Predicting antimicrobial resistance using historical bacterial resistance data with machine learning algorithms

Raquel Urena (raquel.urena@univ-amu.fr)
Aix Marseille Univ https://orcid.org/0000-0002-4099-7437
Camiade Sabine
AlphaBio
Yasser Baalla
Sesstim, Aix Marseille Univ
Martine Piarroux
Centre d'épidémiologie et de santé publique des armées (CESPA)
Philippe HALFON
Laboratoire Alphabio / Hôpital Européen
Jean Gaudart
Sesstim, Aix Marseille Univ / BioSTIC, APHM
Jean Charles Dufour
Sesstim, Aix Marseille Univ / BioSTIC, APHM
Stanislas Rebaudet
Sesstim, Aix Marseille Univ / Hôpital Européen

Article

Keywords:

Posted Date: January 30th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2519978/v1

Abstract

Antibiotic resistance of bacterial pathogens is considered by the World Health Organization as a major threat to global health aggravated by the misuse of antibiotics. In clinical practice results of bacterial cultures and antibiograms can take several days. In the meantime, prescribing an empirical antimicrobial treatment constitutes a challenge in which the practitioner has to strike a balance between antibiotics spectrum and expected susceptibility probability. In this contribution, we report the development and testing of a machine-learning-based system that early predicts the antimicrobial susceptibility probability and provides explanations of the contribution of the different cofactors at 4 different stages prior to the antibiogram (sampling, direct examination, culture, and species identification stages). A comparative analysis of different state of the art machine learning and probabilistic methods was carried out using 7 years of historical bacterial resistance data from the Hôpital Européen Marseille, France. Our results suggest that dense neural network-based models and Bayesian models are suitable to early predict antibiotics susceptibility (average AUC 0.91 at the species identification stage) even for the less frequent situations.

Introduction

According to the World Health Organization, antibiotic resistance of bacterial pathogens is “one of the biggest threats to global health” [1] and it is accelerated by the misuse of antibiotics [1]. In clinical practice, initial antimicrobial treatment is often prescribed empirically until results of bacterial cultures and antibiograms are obtained. This is a challenging gamble [2] based on probabilistic reasoning, which can get complex, even for experienced clinicians. Indeed, it aims to balance the broadest accepted spectrum of the antibiotics, against the lowest required susceptibility probability of suspected germs depending on the severity of illness or risk of treatment failure. To mitigate this risk and save broad-spectrum molecules for critical and resistant infections, prescribers try to rely on guidelines edited by earned societies of infectious diseases [3,4]. They are also supposed to take into account the local microbial ecology, which often proves difficult and gets unrealistic when considering patient subgroups based on demographic or comorbidity characteristics [5]. Considering that one in three patients in European hospitals receive at least one antimicrobial [6], this is a daily issue.

Clinical decision support systems (CDSS) may guide prescribers in their choices [7,8]. But so far, historical resistance data appear only used to create institutional or unit-specific antibiograms for surveillance purposes [8,9], and only few machine-learning CDSSs (ML-CDSS) have been developed to predict drug resistance or guide the choice of antimicrobials [5,8, 10–13]. Besides, most of them integrate numerous variables, which cannot be readily extracted from most electronic health records (EHR), and which analysis in Europe has become strongly controlled since the 2018 General Data Protection Regulation (GDPR) [14].

In this context, we aim to develop a CDSS to help clinicians to choose empiric antimicrobial therapy based on easily available hospital data. The purpose of this first work was thus to benchmark different machine learning algorithms predicting antimicrobial resistance, trained with the sole historical bacterial resistance data of Hôpital Européen Marseille, a French general hospital.

Results

Data description and antibiotic resistance distribution

From January 2014 to December 2020, the bacteriological laboratory performed 44,026 antibiograms (Table 1). After removing likely contaminants, duplicates, and screening specimens for carriage of multidrug-resistant (MDR) bacteria, we analyzed a cleaned dataset of 30,975 antibiograms, isolated from 13,166 patients in 6 different types of wards, mainly the emergency room (25%), critical care (24%), surgery (20%) or medicine (19%). Most bacteria were isolated from urine (41%), lower respiratory tract (23%), blood (14%) or abscess (12%) samples. Isolated bacteria were mostly Gram-negative rods (71%), including enterobacteriaceae-like (56%) and non-fermentative (11%) Gram-negative rods, and Gram-positive cocci (27%), including Staphylococcus-like (13%) and Enterococcus-like (10%) Gram positive cocci (Table 1). Main species included *Escherichia coli* (29%), *Staphylococcus aureus* (12%), *Klebsiella pneumoniae* (10%), *Enterococcus faecalis* (9%) or *Pseudomonas aeruginosa* (8%) (Table 1). The cleaned dataset was subdivided into a (80% training + 20% test)dataset of 26,621 antibiograms from January 2014 to December 2019, including 4,915 antibiograms in 2019 alone, and a validation dataset of 4,354 antibiograms from January to December 2020 (Table 1).

Table 1 – Characteristics of the raw and the cleaned antibiogram datasets from 2014 to 2020

	Raw dataset	Cleaned dataset
No. of antibiograms (% of total)	44,026	30,975 (70%)
No. of patients	14,671	13,166
No. of likely contaminant isolates (% of total)	5,630 (13%)	0
No. of duplicates (% of total)	2,023 (5%)	0
No. of MDR screening specimens (% of total)	5,743 (13%)	0
No. of different wards	38	6
No. of antibiograms per main ward type (% of total)		
Emergency room		7,601 (25%)
Critical care		7,521 (24%)
Surgery		6,295 (20%)
Medicine		5,985 (19%)
Operating endoscopy or radiology rooms		2,907 (9%)
Day hospital		666 (2%)
No. of different specimens	38	13
No. of antibiograms per main specimen type (% of total)		
Urine, or urethra		12,604 (41%)
lower respiratory tract, or pleura		7,014 (23%)
Blood, or blood catheter		4,483 (14%)
Abscess		3,772 (12%)
Skin, or wound		1,564 (5%)
Joint or bone		867 (3%)
Stool, gut, or ascites		283 (1%)
Device		158 (1%)
ear, nose and throat, or eye		143 (0%)
Unspecified biopsy		69 (0%)
Female genital tract		11 (0%)
Cerebrospinal fluid		7 (0%)
No. of antibiograms per main direct type (% of total)		
Gram-negative rods		22,043 (71%)
Gram-positive cocci		8,415 (27%)
Gram-positive rods		361 (1%)
No. of antibiograms per main culture type (% of total)		
Enterobacteriaceae-like Gram-negative bacteria		17,216 (56%)
Staphylococcus-like Gram-positive cocci		3,934 (13%)
Non-fermentative Gram-negative rods		3,392 (11%)
Enterococcus-like Gram positive cocci		3,135 (10%)
Streptococcus-like Gram-positive cocci		1,188 (4%)
Unspecified Gram-negative rods		1,116 (4%)
Anaerobic bacteria		591 (2%)
No. of antibiograms per main species (% of total)		
Escherichia coli		8,861 (29%)
Staphylococcus aureus		3,565 (12%)
Klebsiella pneumoniae		3,085 (10%)
Enterococcus faecalis		2,650 (9%)
Pseudomonas aeruginosa		2,453 (8%)
Proteus mirabilis		973 (3%)
Haemophilus influenzae		670 (2%)
Enterobacter cloacae ssp cloacae		666 (2%)
Klebsiella aerogenes		555 (2%)
Streptococcus agalactiae		538 (2%)
Stenotrophomonas maltophilia		522 (2%)
Klebsiella oxytoca		508 (2%)
Serratia marcescens		484 (2%)
Enterobacter cloacae complex		379 (1%)
No. of antibiograms per history of MDR carriage (% of total)		
MDR carriage during the past 3 months		3,435 (11%)
No MDR carriage during the past 3 months		5,913 (19%)
Unknown patient		21,647 (70%)
No. of antibiograms in the training+test dataset (Jan 2014-Dec 2019)		26,621 (86%)
No. of antibiograms in the last year of the training+test dataset (Jan-Dec 2019)		4,915 (16%)
No. of antibiograms in the validation dataset (Jan-Dec 2020)		4,354 (14%)

