://dx.doi.org/10.1111/j.1538-4632.1992.tb00261.x.

Gorban, A. N, Kegl, B., Wunsch, D. C., Zinovyev, A. (eds) (2008). *Principal Manifolds for Data Visualisation and Dimension Reduction*, LNCSE 58, Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-73750-6.

Hosmer, D., Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd edition, Wiley, New York. http://dx.doi.org/10.1002/0471722146

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, **24**(6), 417-441. http://dx.doi.org/10.1037/h0071325.

Hotelling, H. (1936). Relations between two sets of variates, *Biometrika*, **28**(3/4), pp.321-377. http://dx.doi.org/10.2307/2333955.

Ison, C. A., Deal, C., Unemo, M. (2013). Current and future treatment options for gonorrhoea. *Sexually Transmitted Infections*, **89**(SUPPL 4), 52–57. https://doi.org/10.1136/sextrans-2012-050913

Kim, M. G. (2017). A cautionary note on the use of Cook's distance. *Communications for Statistical Applications and Methods*, **24**(3), 317-324. http://dx.doi.org/10.5351/CSAM.2017.24.3.317.

Kvalseth, T. O. (2017). Coefficient of variation: the second-order alternative, *Journal of Applied Statistics*, **44**(3), 402-415. http://dx.doi.org/10.1080/02664763.2016.1174195.

Lewis, D. A. (2014). Global resistance of Neisseria gonorrhoeae: When theory becomes reality. *Current Opinion in Infectious Diseases*, **27**(1), 62–67. https://doi.org/10.1097/QCO.0000000000000025.

Lewis, D. A. (2019). New treatment options for Neisseria gonorrhoeae in the era of emerging antimicrobial resistance, *Sexual Health*, 16(5), 449–456. http://dx.doi.org/10.1071/SH19034.

Mai, T. T., Lees, J. A., Gladstone, R. A., Corander, J. (2023). Inferring the heritability of bacterial traits in the era of machine learning, *Bioinformatics Advances*, **3**(1). http://dx.doi.org/10.1093/bioadv/vbad027

Meglen, R. R. (1991). Examining Large Databases: A Chemometric Approach Using Principal Component Analysis, *Journal of Chemometrics*, **5**(3), 163-179. http://dx.doi.org/10.1002/cem.1180050305.

Meijering, E. (2002). A chronology of interpolation: from ancient astronomy to modern signal and image processing, *Proceedings of the IEEE*, **90**(3), 319-342. http://dx.doi.org/10.1109/5.993400.

Mittal, S. (2016). A Survey of Techniques for Approximate Computing, *ACM Computing. Surveys*, **48**(4), 1–33. http://dx.doi.org/10.1145/2893356.

Oldenkamp R., Schultsz C., Mancini E., Cappuccio A. (2021). Filling the gaps in the global prevalence map of clinical antimicrobial resistance, *Proceedings of the National Academy of Sciences*, **118**(1), e2013515118. http://dx.doi.org/10.1073/pnas.2013515118.

Sakagianni, A., Koufopoulou, C., Feretzakis, G., Kalles, D., Verykios, V. S., Myrianthefs, P. (2023). Using Machine Learning to Predict Antimicrobial Resistance—A Literature Review, *Antibiotics*, **12**(3), 452. http://dx.doi.org/10.3390/antibiotics12030452

Shalev-Shwartz, S., Ben-David, S. (2014). *Decision Trees, Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press; 212-218. http://dx.doi.org/10.1017/CBO9781107298019.019

Stacklies, W., Redestig, H., Scholz, M., Walther, D., Selbig, J. (2007). pcaMethods—a bioconductor package providing PCA methods for incomplete data, *Bioinformatics*, **23**(9), 1164-1167. http://dx.doi.org/10.1093/bioinformatics/btm069

Tang, R., Luo, R., Tang, S., Song, H., Chen, X. (2022). Machine learning in predicting antimicrobial resistance: a systematic review and meta-analysis, *International Journal of Antimicrobial Agents*, **60**(5–6). https://doi.org/10.1016/j.ijantimicag.2022.106684

Tapsall, J. (2002). Current concepts in the management of gonorrhoea. *Expert Opinion on Pharmacotherapy*, **3**(2), 147–157. http://dx.doi.org/10.1517/14656566.3.2.147.

Unemo, M., Golparian, D., Potočnik, M., Jeverica, S. (2012). Treatment failure of pharyngeal gonorrhoea with internationally recommended first-line ceftriaxone verified in Slovenia, September 2011, *Eurosurveillance*, **17**(25), 1–4. http://dx.doi.org/10.2807/ese.17.25.20200-en.

Unemo, M., Lahra, M. M., Cole, M., Galarza, P., Ndowa, F., Martin, I., Dillon, J. A. R., Ramon-Pardo, P., Bolan, G., Wi, T. (2019). World Health Organization Global Gonococcal Antimicrobial Surveillance Program (WHO GASP): Review of new data and evidence to inform international collaborative actions and research efforts. *Sexual Health*, **16**(5), 412–425. http://dx.doi.org/10.1071/SH19023.

Unemo, M., Nicholas, R. A. (2012). Emergence of multidrug-resistant, extensively drug-resistant and untreatable gonorrhea, *Future Microbiology* **7**(12), 1401–1422. http://dx.doi.org/10.2217/fmb.12.117.

Yang, F., Yan, J. (2020). Antibiotic Resistance and Treatment Options for Multidrug-Resistant Gonorrhea, *Infectious Microbes and Diseases*, **2**(2), 67–76. http://dx.doi.org/10.1097/IM9.0000000000000024.

Yee, T. W. (2010). The VGAM Package for Categorical Data Analysis, *Journal of Statistical Software*, **32**(10), 1–34. https://doi.org/10.18637/jss.v032.i10.

Zimek, A., Schubert, E. (2017). *Outlier Detection, Encyclopedia of Database Systems*, Springer New York, 1-5. http://dx.doi.org/10.1007/978-1-4899-7993-3_80719-1.

WEB (a) Global Antimicrobial Resistance and Use Surveillance System (GLASS), available at https://www.who.int/initiatives/glass.

WEB (b) Multi-drug resistant gonorrhoea, available at https://www.who.int/news-room/fact-sheets/detail/multi-drug-resistant-gonorrhoea.
WEB (c) World DataBank: World Development Indicators, available at https://databank.worldbank.org/reports.aspx?source=world-development-indicators.
WEB (d) World Population Prospects 2019, Volume I: Comprehensive Tables (ST/ESA/SER.A/426), Department of Economic and Social Affairs, 1-395, ISBN: 978-92-1-148327-7, available at https://population.un.org/wpp/Publications/Files/WPP2019_Volume-I_Comprehensive-Tables.pdf
WEB (e) World Health Organization (2021). Global antimicrobial resistance and use surveillance system (GLASS) report 2021. WHO: Geneva, Switzerland, ISBN 9240027335, 9789240027336, available at https://www.who.int/publications/i/item/9789240027336.

Received September 18, 2023, revised January 6, 2024, accepted January 9, 2024