MDR, multidrug resistant bacteria

The mean traditional antibiograms (22 single antibiotics) and combination antibiograms (25 antibiotic combinations) are plotted on Fig. 1 and summarized in Supplementary Table S1, for the main types of each category of isolates. Direct types, culture types and species exhibited an important heterogeneity of susceptibility to single antibiotics (root-mean-square deviation [RMSD], 33%, 37% and 43%, respectively). Susceptibility rates were quite heterogeneous between specimen types too (RMSD, 14%): for instance, susceptibility to Amoxicillin-Clavulanate (Amox-Clav) was 56% in urine, but 35% in lower respiratory tract samples. The global heterogeneity was less important but still discriminant between ward types (7%): susceptibility to Cefotaxime / Ceftriaxone (CTX / CRO) was only 49% for critical care, unlike 77% for emergency room isolates. Similarly, categories of history of MDR carriage (RMSD, 11%) showed important susceptibility differences to certain antibiotics:

susceptibility to CTX / CRO was only 27% in case of MDR carriage during the past 3 months, 54% if no MDR carriage was documented for a known patient, and 72% for isolates from previously unknown patients in the hospital. Finally, the global susceptibility to single antibiotics appeared stable over time (RMSD between periods, 2%), although CTX / CRO, Ertapenem, Amikacin, or Trimethoprim-Sulfamethoxazole (TMP-SMX) exhibited remarkable differences (Fig. 1, Supplementary Table S1).

Comparison of antibiotic susceptibility prediction models for all isolates

Using the 2014-2019 laboratory dataset (80% training + 20% test) and available covariates (specimen origin, type of ward, previous multidrug resistance [MDR] bacteria carriage and sample date), we trained different frequentist (FRQ) and Bayesian (BAY) inference models, as well as machine learning (ML) algorithms (Logistic Regression [LR], AdaBoost [ADA], Gradient Boosting [GBS], Extreme Gradient Boosting [XGB], Bagging [BAG], Random Forest [RF] and neural networks [NN]) to predict antibiotic susceptibility at each stage until antibiogram results. These 4 stages were defined as: (1) “sampling” stage (specific ecology for a body site); (2) “direct” stage, after Gram stain examination of the sample; (3) “culture” stage, after macroscopic and Gram stain examination of positive cultures; and (4) “species” stage, after bacterial identification of positive cultures. We then used these models to predict the susceptibility probability to each of the 22 single antibiotics and 25 antibiotic combinations for isolates of the 2020 dataset (validation). Based on the predictions of susceptibility to single and combined antibiotics of all isolates of the 2020 validation dataset, mean ROC (Receiver Operator Characteristic) curves for each stage of the identification process and each model were plotted (Fig. 2) and mean AUC (Area Under the ROC Curve) were estimated (Table 2). AUC for each stage and each antibiotic were also estimated and displayed (Fig. 3).

Table 2 – Mean areas under the receiver operator characteristic curve (ROC AUC) of frequentist, Bayesian and machine learning susceptibility prediction models at each stage of the identification process.

Model	Stage 1 "sampling"	Stage 2 "Direct"	Stage 3 "Culture"	Stage 4 "Species"	Overall mean
			All isolates of the 2020 validation dataset		
BAY	0.677	0.782	0.861	0.918	0.809
NN	0.693	0.811	0.875	0.915	0.823
FRQ	0.672	0.785	0.856	0.893	0.802
XGB	0.677	0.769	0.847	0.912	0.801
GBS	0.523	0.647	0.712	0.827	0.677
ADA	0.530	0.649	0.725	0.823	0.682
BAG	0.525	0.656	0.728	0.833	0.686
RF	0.529	0.658	0.725	0.828	0.685
LR	0.522	0.617	0.627	0.672	0.610
Overall mean	0.594	0.708	0.773	0.847	0.731
			Isolates of the 2020 validation dataset corresponding to the least frequent situations only*		
BAY	0.584	0.749	0.824	0.917	0.769
NN	0.512	0.658	0.771	0.851	0.698
FRQ	0.586	0.648	0.692	0.662	0.647
XGB	0.483	0.659	0.685	0.848	0.669
GBS	0.516	0.606	0.678	0.776	0.644
ADA	0.518	0.617	0.672	0.760	0.642
BAG	0.508	0.615	0.697	0.776	0.649
RF	0.520	0.616	0.671	0.739	0.637
LR	0.488	0.563	0.573	0.591	0.554
Overall mean	0.524	0.637	0.696	0.769	0.656

BAY, Bayesian; NN, Neural Network; FRQ, frequentist; XGB, Extreme Gradient Boosting; GBS, Gradient Boosting; ADA, AdaBoost; BAG, Bagging; RF, Random Forest; LR, Logistic regression
* 5th percentile of the number of occurrences in the 2014-2019 dataset

The overall mean AUC of the frequentist, Bayesian and the seven ML models increased from 0.594 at stage 1 “sampling”, to 0.708 at stage 2 “direct”, 0.773 at stage 3 “culture” and 0.847 at stage 4 “species”, with important differences between antibiotics and models (Fig. 2, Table 2 and Fig. 3).

At all stages, the BAY, NN, FRQ and XGB models exhibited much better prediction performances than other models (GBS, ADA, BAG, RF and LR) (Fig. 2 and Table 2). Notably, NN models showed the highest mean AUC (0.823) and the highest AUC at each stage but stage 4, where the BAY model AUC reached 0.918. Conversely, mean AUCs of FRQ models were slightly impaired at late stages, as prediction failures increased

for situations of the 2020 dataset that were unmet in the 2019 dataset: from 0.7% at stage 1, to 1.4% at stage 2, 2.6% at stage 3 and 10.2% at stage 4 (Supplementary Table S1). Failure rates remained null for Bayesian and ML models.

Predictability was very heterogeneous between antibiotics (Fig. 3), with model-specific standard deviations of AUCs between 0.03 and 0.17 (Fig. 2), especially at stage 2 and 3. Among all models, NN exhibited the lowest standard deviations (from 0.06 to 0.09 depending on the stage). Prediction performances of BAY models were quite consistent between antibiotics too. For instance, AUC of BAY models ranged from 0.566 for TMP-SMX to 0.850 for Meropenem + Vancomycin at stage 1, and from 0.719 for Levofloxacin to 1.000 for Metronidazole at stage 4 (Supplementary Table S1).

Comparison of antibiotic susceptibility prediction models for the least frequent situations

To assess the prediction performances of models for rare situations, distribution of AUC for the least frequent situations only (5^{th} percentile of the number of occurrences in the 2014-2019 dataset) was plotted (Fig. 4), and mean AUC were estimated (Table 2).

The combined overall AUC of all models decreased from 0.731 to 0.656 when restricting to the rarest situations (Table 2). LR remained the worst model at all stages (mean AUC at 0.554) (Table 2 and Fig. 4). Conversely, BAY models kept good performances, with mean AUC as high as 0.824 and 0.917 for stages 3 and 4, respectively. Mean AUC for NN models were less preserved, especially for stage 1 and 2 where they dropped from 0.693 to 0.512 and from 0.811 to 0.658, respectively. For FRQ models, mean AUC at stage 4 fell down from 0.893 to 0.662, as many situations were unmet in the dataset used to train them (Table 2, Fig. 4 and Supplementary Table S1).

Overall, based on these comparisons, the best models appeared to be the BAY and NN models.

Contribution of cofactors and model interpretability

To interpret predictions of NN models, we applied the SHapley Additive exPlanations (SHAP), and first assessed the importance of each cofactor in the NN model at each stage. Fig 5 depicts the importance of each covariable at each stage in decreasing order of importance. The color indicates the cofactors' value, for example for the case of BMR, red indicates positive known BMR carriage, gray indicates not known BMR carriage and blue indicates negative BMR known carriage. As we can observe, the past history of BMR carriage contributed much to the performance of each model, and even was the most important cofactor at stage 1 (Sampling) and 2 (Direct). Direct type also had a very significant impact on the prediction at stage 2. The culture type contributed most to the NN model performance at stage 3 (Culture) and it keeps being relevant at stage 4, after the species. The date and the sample type exhibited less significant impact on predictions at all stages.

As an illustration of how cofactors can influence the prediction of antibiotic susceptibility in the NN models, we plotted SHAP values at the individual level using a didactic scenario (Fig. 6). In the first example, the predicted susceptibility to CTX / CRO at stage 3 (culture) is 0.87 (Fig. 6A); the culture type (Enterobacteriaceae-like Gram-negative bacteria) and the unknown history of BMR carriage (in red) explain this good probability. In the last scenario, the predicted susceptibility to CTX / CRO at stage 4 (species) is 0.40 (Fig. 6E); the species (*Citrobacter koseri*), the recent history of a BMR carriage and the critical care ward (in blue) explain this bad probability; they are in part compensated by the Enterobacteriaceae-like Gram-negative bacteria type of this species (in red).

Using the same scenario, NN and BAY models exhibited close susceptibility predictions, except for least frequent situations (Fig. 6): for the last prediction, the training+test dataset included only one similar situation, and the predicted susceptibility to CTX / CRO was 40% with NN and 78% with BAY models (Fig. 6E). Covariate-specific SHAP values (NN) and corresponding likelihood ratios (BAY) appeared moderately consistent: on these examples, overall Spearman's rank correlation , 0.57 (p-value, 0.0033). Correlation appeared the best for the BMR carriage history (specific , 0.92), or the ward (0.89) (Fig. 6). But for culture types and species at stage 4, both NN and BAY showed more contrasted SHAP values and likelihood ratios, as NN models included both cofactors whereas BAY models only included the species (Fig. 6B and 6E).

Discussion

To our knowledge, we present here the first study aiming to predict the susceptibility to antibiotics in a hospital setting and throughout the successive stages of the bacteriological process (sampling, direct examination, culture, and species identification). Overall, models showed a good predictability of antibiotic susceptibility or resistance, even at early stages of the identification process (stages 1 "sampling" or 2 "direct"), with mean AUCs for all isolates of the 2020 validation dataset reaching 0.693 and 0.811, respectively. Performances got even better at subsequent stages, at 0.875 and 0.918, at stages 3 "Culture" and 4 "Species", respectively. But predictability was heterogeneous between antibiotics, and prediction performances were heterogeneous between models as well. Neural networks (NN) and Bayesian inference (BAY)

models exhibited the most consistent performances. Both exhibited the highest mean AUC at early (sampling and direct) and late (culture and species) stages, even for the least frequent situations. Their AUCs also exhibited the lowest standard deviations between antibiotics. In particular, BAY models kept good and steady AUCs in rare situations, as these models were built to predict antibiotic susceptibility using pretest probabilities and likelihood ratios estimated from more general data. But this tailored model family was not easy to program, whereas NN models could be implemented and tuned using open source packages and they are scalable to new variables.

Conversely, performances of frequentist inference models (FRQ) were impaired at late stages, as prediction failures increased for rare specific situations. Logistic regression (LR) showed very bad performances, even at late stages. All ensemble models (AdaBoost (ADA), Gradient Boosting (GBS), Bagging (BAG) and Random Forest (RF)) but Extreme Gradient Boosting (XGB) showed markedly lower performances than BAY and NN models. The fact that XGB performs more satisfactory than GBS could be motived by the higher regularization measures applied in XGB, that both prevents the overfitting and reduces the training times.

All covariates did not exhibit similar predicting weight. In the whole dataset, direct types, culture types and species categories exhibited an important heterogeneity of susceptibility to antibiotics, with high RMSD: this supports our choice to develop specific prediction models for each stage of the identification process at the laboratory. Categories of specimen types, previous multidrug resistant (MDR) bacteria carriage, and ward types exhibited intermediate levels of susceptibility heterogeneity, which justifies integration of these covariates in our models. Although the global susceptibility to antibiotics showed a low heterogeneity between period categories, it remained important to include this temporal covariate in models because resistance to certain molecules like Ceftriaxone / Cefotaxime, Ciprofloxacin or TMP-SMX is known to rapidly evolve over time, which was also observed in our dataset.

This heterogeneity was included within the BAY models through the combination of feature-specific likelihood ratios, which can provide a direct interpretability of predictions. For NN and other ML models, we used the properties of Shapley values to visualize how each covariate contributes to push the predicted antibiotic susceptibility from the base value at both population and individual level. All covariates contributed to prediction results with highest weights for direct, culture, species and history of BMR carriage categories depending on the lab stages. Influence of covariates on the successive predictions of our example was consistent between the NN models (SHAP values) and the corresponding BAY models (likelihood ratios), although discrepancies suggested distinct learning strategies, especially at stage 4 because of the respective architecture of NN and BAY models. Both also exhibited sometimes divergent predictions, especially for rare situations. Overall, these results comfort the partial interpretability of our ML results.

For each stage of the identification process, our models included covariates available in the laboratory database such as sample types, ward types and history of past sample and BMR carriage. However, they did not include demographics, comorbidities and antibiotic treatment history although such covariates have been often associated with antibiotic resistance [5]. Including these covariables could certainly improve models' performances, especially at early stages of the identification process, as suggested by the AUC reported by Yelin et al [5] in community-acquired urinary tract infections: demographics and records of past urine cultures and history of drug purchases of the patients showed high levels of relative importance in their gradient-boosting decision trees; and their complete logistic regression model with these covariates provided AUC from 0.7 (for Amox-Clav) to 0.83 (for ciprofloxacin), whereas AUC of our BAY models at stage 1 for urine specimens were only 0.57 for amoxicillin-CA, 0.58 for ciprofloxacin, and only reached 0.67 for CTX / CRO (data not shown). We plan future analyses including such hospital-level health records in NN and other ML models, in order to evaluate prediction gains.

Nevertheless, as such patient-related sensitive data have become strictly controlled by the European GDPR and are stored in electronic patient files of various formats and accessibility levels, they cannot be readily available in every hospital facility.

An important advantage of the models in this contribution is that they are performant using only minimal health records, which can be easily extracted as spreadsheet files at every bacteriological laboratory, with no GDPR restriction after onsite anonymization. They could thus be deployed in other hospitals with limited effort, as a standalone mobile application, which could be used at the patients' bedside by clinician doctors without the need of filling long online forms. They could also be directly embedded in the Lab Information Management System (LIMS) to provide preliminary probabilistic antibiograms, or in electronic medical record (EMR) systems or health data warehouses [15]. This multisite deployment, combined with a regular update of lab databases and models re-training to keep then updated to the bacterial ecology, would allow pooling data and increasing prediction performances, considering time and space heterogeneity of resistance to antimicrobials.

In this contribution we have chosen to calculate the probability of susceptibility instead of the most likely result (i.e. susceptible or resistant), since, in our opinion, the proposed tool shall not replace prescribers' decisions but remain a simple clinical decision support to their complex probabilistic reasoning and risk management [7,8]. Such tool shall also not replace practice guidelines [3,4], but help prescribers to adapt these empiric antimicrobial therapy recommendations to the finest bacterial ecology available data, throughout the stages of the bacteriological process prior to the antibiogram. This could mitigate the risk of treatment inadequacy and failure, save broad-spectrum molecules, or prevent bacterial resistance selection. We thus believe that the routine use of such prediction models trained with lab-data for the early initiation and

re-evaluation of empiric antimicrobial therapy could significantly improve the relevance of empiric antimicrobial prescriptions, even at early stages of the identification process, and substantially improve the current standard of care. They may be of particular help for on-call infectiologists requested for antimicrobial stewardship advice from unfamiliar hospitals, or in countries where antibiograms are not routinely available and where antimicrobial resistance creates the most serious problems [16]. A demo app based on Gradio Python package is being tested at Hôpital Européen Marseille. Several studies are underway to assess the potential interest of such algorithms for antimicrobial stewardship, by comparing actually prescribed or theoretically recommended antimicrobial regimens to susceptibility predic tions. Together with the present work, we hope they will contribute to the struggle against the antibiotic resistance pandemic.

Online Methods

Settings and study design

The study was conducted in Hôpital Européen Marseille, a French general hospital located in the popular city-center of a nearly 900,000 inhabitant city in southeastern France. The hospital has been fully paperless since its opening in 2013, and all inpatient and outpatient information is electronically stored in the same comprehensive electronic health record (EHR) called QCare ICU Manager (Health Information Management GmbH, Bad Homburg, Germany).

The study was designed as a monocentric retrospective analysis of positive bacterial culture and susceptibility data (antibiograms) performed at Hôpital Européen Marseille between January 2014 and December 2020.

Using covariates available in the laboratory database (specimen origin, type of ward, previous multidrug resistance [MDR] bacteria carriage and sample date), we first designed different frequentist, Bayesian and state of the art machine learning models predicting antibiotic susceptibility (Supplementary Table S2), for each stage preceding antibiogram results: (1) “sampling” stage (specific ecology for a body site); (2) “direct” stage, after Gram stain examination of the sample when available at day 0 or 1; (3) “culture” stage, after macroscopic and Gram stain examination of positive cultures, usually from day 1 to 3 ; and (4) “species” stage, after bacterial identification of positive cultures, usually at day 2 or 3. We trained models on the 2014–2019 dataset (training + test), and used them to predict the susceptibility probability to each of the 22 single antibiotics and 25 antibiotic combinations for isolates of the 2020 dataset (validation).

Data

We extracted antibiograms from microbiology electronic records of the accredited bacteriology laboratory of the hospital between January 2014 and December 2020. Specimen culturing followed the latest guidelines of French microbiology earned societies [17]. Bacterial identification was performed using matrix assisted laser desorption ionisation/time of flight mass spectrometry (MALDI/ToF MS) with VITEK® MS (bioMérieux France, Craponne, France). Antibiotic susceptibility testing was performed using VITEK® 2 automated system (bioMérieux France) or diffusion techniques on Agar plates when relevant, and SIRxpert Master® software (i2a, Montpellier, France) based on the current French and European guidelines (CA-SFM EUCAST) [18].

Covariate definition and data preparation

Based on metadata and results of each antibiogram, we characterized: (1) the type of ward (i.e. emergency room, intensive care, medicine, surgery, day care unit…); (2) the body site origin of the specimen (i.e. blood or intravenous catheter, urine, lower respiratory, joint/bone, digestive, genital, cerebrospinal…); (3) the history of a previous multidrug resistant (MDR) bacteria carriage for the same patient within the past three months (i.e. usual threshold for risk factors of multiresistant bacterial infections) [19]; (4) the clinical relevance of bacterial isolates considering the body site (i.e. relevant or likely contaminant such as a unique blood culture positive for common bacteria of the skin flora like *Staphylococcus epidermidis*) [20]; (5) duplicates (i.e. similar isolates and antibiogram within the past two days). To provide antibiotic susceptibility predictions based on information available at the “direct” and “culture” stage preceding species identification usually at D0-D2, we also characterized: (6) the typical Gram stain features on direct examination of the sample (i.e. Gram positive cocci, Gram-negative rods…); and (7) the typical macroscopic and Gram stain features of positive cultures (i.e. Staphylococcus-like Gram-positive cocci, enterobacteriaceae-like Gram-negative bacteria, non-fermentative Gram-negative rods…) (Supplementary Table S3).

Finally, we interpreted (8) the susceptibility to 22 single antibiotics (traditional antibiogram) and 25 antibiotic combinations (combination antibiogram) of clinical interest (Table 3). Isolates with intermediate susceptibility were classified as resistant to the antibiotic. For strains whose susceptibility results were not available on a specific antibiotic, we interpreted susceptibility based on recommended expert rules on intrinsic and cross resistances [18,21].

Table 3
– List of analyzed single antibiotics and antibiotic combinations of clinical interest

Single antibiotics	Antibiotic combinations
1. Amoxicillin	1. Amox-Clav + Amikacin
2. Amox-Clav	2. Pip-Tazo + Amikacin
3. Oxacillin / Cefazolin	3. CTX / CRO + Amikacin
4. Pip-Tazo	4. Ceftazidime + Amikacin
5. Cefotaxime / Ceftriaxone	5. Cefepime + Amikacin
6. Ceftazidime	6. Aztreonam + Amikacin
7. Cefepime	7. Imipenem + Amikacin
8. Aztreonam	8. Meropenem + Amikacin
9. Imipenem	9. Amox-Clav + Gentamicin
10. Meropenem	10. Pip-Tazo + Gentamicin
11. Ertapenem	11. Imipenem + Gentamicin
12. Amikacin	12. Vancomycin + Gentamicin
13. Gentamicin	13. Amox-Clav + Ciprofloxacin
14. Ciprofloxacin	14. Pip-Tazo + Ciprofloxacin
15. Levofloxacin	15. CTX / CRO + Ciprofloxacin
16. TMP-SMX	16. Ceftazidime + Ciprofloxacin
17. Vancomycin	17. Cefepime + Ciprofloxacin
18. Rifampicin	18. Aztreonam + Ciprofloxacin
19. Clindamycin	19. Imipenem + Ciprofloxacin
20. Macrolides	20. Meropenem + Ciprofloxacin
21. Linezolid	21. Pip-Tazo + Vancomycin
22. Metronidazole	22. Cefepime + Vancomycin
	23. Meropenem + Vancomycin
	24. CTX / CRO + Metronidazole
	25. Cefepime + Metronidazole
Amox-Clav, Amoxicillin-Clavulanate (Augmentin*)	
Pip-Tazo, Piperacillin-Tazobactam (Tazocilline*)	
TMP-SMX, Trimethoprim-Sulfamethoxazole (Bactrim*)	
CTX / CRO, Cefotaxime / Ceftriaxone	

After removing likely contaminants, duplicates and MDR bacteria carriage screening specimens (rectal and nose swabs), we estimated mean antibiotic susceptibility rates and constructed mean antibiograms for single antibiotics (traditional antibiograms) or antibiotic combinations (combined antibiograms) [22], for each type of each category of isolates: bacterial species; bacterial type on direct examination; bacterial type on culture; specimen type; ward type; and past history of MDR carriage. To illustrate the susceptibility differences of each category of isolates, these mean antibiograms were plotted. Root-mean-square deviations (RMSD) were also estimated for each category, using the mean susceptibility rates of overall isolates as a reference.

We converted all the categorical variables into numerical ones, and then split the dataset in three parts: training, test and validation. The training and the tests dataset constitutes 80% and 20% respectively of the data registered between 2014 and 2019, whereas the validation

data set was all the data gathered in 2020. For machine learning models, the date of the sample was considered as a distinct variable in order to take into consideration the variation of the antibiotic resistance over time.

In the following sections, we describe the details of each model.

Frequentist inference models

For each antibiotic and each stage preceding antibiogram results, we first trained very simple frequentist inference models (FRQ) using the last year of the training + test dataset (2019) and validated them using the validation dataset (2020). These frequentist models directly estimated the posterior susceptibility probability, as the proportion of antibiotic susceptibility within isolates with similar features within the last year of the training + test dataset (2019) (Eq. 1, example for the “species” stage).

Eq. 1: $$P(S \mid Species, Specimen, Ward, MDR)_{2019} = \frac{N(S|Species,Specimen,Ward,MDR)_{2019}}{N(Species,Specimen,Ward,MDR)_{2019}}$$

where $P\left(S \mid Species, Specimen, Ward, MDR\right)_{2019}$ is the probability of susceptibility $P\left(S\right)$ to a given antibiotic group or combination, within the last year of the training + test dataset (2019), given the $Species$ of bacteria grown from positive cultures, the $Specimen$ origin, the type of hospital $Ward$, and the previous history of MDR carriage (Supplementary Table S2). In case no similar situation was available in the training + test dataset, we chose to attribute a probability of susceptibility of 0.5, so that these isolates were not excluded from AUC estimations. With this kind of “intention to treat” analysis, AUCs were thus impaired by prediction failures, whereas a kind of “per protocol” analysis excluding prediction failures would have artificially overestimated AUCs.

Bayesian inference models

Using the whole training + test dataset (2014–2019), we trained one Bayesian inference model (BAY) per antibiotic and stage and validated them using the validation dataset (2020). We first estimated likelihood ratios, LRT, from prior and posterior susceptibility probabilities, after converting probabilities $P\left(S\right)$into odds $O\left(S\right)$ (Eq. 2 and Eq. 3) [23]. For all combinations of isolates features, posterior probabilities were then approximated, using prior probabilities and corresponding likelihood ratios (Eq. 4), and conversions from odds $O\left(S\right)$ to probabilities $P\left(S\right)$ (Eq. 5).

Eq. 2: $$O\left(S\right) = \frac{P(S)}{(1-P(S))}$$

Eq. 3: $$LRT_{Species} = \frac{O(S/Species)}{O(S)}$$

Eq. 4:
$$O\left(S|\, Species, Specimen, Ward, MDR\right)_{2019} = O\left(S\right)_{2014-19} \times LRT_{2019} \times LRT_{Species} \times LRT_{Specimen} \times LRT_{Ward} \times LRT_{MDR}$$

Eq. 5: $$P\left(S\right) = \frac{O(S)}{(1+O(S))}$$

where $O\left(S\right)_{2014-19}$ is the prior probability of susceptibility S to a given antibiotic group or combination, within the whole training + test dataset (2014-2019), LRT_{2019} is the likelihood ratio for the last year of the training + test dataset (2019), and $LRT_{Species}$, $LRT_{Specimen}$, LRT_{Ward} and LRT_{MDR} are the likelihood ratios for the $Species$, the $Specimen$, the hospital $Ward$, and the previous MDR carriage, respectively (Supplementary Table S2). Models were implemented using Python v3.6.

Machine learning models

We have trained and tested state of the art supervised machine learning models using the training + test dataset (2014–2019), and validated them using the validation dataset (2020): Logistic Regression (LR), as abaseline; ensemble models (AdaBoost (ADA), Gradient Boosting (GBS), Extreme Gradient Boosting (XGB), Bagging (BAG) [24], Random Forest (RF) [25], and neural networks (NN). Support-vector machine (SVM) models were discarded in the exploratory analysis since they did not present a good performance, which was not surprising given that our dataset included only few categorical features.

Logistic regression is a generalized linear model that computes a weighted sum of the input features (plus a bias term) in order to estimate the probability that an instance belongs to a particular class.

Ensemble models are a group of learning methods that can combine several weak learners into a strong learner, and that provide a decision based on a majority voting of all the ensemble. We chose models using decision tree learners given their suitability to deal with categorical variables. We first considered three boosting methods, which train predictors sequentially so each predictor aims to correct its predecessor:

AdaBoost (ADA) [26] predictors focus specifically on cases that previous learners have underfit, and thus progressively become better in classifying difficult cases; Gradient Boosting (GBS) [27] tries to fit the new predictor to the residual errors made by the previous one; and Extreme Gradient Boosting (XGB) [28] uses a more regularized model formalization than GBS to control over-fitting, which generally improves performance. Bagging (BAG) [24] is another ensemble learning approach, which applies the same training model for every predictor, and trains each predictor on different random subsets of the training set. An example of this type of model is Random Forest (RF) [25] is considered as an improvement of bagging models that changes the way the sub-trees are learned, so that predictions resulting from all of the subtrees have less correlation.

Finally, artificial neural networks (NN) [29] are based on a collection of connected unit functions or nodes called artificial neurons. In this contribution we have implemented one dense neural network for each stage. The implemented NN based models are composed of one input layer, three hidden dense layers, one batch normalization layer before each hidden dense layer, one dropout layer after each dense hidden layer to reduce overfitting and one output layer with one node per antibiotic or antibiotic combination. The number of nodes per hidden layer ranges between 50 and 150 depending on the position of the hidden layer and the stage. The number of nodes in each layer, the number of layers and the dropout rate were chosen following a procedure of grid search with cross validation. The activation functions are set to 'ReLU' (Rectified Linear Unit) for the intermediates layer and to 'sigmoid' for the final layer.

To train the different ML models we split the 2014–2019 dataset into 80% for training and 20% for the test. For each of the considered models, we implemented one classifier per stage with one output per antibiotic. Cofactors or features included in each model are summarized in (Supplementary Table S2). For stage 4 (Species), we also included the culture type as this cofactor markedly improved predictions for rare situations, probably because the antibiotic susceptibility of rare species is generally close to the susceptibility of species from the same culture type (for instance Non-fermentative Gram-negative rods). However, we chose not to include the direct type as a cofactor within stages 3 and 4 models as such types (for instance Gram-positive cocci) include numerous species with contrasting susceptibility patterns

Models were implemented using Python v3.6, and the Scikit-Learn Python library [30]. For the neural networks, we used the Keras library and KerasClassifier, a scikit-learn API wrapper [31]. To ensure model generalization, all models were trained using a 10-fold cross-validation procedure over the train dataset, then tested using the test dataset and validated using the validation dataset, the displayed results are the one obtained with validation dataset. To select the hyperparameters of each model, we implemented a grid search procedure with cross validation.

Model comparisons

We then compared the prediction performance of frequentist, Bayesian and machine learning models by conducting a receiver operator characteristic (ROC) analysis, with antibiotic susceptibility probabilities as binary classifiers and antibiograms as outcomes using the validation dataset. We estimated areas under the ROC curve (AUC) for each antibiotic and antibiotic combination and at each stage, and estimated the mean AUC for each model at each stage. We plotted mean ROC curves as well as AUC distributions. We then compared models by their mean AUC and by their prediction failure rate (for FRQ models): (1) globally; and (2) for the least frequent situations only (5th percentile of the number of occurrences in the 2014–2019 dataset: e.g. a rare species in a rare specimen in a rare ward etc…). Finally, we chose the best model family.

Model interpretability

Understanding predictions of the models is key in any application for health purposes. This was not an issue for our FRQ and BAY models, which reflect the mean antibiograms shown in Fig. 1, through the direct combination of features (FRQ) or feature-specific likelihood ratios (BAY). In LR influence of covariates is included in odds ratios. RF models also allow feature interpretation by providing the degree of contribution of each variable to the final decision. But this gets more challenging with other methods that are not directly interpretable and are often considered as "black boxes", especially for the case of Neural Networks. For such cases, we thus used the SHapley Additive exPlanations (SHAP) as done in other machine learning implementations for health purpose predictions [33,34]. SHAP is based on the Shapley value, a game theoretic approach that calculates the average marginal contribution of a feature value across all possible coalitions. We have applied this SHAP analysis, using the SHAP Python package [32], to the machine learning models chosen after comparing prediction performances, and obtained the relative contribution of each cofactor, which was plotted for each stage of the identification process. Interpretability of each cofactor can then be obtained and visualized. As an illustration of how cofactors can influence the prediction of antibiotic susceptibility, we then plotted SHAP values at the individual level using a didactic scenario. Using the same scenario, we finally compared NN predictions and covariate-specific SHAP values with the corresponding BAY predictions and likelihood ratios.

Ethics

This study exclusively used electronic health records routinely collected by the bacteriology lab of Hôpital Européen. All data were pseudonymized. Authors received authorization #DR-2020-047 of the French national commission for data protection (CNIL, Commission nationale de l'informatique et des libertés).

Declarations

Data and code availability

The original data used in this study are derived from patient health records and are not publicly available. Anonymized data are, however, available from Hôpital Européen Marseille upon reasonable request and signing a material transfer agreement.

The code used for data analysis is available upon request.

Acknowledgement

We thank Swapnesh Panigrahi, Leon Espinosa, Tâm Mignot and Renaud Piarroux for their feedback and suggestions in the analysis plan. This work was directly supported by Hôpital Européen Marseille and SESSTIM and was not conducted using specific funding.

Author contributions

SR developed the study idea. SR, RU, JCD and JG contributed to the analysis plan. SR and RU performed the literature review. SR and SC did the data investigation and management. All authors had full access to the data and analyses and were responsible for the final decision to submit for publication. SR, YB, and RU participated in the data management and analysis. SR and RU wrote the first draft. All authors contributed to the interpretation of the findings, reviewed the analysis, wrote the manuscript, and approved the final manuscript. SR and RU shared the final responsibility for the decision to submit for publication.

Competing interests

We declare no competing interests.

References

1. Antibiotic resistance. World Health Organization (WHO) https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance (2018).
2. Strich, J. R., Heil, E. L. & Masur, H. Considerations for Empiric Antimicrobial Therapy in Sepsis and Septic Shock in an Era of Antimicrobial Resistance. J. Infect. Dis. **222**, S119–S131 (2020).
3. IDSA Practice Guidelines. Infectious Diseases Society of America (IDSA) https://www.idsociety.org/practiceguidelines#/name_na_str/ASC/0/+/.
4. Société de Pathologie Infectieuse de Langue Française (SPILF). Recommandations. *Infectiologie.com* https://www.infectiologie.com/fr/recommandations.html.
5. Yelin, I. *et al.* Personal clinical history predicts antibiotic resistance of urinary tract infections. Nat. Med. **25**, 1143 (2019).
6. Plachouras, D. *et al.* Antimicrobial use in European acute care hospitals: results from the second point prevalence survey (PPS) of healthcare-associated infections and antimicrobial use, 2016 to 2017. Eurosurveillance **23**, 1800393 (2018).
7. Bremmer, D. N., Trienski, T. L., Walsh, T. L. & Moffa, M. A. Role of Technology in Antimicrobial Stewardship. Med. Clin. North Am. **102**, 955–963 (2018).
8. Peiffer-Smadja, N. *et al.* Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clin. Microbiol. Infect. **0**, (2019).
9. Forrest, G. N. *et al.* Use of electronic health records and clinical decision support systems for antimicrobial stewardship. Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am. **59 Suppl 3**, S122–133 (2014).
10. Vasudevan, A., Mukhopadhyay, A., Li, J., Yuen, E. G. Y. & Tambyah, P. A. A prediction tool for nosocomial multi-drug Resistant Gram-Negative Bacilli infections in critically ill patients - prospective observational study. BMC Infect. Dis. **14**, 615 (2014).
11. Oonsivilai, M. *et al.* Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in Cambodia. Wellcome Open Res. **3**, 131 (2018).
12. MacFadden, D. R. *et al.* Utility of prior cultures in predicting antibiotic resistance of bloodstream infections due to Gram-negative pathogens: a multicentre observational cohort study. Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis. **24**, 493–499 (2018).

13. MacFadden, D. R. *et al.* Decision-support models for empiric antibiotic selection in Gram-negative bloodstream infections. *Clin. Microbiol. Infect.* **25**, 108.e1-108.e7 (2019).
14. Chico, V. The impact of the General Data Protection Regulation on health research. Br. Med. Bull. **128**, 109–118 (2018).
15. Vorisek, C. N. *et al.* Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review. JMIR Med. Inform. **10**, e35724 (2022).
16. Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet Lond. Engl. **399**, 629–655 (2022).
17. Société Française de Microbiologie (SFM), Société Française de Mycologie Médicale (SFMM) & Société Française de Parasitologie. *Rémic - Référentiel en microbiologie Médicale*. (2018).
18. Société Française de Microbiologie. *CASFM / EUCAST Recommandations 2021 V.1.0 Avril*. (2021).
19. Haute Autorité de Santé. Antibiothérapie des infections à entérobactéries et à Pseudomonas aeruginosa chez l'adulte: place des carbapénèmes et de leurs alternatives. Recommandation de bonne pratique. *Haute Autorité de Santé* https://has-sante.fr/jcms/c_2968915/fr/antibiotherapie-des-infections-a-enterobacteries-et-a-pseudomonas-aeruginosa-chez-l-adulte-place-des-carbapenemes-et-de-leurs-alternatives (2019).
20. Murray, P. R. The Clinician and the Microbiology Laboratory. Mand. Douglas Bennetts Princ. Pract. Infect. Dis. 191–223 (2015) doi:10.1016/B978-1-4557-4801-3.00016-3.
21. European Committee on Antimicrobial Susceptibility Testing (EUCAST). EUCAST: Expert rules and intrinsic resistance. http://www.eucast.org/expert_rules_and_intrinsic_resistance/.
22. Klinker, K. P. *et al.* Antimicrobial stewardship and antibiograms: importance of moving beyond traditional antibiograms. Ther. Adv. Infect. Dis. **8**, 20499361211011372 (2021).
23. Deeks, J. J. & Altman, D. G. Diagnostic tests 4: likelihood ratios. BMJ **329**, 168–169 (2004).
24. Breiman, L. Bagging predictors. Mach. Learn. **24**, 123–140 (1996).
25. Ho, T. K. Random decision forests. in Proceedings of 3rd International Conference on Document Analysis and Recognition vol. 1 278–282 vol.1 (1995).
26. Schapire, R. E. Explaining AdaBoost. in Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik (eds. Schölkopf, B., Luo, Z. & Vovk, V.) 37–52 (Springer, 2013). doi:10.1007/978-3-642-41136-6_5.
27. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. Ann. Stat. **29**, 1189–1232 (2001).
28. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (2016). doi:10.1145/2939672.2939785.
29. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 2554–2558 (1982).
30. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011).
31. Keras: Deep Learning for humans. (2022).
32. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in Advances in Neural Information Processing Systems vol. 30 (Curran Associates, Inc., 2017).
33. Lundberg, S. M. *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat. Biomed. Eng. **2**, 749–760 (2018).
34. Bibault, J.-E. *et al.* Development and Validation of an Interpretable Artificial Intelligence Model to Predict 10-Year Prostate Cancer Mortality. Cancers **13**, 3064 (2021).

Supplementary Tables

Supplementary Tables 1-3 not available with this version.

Figures

Figure 1

Mean antibiograms for single antibiotics (A and B, traditional antibiogram) or antibiotic combinations (C and D, combination antibiograms), for main direct types, culture types and species (A and C), and for main specimen types, main ward types, past multidrug bacteria carriage history and periods (B and D), including root-mean-square deviation (RMSD) for each category.

Amox-Clav, Amoxicillin-Clavulanate; Pip-Tazo, Piperacillin-Tazobactam; TMP-SMX, Trimethoprim-Sulfamethoxazole; CTX / CRO, Cefotaxime / Ceftriaxone.

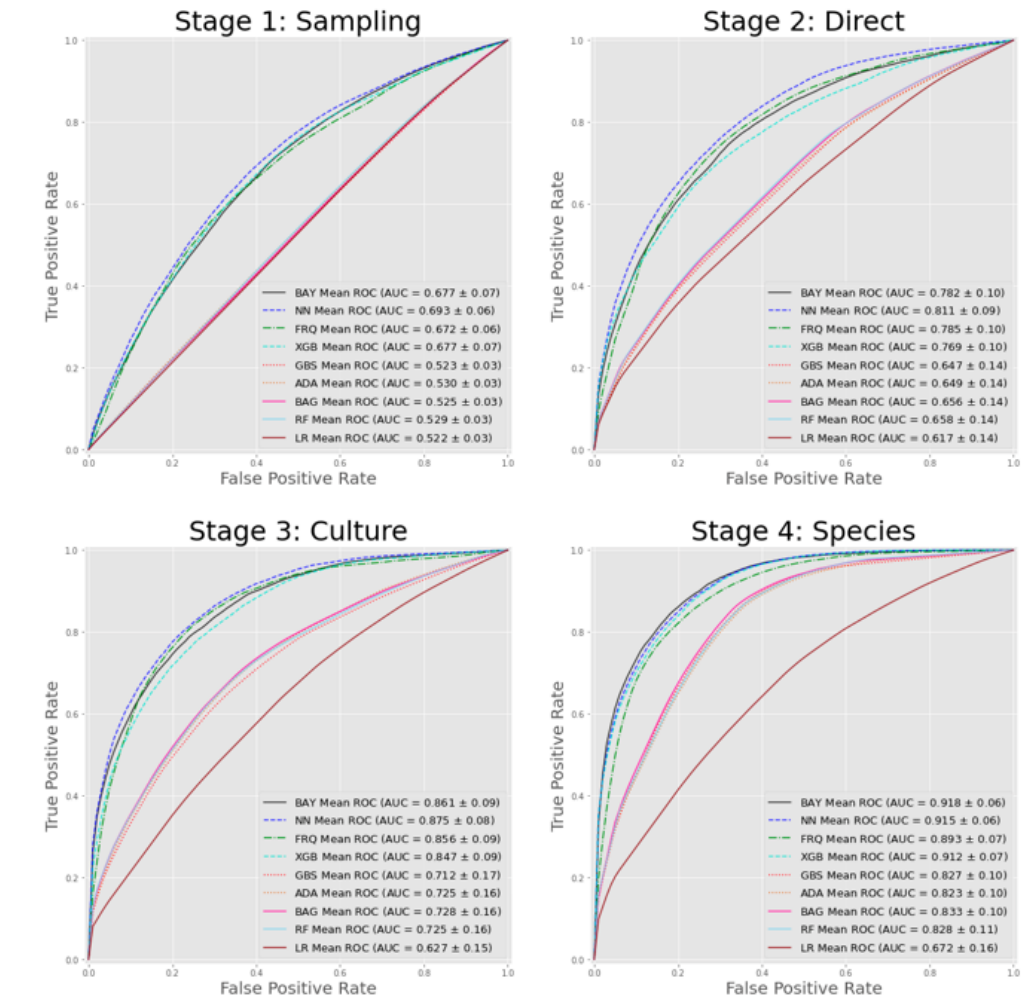

Figure 2

Mean global receiver operator characteristic (ROC) curves (and standard deviations) of frequentist, Bayesian and machine learning antibiotic susceptibility prediction models at each stage of the identification process and for all isolates of the 2020 validation dataset. Models: BAY, Bayesian; NN, Neural Network; FRQ, frequentist; XGB, Extreme Gradient Boosting; GBS, Gradient Boosting; ADA, AdaBoost; BAG, Bagging; RF, Random Forest; LR, Logistic regression.

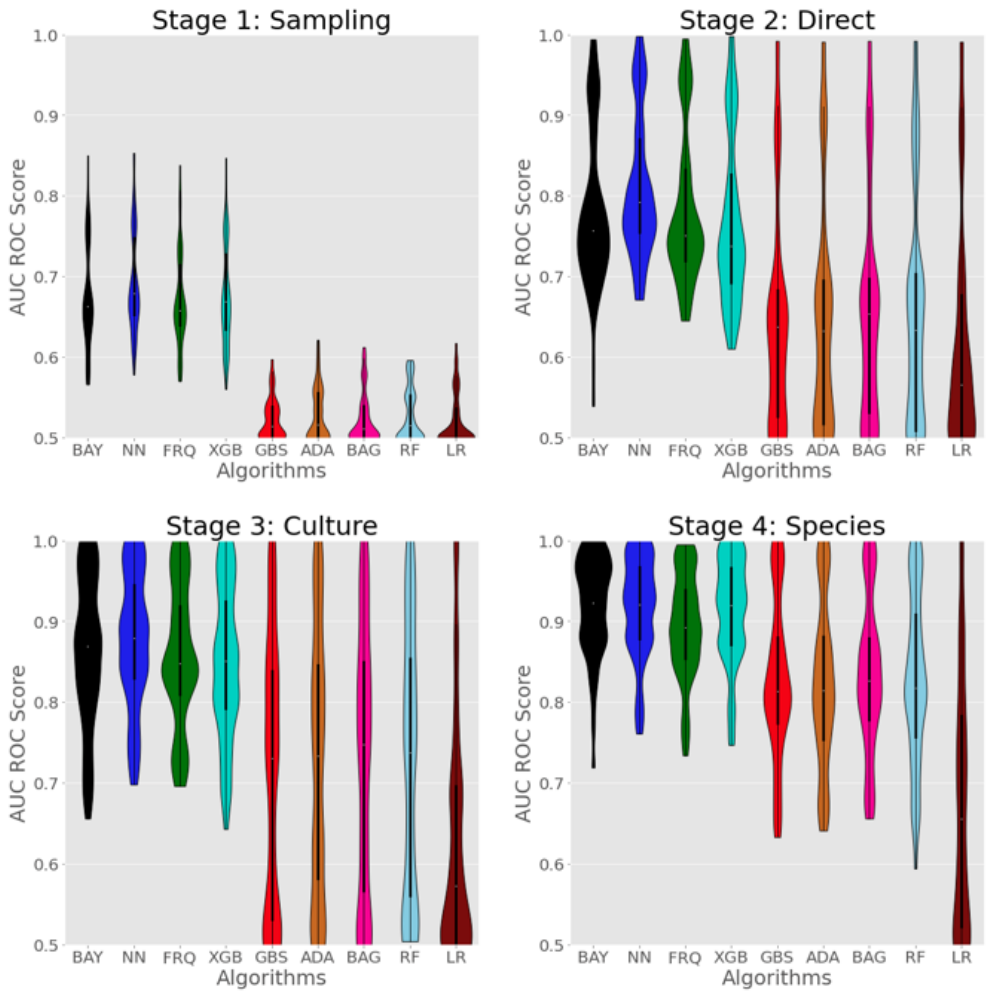

Figure 3

Distribution of individual areas under the receiver operator characteristic curve (ROC AUC) of frequentist, Bayesian and machine learning susceptibility prediction models for each antibiotic, at each stage of the identification process and for all isolates of the 2020 validation dataset. Models: BAY, Bayesian; NN, Neural Network; FRQ, frequentist; XGB, Extreme Gradient Boosting; GBS, Gradient Boosting; ADA, AdaBoost; BAG, Bagging; RF, Random Forest; LR, Logistic regression.

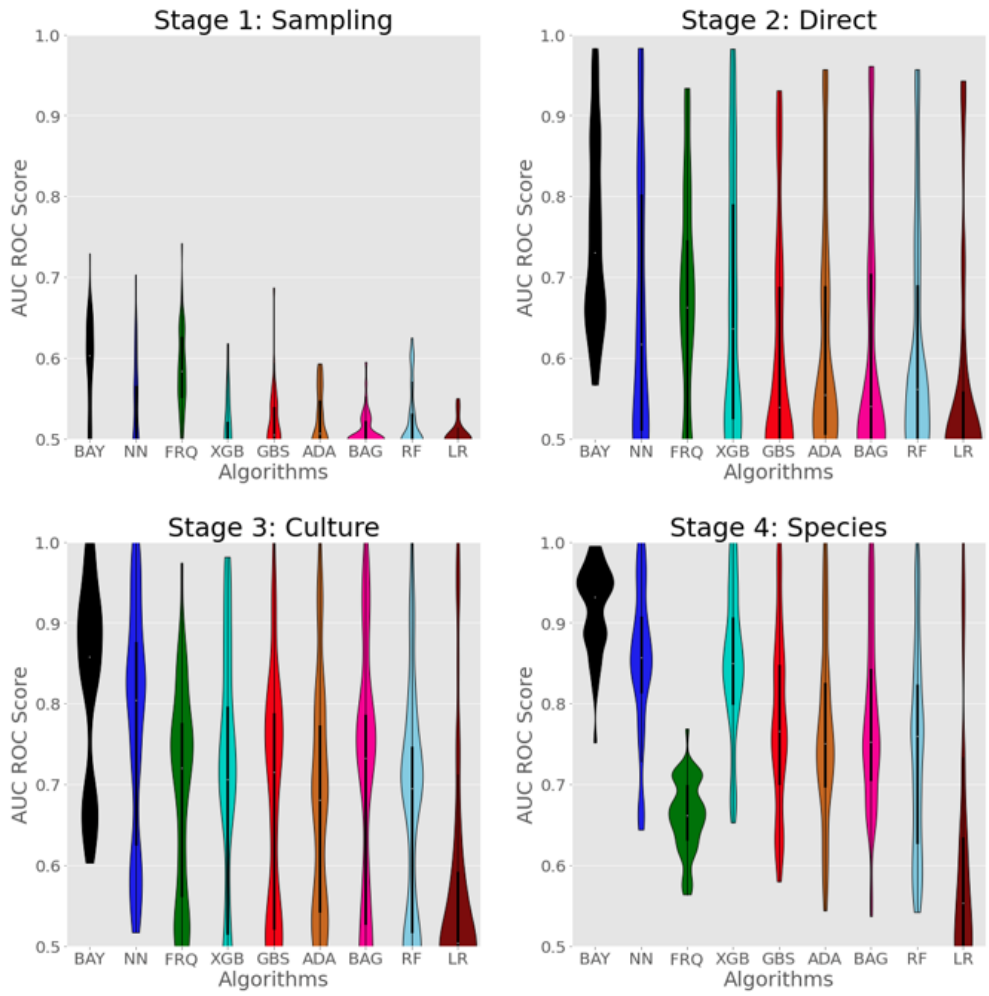

Figure 4

Distribution of individual areas under the receiver operator characteristic curve (ROC AUC) of frequentist, Bayesian and machine learning susceptibility prediction models for each antibiotic, at each stage of the identification process and for isolates of the 2020 validation dataset corresponding to the least frequent situations only (5^{th} percentile of the number of occurrences in the 2014-2019 dataset). Models: BAY, Bayesian; NN, Neural Network; FRQ, frequentist; XGB, Extreme Gradient Boosting; GBS, Gradient Boosting; ADA, AdaBoost; BAG, Bagging; RF, Random Forest; LR, Logistic regression.

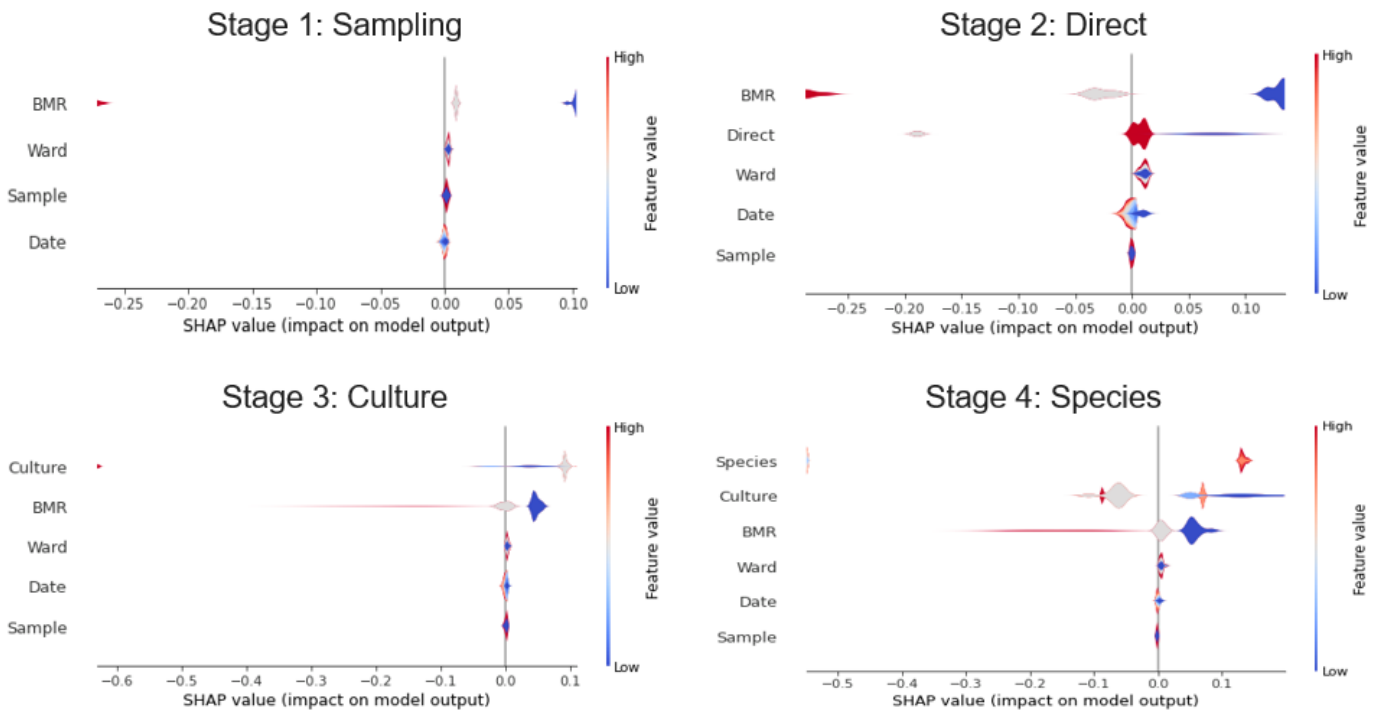

Figure 5

Relative importance of cofactors in the NN antibiotic susceptibility prediction models, at each stage of the identification process at population level, using the SHapley Additive exPlanations (SHAP). A cofactor with a negative SHAP value will decrease the probability of antibiotic susceptibility.

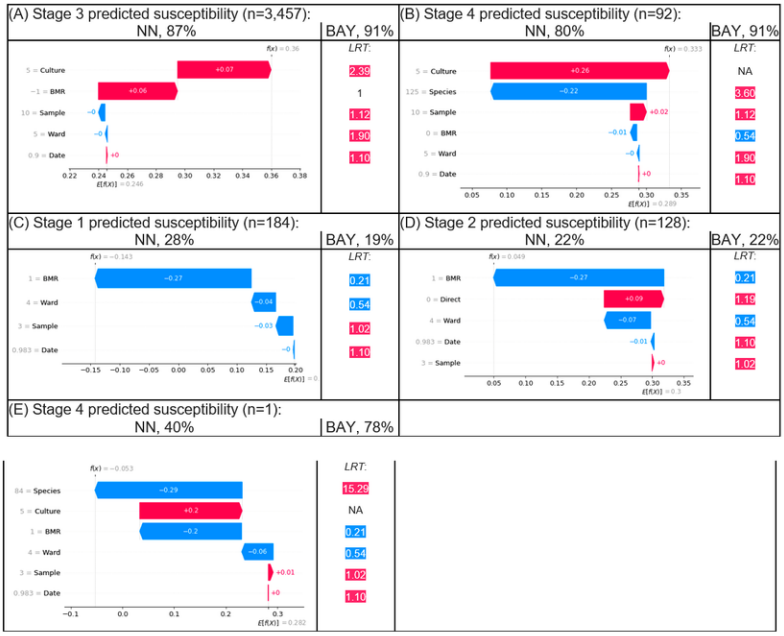

Figure 6 – Illustration of the interpretability of the NN and BAY models at individual level, using a didactic scenario with prediction of the susceptibility to Cefotaxime / Ceftriaxone (CTX / CRO) at various stages. (A) A patient unknown from the hospital is managed at the emergency room for a urinary tract infection; a urine sample is sent to the laboratory and grows an Enterobacteriaceae-like Gram-negative bacteria (stage 3). (B) In reality she has no past history of BMR carriage, and the urine sample grows *Escherichia coli* (stage 4). According to the antibiogram, this *E. coli* produces an extended-spectrum beta-lactamase (ESBL). (C) One month later, the same patient is admitted in critical care for a septic shock and blood cultures are sampled (stage 1). (D) These blood cultures grow a Gram-negative rod (stage 2). (E) A *Citrobacter koseri* is identified (stage 4). For each stage, the panel lists the number of similar situations in the training+test dataset, the susceptibility predictions with NN and BAY models. For NN models, the waterfall plot shows how each cofactor contributes to push the model output from a base value $E[f(X)]$ (the average model output over the training dataset) to the individual output $f(x)$. Covariates pushing the prediction of the susceptibility to CTX / CRO higher are shown in red, those pushing the prediction lower are in blue. For BAY models, each panel presents the likelihood ratio (*LRT*) associated with each covariate. *LRT* >1 pushes the susceptibility prediction higher, *LRT* <1 pushes the prediction lower.

Figure 6

See above image for figure legend.